El problema de la fundamentación del clustering

Una investigación sobre la fundamentación del clustering como teoría matemática.

Lic. De Oliveira Gaffrée, Candido Lucas

Tesis de maestría
Maestría en matemática

Orientador:
José Rafael León Ramos

UNIVERSIDAD
DE LA REPÚBLICA
URUGUAY

Facultad de Ciencias
Universidad de la República

Montevideo-Uruguay
Noviembre

Abstract

This work lies within the theoretical study of clustering or unsupervised classification, a field concerned with grouping data without any prior labeling or training information. Although many practical methods have been developed—such as k-means, single-linkage, and spectral clustering—the construction of a general theory of clustering remains an open problem. From Wright's early formalization attempt (1973) and Kleinberg's well-known impossibility theorem (2002) to Margareta Ackerman's doctoral work (2012), the main challenge persists: defining, in formal terms, what a clustering function truly is.

The aim of this work is to offer an initial contribution toward such a theoretical foundation. The proposed approach proceeds inductively, starting from specific methods of unsupervised classification and moving toward a more general and abstract framework. The central thesis argues that the perceived ambiguity in defining clustering stems from the lack of a clear definition of its proper domain of study. It is maintained that what is often referred to as clustering or unsupervised learning is, more precisely, a problem of classification by similarity. Once this distinction is made explicit, much of the conceptual ambiguity surrounding the notion of clustering disappears.

Resumen

Este trabajo se enmarca dentro del estudio teórico del clustering o clasificación no supervisada, disciplina que busca agrupar datos sin disponer de información previa o etiquetas de entrenamiento. Si bien existen numerosos métodos prácticos —como k-medias, single-linkage o clustering espectral—, el desarrollo de una teoría general del clustering continúa siendo una cuestión abierta. Desde los intentos iniciales de Wright (1973) y el conocido teorema de imposibilidad de Kleinberg (2002), hasta los avances de Margareta Ackerman (2012), persiste la dificultad de ofrecer una fundamentación axiomática coherente que defina qué es, en sentido formal, una función de clustering.

El objetivo de este trabajo es ofrecer una contribución incipiente a dicha fundamentación. Se propone abordar el problema de forma ascendente: partir del análisis de métodos concretos de clasificación no supervisada, para luego ascender hacia una visión más general y abstracta. La tesis central que se argumentará es que la dificultad de formalizar el clustering proviene de la

ambigüedad en la delimitación del propio campo de estudio. Se sostendrá que lo que habitualmente se denomina clustering o aprendizaje no supervisado corresponde, en realidad, a un problema de clasificación por similitud, y que reconocer esto permite superar gran parte de la ambigüedad conceptual que ha obstaculizado el desarrollo de una teoría general del clustering.

Agradecimientos

Quisiera agradecer a mi Abbá celestial, a mi Padre que está en los cielos, que me hizo hijo en el Hijo, el Padre de las luces por iluminarme con esta temática que tanto le rogué. A mi Señor y mi Dios, Jesucristo, por ser mi Guía y mi Amigo, mi Sabiduría y mi Luz, quien me ha guiado a lo largo de esta tesis y maestría. Al Don del cielo, mi Maestro interior, el Espíritu Santo, que me guió interiormente con sus suaves y divinas mociones dándome las gracias que necesité para escribir este texto. A ellos sea la gloria por siempre y el principal mérito de esta obra, en todo lo que contenga de verdadero y bueno. Y también a mi Madre María Santísima, por todos sus cuidados y afectos maternales para conmigo, que ella eleve esta obra al Señor para que sea agradable a Él; y a su bienaventurado esposo San José, de quien aprendo cada día a ser un mejor varón, esposo y trabajador.

Agradezco a mi amada esposa Yanina Fernández por todo el amor, soporte y por los sacrificios que tuvo que hacer para que esta obra pudiera completarse. Agradezco también a mi hija, cuya sola presencia fue motivo de alegría y perseverancia durante este camino.

Agradezco a mis padres, fieles guardianes y auxiliadores en mis necesidades y dificultades por las que he atravesado; sin ellos esta obra no sería posible.

Agradezco a mi querido tutor: José Rafael León, por toda su guía intelectual, su tiempo, dedicación y esfuerzo; así como la paciencia que me tuvo todos estos años.

Agradezco a mis amigos y hermanos en la fe, por sus oraciones, apoyo y compañía durante estos años.

Agradezco también al tribunal que juzgó esta obra, por el tiempo y dedicación.

También a la Universidad de la República, gracias a la cual tuve la oportunidad de realizar este trabajo y cursar la maestría.

Y por último un agradecimiento a todos los investigadores y autores que han contribuido a esta hermosa teoría, con el destaque especial a Jon Kleinberg y Margareta Ackerman, quienes inspiraron este trabajo.

Índice general

Capítulo 1

Introducción

1.1. Planteamiento del problema

Cuando se habla de clasificación estadística y matemática se hace referencia al estudio de cómo clasificar datos cuantitativos. Dentro de las diversas técnicas, metodologías y estudios de la clasificación, una distinción habitual es diferenciar por clasificación supervisada, semisupervisada y no supervisada (también llamada de "Clustering"). La primera consiste en lo siguiente: clasificar nuevos datos a partir de un conjunto de datos ya clasificados previamente, generalmente muy voluminoso. Un ejemplo sería el de una base de datos con información de pacientes de un hospital, clasificados según tengan o no una enfermedad particular. Supóngase que ingresa un nuevo paciente al sistema, ¿hay alguna forma de deducir qué tan probable es que tenga dicha enfermedad, a partir de los datos previos? Esta es la pregunta que intenta responder la clasificación supervisada. Por eso mismo, es muy útil para lo que al día de hoy se conoce como inteligencia artificial o "machine learning". La clasificación semisupervisada consiste en algo semejante, pero reduciendo considerablemente el conjunto de datos iniciales ya clasificados (también conocidos como: datos de entrenamiento). La idea es clasificar nuevos datos a partir de pocos ya previamente clasificados. Por último, y el que interesa a este trabajo, está la clasificación no supervisada. Su nombre significa que no tengo datos de entrenamiento, no se dispone de una base de datos ya clasificada. Se desea clasificar un conjunto de datos a partir de ellos mismos. Por eso también se le llama "clustering", que en inglés significa "agrupamiento". La idea es agrupar los datos en "clusters" ("grupos"), es decir, clasificarlos, pero sin datos de entrenamiento. Se han desarrollado diversas técnicas que se expondrán en este documento, desde métodos como k-medias y single-linkage hasta clustering espectral, pero a pesar de ellas hay muy poco desarrollo en

lo que respecta a una teoría general de la clasificación no supervisada.

A este respecto, uno de los primeros en comentar esto fue William E. Wright en su artículo "A Formalization of Cluster Analysis" [36] publicado en 1973. En su introducción dice: "Thus it seems clear that there is a strong need for the development of a *theoretical* basis for this analytical method known as clustering, in contrast to the almost completely *heuristic* approach which has been so dominant up to this time. That is the purpose of this paper, to develop a theoretical clustering model, or to present a formalized notion of clustering analysis" ("Por lo tanto parece claro que existe una fuerte necesidad para el desarrollo de una base *teorética* para el análisis del método conocido como "clustering", en contraste con el método prácticamente *heurístico*, que ha sido el dominante hasta ahora").

Más tarde, Jon Kleinberg en su célebre artículo [1], publicado en "Advances in Neural Information Processing Systems 15" en 2002, titulado "An Impossibility Theorem for Clustering" describe en su "abstract" que si bien la teoría del clustering se centra en una idea muy intuitiva, a pesar de esto ha sido muy difícil desarrollar un paradigma común de razonamiento sobre el mismo a un nivel técnico.

Esto se debe al hecho de que para él, la idea intuitiva consiste en: "datos semejantes deben estar agrupados y datos muy distintos, separados". Pero a partir de ella es difícil dar con un paradigma axiomático como punto de partida para un desarrollo teórico. Más aún, en su propio artículo él intenta dar con uno proponiendo tres axiomas, los cuales resultan ser inconsistentes; de ahí el nombre de su artículo.

Siguiendo esta línea, Margareta Ackerman desarrolló en su tesis doctoral, posiblemente de toda la literatura existente, la mayor contribución hasta la fecha: "Towards Theoretical Foundations of Clustering" (Hacia una fundamentación teórica del clustering) [12], de 2012. En esta tesis, Margareta se propone contribuir al problema de estudiar matemáticamente el clustering desde el punto de vista más general posible.

Después de su tesis, se ha avanzado poco en la dirección de desarrollar una teoría general del clustering. Es sobre este problema que se enmarca esta tesis.

1.2. Objetivo

Dar una solución, al menos incipiente, al problema de qué es el clustering, o el aprendizaje no supervisado, cuáles axiomas debería tener y dónde estuvo la dificultad en la formalización, hasta el momento. Se entenderá por "clustering" lo que los autores citados anteriormente han entendido. La diferencia

con los intentos anteriores está en el procedimiento. Se pretende proceder por vía de "ascenso", desde lo más particular hasta subir a lo más general. Primero se estudiarán los diversos métodos reales y concretos de clasificar datos sin supervisión, para luego pasar a exponer las contribuciones de los diversos autores sobre una teoría general del clustering, y a partir de todo dar con una conclusión que busque ayudar en la resolución de este problema.

Un segundo objetivo es organizar las ideas y desarrollos en sus aspectos más fundamentales e importantes. Brindar una revisión y crítica a lo investigado y dar ideas de cómo continuar una futura investigación.

Un último objetivo es servir como referencia a futuros investigadores sobre esta temática. Se pretende compendiar la información más relevante considerada, para servir como base para futuras investigaciones y desarrollos sobre este tema. Se busca facilitar la introducción a esta temática, e investigación, para futuros investigadores interesados en la misma.

1.3. Tesis

La conclusión que se piensa argumentar es la siguiente: *el obstáculo por el cual se considera "ambiguo", o difícil, de definir qué es una función de clustering, o qué es un clustering, está en la ausencia de una definición formal del campo de estudio.* Si bien diversos autores han intentado formular una teoría general, esta no resulta ser verdaderamente una teoría general del aprendizaje no supervisado. Se propondrá que una posible solución consiste en clarificar qué se está estudiando y qué se pretende lograr. Se afirmará que el término que describe con mayor precisión lo que esos autores han desarrollado y estudiado es "clasificación por similitud", pues este refleja con claridad el objeto real de estudio; y que una vez eso está presente la ambigüedad disminuye.

1.4. Relevancia

¿Cuál es el interés de desarrollar una teoría general del clustering? Y más en específico: ¿qué aporte puede dar esta tesis a un estadístico, o a un científico de datos? Para dar con la respuesta, se pretende sintetizarla en los siguientes puntos:

Conocimiento general

A diferencia de cada teoría o conocimiento particular del clustering, como lo puede ser el estudiar métodos asociativos, o basados en centros, entre otros; una teoría general pretende dar definiciones, propiedades, características que permitan comparar diversos métodos. Los seres humanos aprendemos por comparación, y por notar diferencias entre las cosas, luego, es de suma importancia una teoría que evidencie esas diferencias entre algoritmos, métodos y situaciones diversas.

Conocimiento ordenado

Las teorías generales y fundamentales son el principio del conocimiento. No solo dan un punto de partida para cualquier investigador, sino que ordenan el conocimiento subsecuente. El conocimiento ordenado permite tener un mayor entendimiento, no solo de cada algoritmo particular, o método, sino de la diversidad de situaciones, propiedades, análisis y complejidades de esta herramienta.

Poco desarrollo

Un aspecto interesante es el escaso desarrollo sobre estos temas. Dentro de mi punto de vista, esto lo hace un tema muy interesante para contribuir en el conocimiento humano, especialmente en un área con pocas contribuciones y de gran relevancia potencial.

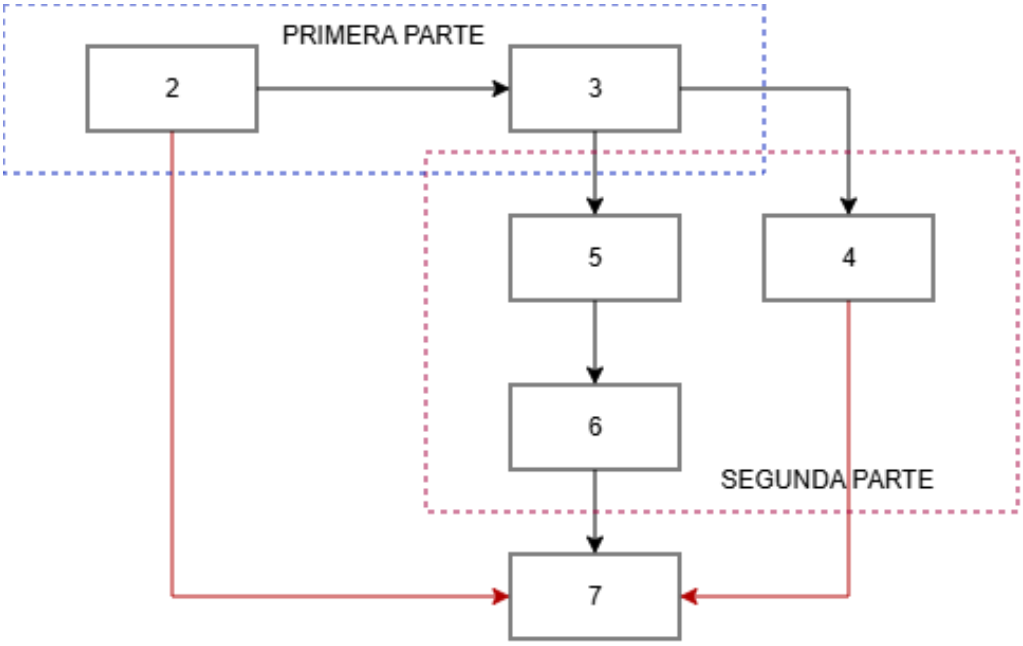
1.5. Metodología

Esta investigación se estructuró en dos partes con finalidades diferentes, pero ambas apuntando a contribuir a la conclusión final, expuesta en la tercera parte de este documento. La primera de ellas se titula: *fenomenología de la clasificación no supervisada*. El objeto de la misma está en estudiar el **fenómeno** del clustering tal cual se presenta, sin teorizar a partir de él. Es el pie en la realidad más concreta. Esto, porque se parte del siguiente principio: una teoría general del clustering debe ser aplicable a los casos particulares, y no una abstracción sin un soporte en lo real. Por este motivo, en esa sección se estudian diversos métodos de clustering. La información fue retirada de una multitud de fuentes: artículos científicos, libros, capítulos de libros, etc. Cada una es citada en su respectivo lugar, y en la bibliografía se encuentra todo detallado.

La segunda parte se titula: *Desarrollos de una teoría general del clustering*. En ella, se exponen los diversos intentos de generar una teoría general del clustering. El centro de dicho apartado es la tesis doctoral de Margareta Ackerman. Es a partir de su tesis que se obtiene la línea temporal de los desarrollos anteriores hasta la fecha de su trabajo. Posterior a ella, hay muy pocos avances, y los que hay son de poca relevancia para este trabajo; el más relevante encontrado es [39], por lo que se dará un breve comentario sobre él al final de la tercera parte.

Una vez concluidas esas dos partes, y a partir de ellas, se brindará la conclusión antes comentada. A su vez, se pasará a dar una crítica a cada uno de los autores expuestos en la segunda parte. Se concluirá brindando líneas de futuros desarrollos. La idea es que la fuerza de la conclusión va a derivar en virtud de todo lo expuesto en las primeras dos secciones.

1.6. Estructura de la tesis



Cada rectángulo representa un capítulo. Las flechas negras significan que dicho capítulo es muy importante para entender el siguiente. Las flechas rojas significan que es recomendado.

Sin considerar este primer capítulo introductorio, la tesis se divide en seis capítulos distribuidos en tres partes. La primera parte titulada **fenomenología del clustering** se divide en los capítulos segundo y tercero. Como el

objeto de esa sección es mostrar las técnicas reales más usadas y el fenómeno del aprendizaje no supervisado en la práctica, esta se divide en un primer capítulo dedicado a las **aplicaciones del clustering** (capítulo 2), donde se revisan las principales aplicaciones de lo que tanto en la industria como los investigadores entienden por aplicar técnicas de clasificación no supervisada. Para ese capítulo, la inspiración fue el libro "Cluster Analysis and Applications" ([2]), y algunos artículos retirados principalmente de la bibliografía de ese libro. El siguiente capítulo se llama **algoritmos más importantes del clustering** (capítulo 3); es una exposición sobre las principales técnicas para realizar clasificación no supervisada. Con ese capítulo se pretende presentar, además de su realidad aplicada vista en el capítulo precedente, las diversas técnicas y metodologías matemáticas para estudiar el clustering. Para dicho capítulo, algunas partes se continuó con el mismo libro, pero en otras está inspirado en "Handbook of Cluster Analysis" ([8]) y algunos artículos que se citan en sus partes correspondientes. Si bien en la imagen se propone que es necesario para leer el tercer capítulo haber leído el segundo, en este caso esta sugerencia es para el público que no conoce nada de clustering o clasificación no supervisada. Para una persona con alguna noción, dichos capítulos son independientes y pueden leerse por separado. *El capítulo tercero es el centro y eje de la primera parte.* Si bien es expositivo, no se sienta obligado el lector a entender todo en una primera lectura para proseguir a los siguientes capítulos. Lo necesario para avanzar a la segunda parte es simplemente tener una noción, aunque sea superficial, de lo comentado en dicho capítulo.

La segunda parte, titulada **desarrollos de una teoría general del clustering**, se divide en tres capítulos, organizados en orden cronológico. Por eso, no es necesario leer el capítulo cuarto antes del quinto; si bien es lo que se recomienda. En el cuarto capítulo se trata de las **contribuciones anteriores a Jon Kleinberg** (capítulo 4). Hay varias, pero las principales, en mi opinión, son la de William E. Wright y las de Puzicha, Hofmann y Buhmann. La razón de este capítulo es más a modo expositivo, para que el lector pueda tener más noción del peso e importancia de las contribuciones posteriores. El quinto capítulo trata de exponer el artículo de Jon Kleinberg donde expone su famoso **teorema de imposibilidad** (capítulo 5). El sexto, y más importante en orden a la conclusión, versa sobre **la tesis doctoral de Margareta Ackerman** (capítulo 6). En él, se expone la tesis a modo de resumen comentando sus principales desarrollos.

Como se puede observar en la imagen, los capítulos cuarto y quinto son independientes. Por lo dicho anteriormente, la numeración sigue un orden cronológico. Es indispensable leer el quinto capítulo antes del sexto. En primer lugar, por la notación, se mantiene la notación del quinto capítulo en el sexto. Y en segundo, por una cuestión lógica: Margareta Ackerman en su

tesis continúa el trabajo de Kleinberg y lo profundiza.

Es importante notar que la columna vertebral de esta tesis está en la secuencia: tercer, quinto y sexto capítulo. Estos tres capítulos son los más importantes, funcionando los capítulos segundo y cuarto como auxiliares que amplían el panorama y brindan más riqueza.

Por último, en el capítulo séptimo se da la **propuesta de fundamentación** (capítulo 7). En él, se describe el principal obstáculo que dificultó dar con una teoría general de la clasificación no supervisada. Se sostendrá que esto presenta dificultades fundamentales debido a la naturaleza de lo que significa clasificación no supervisada, y se argumenta que el término más preciso para describir lo estudiado por Kleinberg y Ackerman es: clasificación por similitud. Una vez establecida la distinción entre clasificación no supervisada y clasificación por similitud, se pasa a dar una propuesta axiomática y a revisar las diversas posibilidades a partir del estudio realizado por Ackerman. También se propone qué debería estudiar la clasificación por similitud. Finalmente, se da una crítica a los diversos autores que han tratado este tema. Por esta razón, es necesaria la lectura previa de los capítulos tercero, quinto y sexto. Los capítulos segundo y cuarto, si bien no son centrales, son importantes y muy recomendados, para el séptimo. El segundo capítulo toma relevancia en el argumento principal de esta tesis, y el capítulo cuarto es importante para entender la crítica que se realiza a esos autores.

1.7. Requerimientos previos para la lectura de esta tesis

Los contenidos matemáticos manejados en este documento no son de una profundidad muy elevada. Cualquier persona que esté o haya cursado un grado en matemática es capaz de entender todos los capítulos. Para una lectura más profunda, que busque ir a las fuentes citadas, se requiere experiencia leyendo artículos matemáticos, pero no más que eso. El último capítulo es el que puede generar más complicaciones en un público matemático sin conocimientos en lógica filosófica, especialmente con lo que se conoce como *lógica material*. En él, utilizo el concepto de objeto material y formal de una ciencia, y otros conceptos. Sobre esto, el lector puede consultar el primer capítulo de la cuarta parte de [38].

Parte I

Fenomenología de la clasificación no supervisada

Capítulo 2

Aplicaciones del clustering

Para iniciar nuestro estudio sobre los fundamentos del clustering a nivel matemático comenzaremos analizando las distintas aplicaciones del clustering al día de hoy. Este capítulo es un resumen del capítulo 8 del libro: *"Cluster Analysis and Applications"*[2], junto con otras fuentes que serán citadas en sus respectivos apartados.

2.0.1. Motivación y objetivo

El objetivo de este capítulo es presentar diversas aplicaciones del clustering, para que, una vez familiarizado con ellas, sea más fácil juzgar las diversas propuestas de fundamentación de una teoría general del clustering; así como poder fundamentar la tesis de este documento (Ver el capítulo introductorio).

2.0.2. Resumen de este capítulo

Empezaremos definiendo qué es el clustering según las ciencias de la información y el análisis de datos, para luego pasar a analizar los problemas reales, de hoy en día, que utilizan técnicas basadas en clustering.

2.1. ¿Qué es el clustering?

Primeramente, empezaremos con definiciones dentro del marco de la teoría informática. Por eso mismo, no se hablará del clustering desde el punto de vista *meramente matemático* sino desde su aplicación real.

Definición 2.1.1. (Clustering) Clustering es una técnica estadística que tiene como principal finalidad particionar los datos de forma significativa en

subconjuntos llamados *clusters* y como finalidad secundaria, dar un criterio de clasificación en caso de que haya un nuevo dato. [2][9]

Definición 2.1.2. (Función de clustering) Decimos que una función de clustering es toda función diseñada con la finalidad de realizar un clustering de los datos.

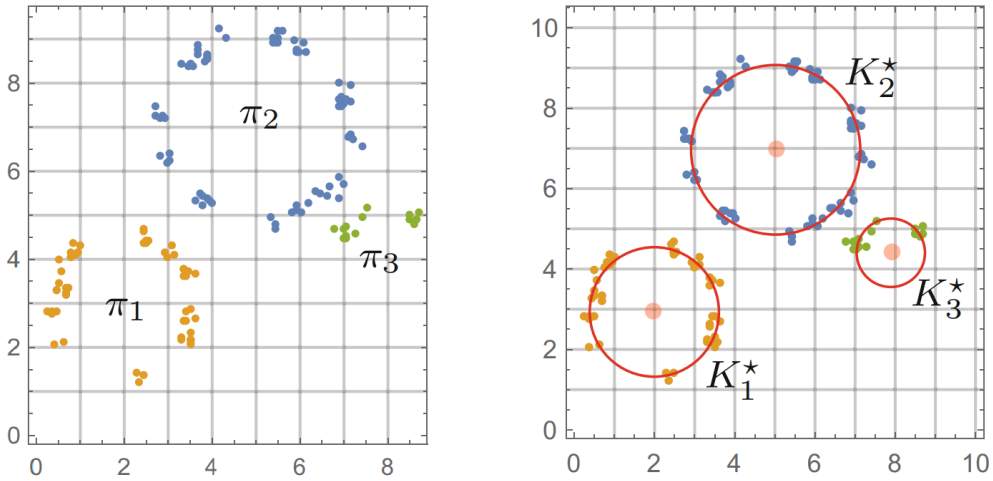
Definición 2.1.3. (Función de costo, u objetivo) Una función de costo es una función que: 1. se busca minimizar para realizar el clustering; 2. está relacionada con el objetivo del investigador; 3. está relacionada con el tipo de clustering a realizar.

Veremos en el capítulo siguiente que existen varios algoritmos distintos para particionar los datos basados en funciones de costo (ver la sección 3.1).

2.2. Detección de objetos geométricos

Una de las aplicaciones más destacadas del clustering al día de hoy está en la detección de patrones y objetos en imágenes y videos. Hay diversas formas de detectar objetos en un video, desde entrenar redes neuronales específicas capaces de detectar ciertos patrones, usar ciertas redes convolucionales, modificar la imagen para luego con ciertos algoritmos detectar los patrones deseados, etc. Pero una aplicación del clustering es detectar ciertos objetos geométricos presentes en una imagen, usando técnicas clásicas de clustering. Dentro de toda la diversidad de objetos geométricos que uno puede desarrollar, destacaré tres: círculos, elipses y líneas [3][4][7][8][9].

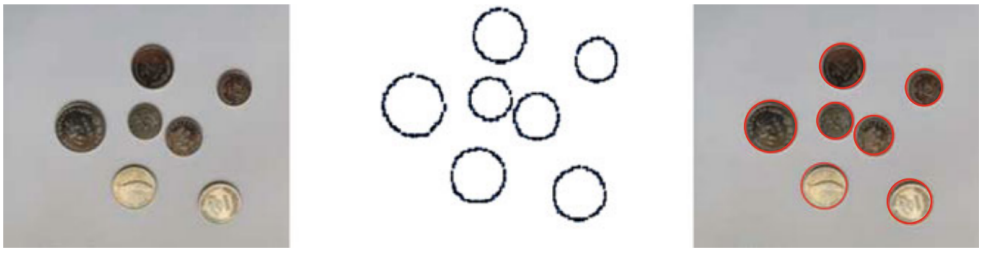
2.2.1. Detección de círculos



Cuando tenemos datos que sabemos que generarán un patrón circular se pue-

den utilizar técnicas de clustering para agruparlos bajo el criterio de pertenecer al mismo círculo (ver la sección 3.1.4 para los detalles matemáticos). Estos datos pueden provenir de diversas fuentes y la aplicación puede servir para diversos fines. Por ejemplo, pueden provenir de imágenes de monedas. Es posible transformar una moneda para dejar únicamente los pixeles que corresponden al borde. Estos pixeles obviamente no tienen por qué formar un círculo perfecto, sino más bien imperfecto. Una aplicación del clustering es utilizarlo para agrupar datos que pertenecen a un mismo círculo en un mismo clúster. Nótese que la razón por la cual se agrupan los datos es geométrica; porque forman un círculo. Más sobre este asunto puede encontrarse en [2]. Un ejemplo puede verse en la siguiente imagen:

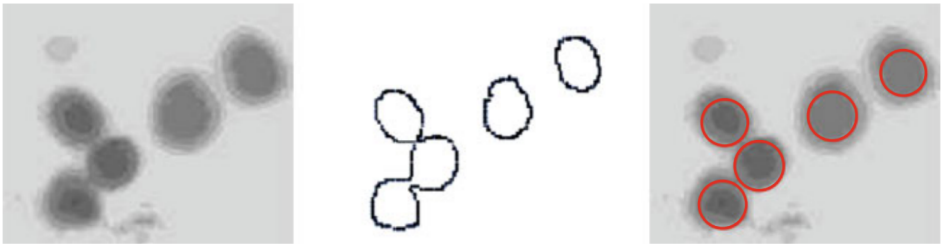
(a) Coins (b) Edge detection (c) Detection



Todas las imágenes de este apartado 2.2.1 son de *"Cluster Analysis and Applications"*[2].

Aplicación a la medicina para la deteccion de cáncer de mama

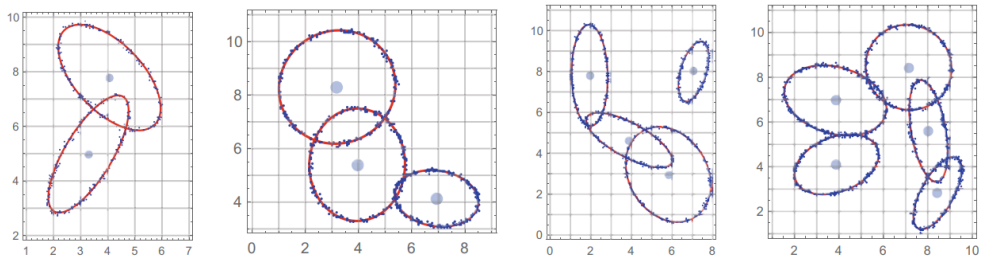
(a) Breast cancer (b) Edge detection (c) Detection



Aplicación a la agronomía para detectar tomates

(a) Tomatoes (b) Edge detection (c) Detection

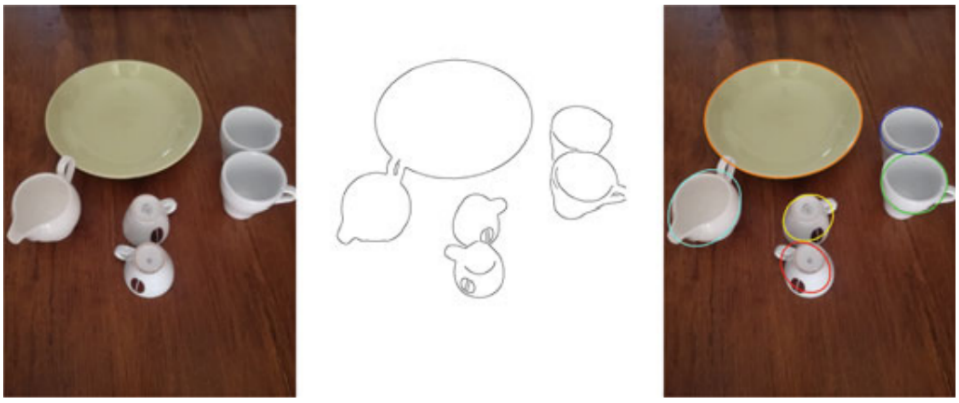
2.2.2. Detección de elipses



Dentro de la aplicación de detectar círculos se encuentra la modificación más coherente que es detectar elipses; ya que una elipse es lo más próximo a un círculo y lo más esperable en lo que se refiere a detectar figuras circulares. Los detalles matemáticos pueden verse en la sección 3.1.4 del capítulo 3 o en el capítulo 8 de [2]. Nuevamente, las imágenes de este apartado son de *"Cluster Analysis and Applications"*[2]. Estas técnicas son similares a la de detección de círculos y permiten detectar una multitud de objetos, sean estos circulares o no. A continuación listaré algunos ejemplos a modo ilustrativo para esta tesis, pero se pueden encontrar más en el capítulo 8 de [2].

Detección de objetos domésticos

(a) Cups (b) Edge detection (c) Detection



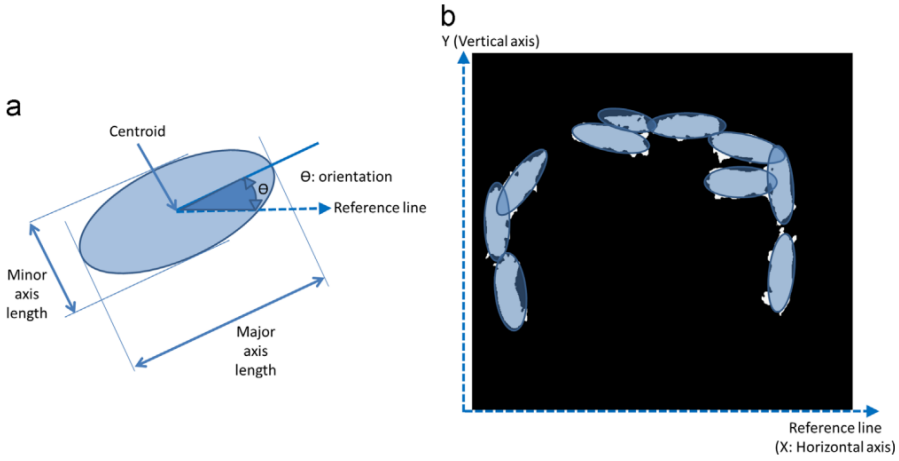
Aplicación a medicina para la detección de la cabeza de un bebé

(a) Head (b) Edge detection (c) Detection

La imagen además de estar en [2] pertenece al siguiente artículo [5].

Aplicación a ganadería para la detección de la posición de cerdos

Otra aplicación es en la ganadería [4], se utiliza para la detección de la ubicación de los cerdos, así como sus dimensiones. La imagen fue retirada de ese mismo estudio.



2.2.3. Detección de líneas

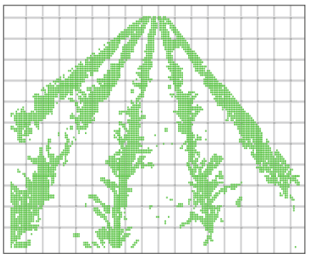
Siguiendo con las aplicaciones geométricas, hay una variación para la detección de líneas. La base matemática es muy semejante a la detección de círculos y elipses. Esta puede encontrarse en la sección 8.1.7 de [2].

Aplicación a agronomía

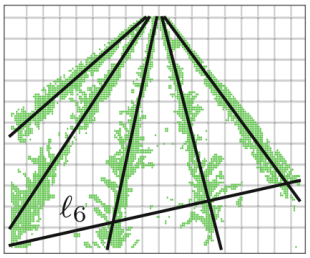
Se puede utilizar la detección de líneas para la detección de cultivos:

a) Maize field

b) Prepared data



c) Recognized 6 lines



En la imagen del ejemplo es importante aclarar que el algoritmo no es perfecto y puede tener errores, como ocurre con la sexta línea. La imagen fue retirada de [2].

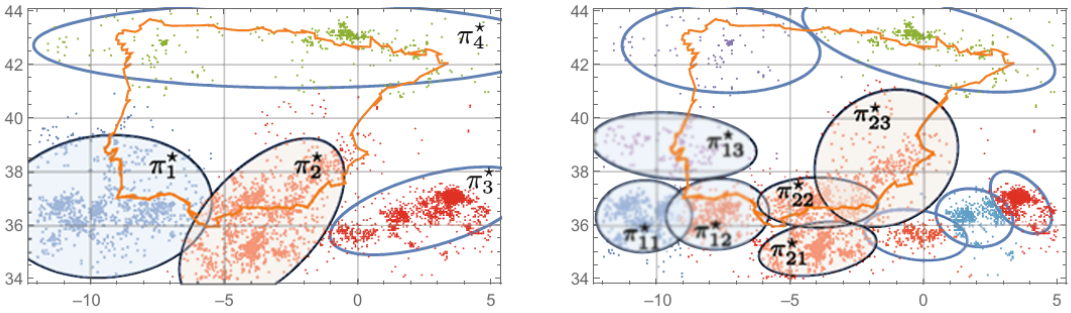
2.2.4. Conclusión

En general se pueden adaptar los algoritmos ya conocidos para reconocer diversas formas geométricas, desde las ya vistas: círculos, elipses, líneas, hasta incluso no vistas: cuadrados, triángulos, óvalos, etc. Podemos concluir que

el clustering se puede utilizar para la detección de la presencia de ciertos objetos geométricos en la imagen, ya sea porque realmente están presentes, o porque el objeto se asemeja; así como es de interés encontrar objetos en las imágenes utilizando una forma que se les parezca, como el caso de las tazas que se detectaron con elipses. Es importante notar que en todos estos casos la razón por la cual se decide agrupar los datos es de índole geométrica. Sobre este asunto volveremos a tratar en el capítulo conclusivo de la tesis. Por lo tanto, los datos son agrupados en virtud de la figura que generan, y este es el criterio para su agrupación, a diferencia de las funciones de clustering estudiadas por Ackerman y Kleinberg, que trataremos con más detalles en sus respectivos capítulos 4 y 5 de la segunda parte.

2.3. Detección de zonas sísmicas

Una de las aplicaciones que se ha hecho al clustering es para la detección de zonas sísmicas de un cierto territorio. El objetivo es: dada una cierta región con actividad sísmica y un registro de los lugares y tiempos de las actividades, particionar el territorio en diversas zonas de actividad sísmica. Un ejemplo real se puede ver en el paper: "A fast partitioning algorithm using adaptive Mahalanobis clustering with application to seismic zoning" [20]; proponen una partición de zonas sísmicas del territorio de Croacia. Estas investigaciones, además de servir para proponer nuevas técnicas de clustering (véase [18]), han servido para el desarrollo de detección de zonas sísmicas. A mi entender, ya que no soy un experto en estas aplicaciones del clustering, y que esta investigación busca servir a otro propósito, la detección de zonas sísmicas tiene como objeto detectar zonas que luego sean propicias para hacer predicciones usando técnicas de machine learning como lo hacen en el paper: "Temporal analysis of croatian seismogenic zones to improve earthquake magnitude prediction" [19] utilizando las mismas zonas sísmicas halladas en el paper que cité anteriormente. También se ha realizado lo mismo para la Península Ibérica.

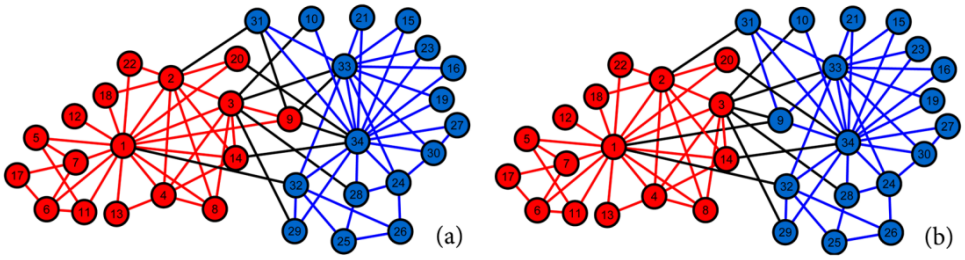


A la izquierda se realizó una división de los datos en 4 clústers. A la derecha se dividió en 11. La imagen fue retirada de [2].

La estrategia empleada para obtener los clústers fue un fraccionamiento territorial basado en una adaptación del algoritmo de k-medias (sección 3.1, especialmente 3.1.4). Por lo que se buscan las zonas, en este caso de formato elíptico, que la distancia entre los datos y el centro de las elipses sea mínima. Notemos aquí que la razón para agrupar los datos es geométrica, porque se busca agrupar los datos en razón de la figura geométrica generada por ellos.

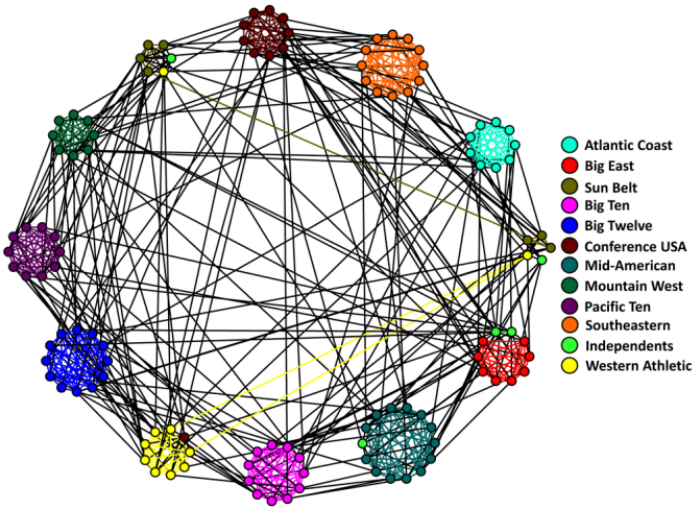
2.4. Detección de comunidades

Otra de las aplicaciones que se puede dar al clustering es la detección de comunidades en grafos. En un grafo se pueden particionar los vértices formando clústers. Entre distintos vértices se puede crear una noción de similitud, y se busca particionar el grafo en diversos clústers que maximicen la similitud intraclúster y minimicen la extraclúster. En este tipo de problemas, a diferencia de las aplicaciones en política o detección de imagen, no particiona en razón de propiedades geométricas, sino por sus similitudes. Un ejemplo es el presentado por el siguiente artículo [21] ***del club de Karate de Zachary***. Los investigadores aplicaron sus métodos de clustering a la red social del club de Karate de Zachary. Los nodos en dicha red representan personas y los links la amistad dentro de la red social. Dicho club de Karate en un determinado momento tiene un problema interno que termina fragmentándolo en dos; causado por el presidente del club (nodo 34) y el instructor (nodo 1). A la izquierda se observa cómo se dividió el club realmente y a la derecha la detección de comunidades realizada por los investigadores.



La red de amistades del club de Karate de Zachary. **(a)** Las comunidades que se formaron luego de la división. **(b)** Las comunidades detectadas por el método de clustering. La imagen es retirada de [21].

Otra aplicación que presenta el mismo artículo es una aplicada sobre una red real más compleja. La red representa el calendario de juegos de fútbol americano entre equipos de la división IA de colegios durante la temporada regular de otoño del año 2000. Los nodos representan a los equipos y los links juegos regulares entre los dos equipos. Al igual que el ejemplo anterior, aquí se compara la realidad con la predicción del algoritmo. Los investigadores ya conocen las comunidades reales y aplican su método propuesto en el artículo a esta red.



Los colores representan las comunidades reales y las comunidades obtenidas por el algoritmo son representadas geométricamente. La imagen es retirada de [21].

2.5. Conclusión

Estas son algunas de las aplicaciones en que se pueden utilizar técnicas de clustering. Como conclusión próxima de lo visto podemos destacar que en todos los casos, diversos, el investigador tiene un *criterio* por el cual va a agrupar los datos. Me explico: en el caso de detección de imágenes ya se sabe de antemano que los clústers tendrán una forma geométrica particular; es decir, la clasificación se realizará en virtud de la geometría de los datos. Para el caso de detección de zonas sísmicas, se sabe que las zonas tienen que ser verdaderas zonas, que tengan área y si es posible que la forma de ellas sea convexa; nuevamente aquí la razón es por consideraciones geométricas. En el caso de detección de comunidades se asume que en un clúster habrá una alta actividad (o similitud) entre vértices y la actividad (o similitud) entre vértices de distintos clústers será baja; aquí el criterio no es geométrico sino por similitud. Esto es de suma importancia, porque si bien los diversos autores pueden utilizar el mismo nombre de "clustering" podemos notar que las razones por las cuales se toma una decisión son distintas; incluso en el capítulo siguiente en la sección 3.5 veremos otro criterio. Esta distinción de motivos es la piedra angular de la confusión en los diversos intentos de axiomatizar una teoría general del clustering (sobre esto, será tratado en la segunda parte y en la tercera), porque se utiliza el mismo nombre para cosas distintas. Todos estos casos tienen en común que son no supervisados: esto es, sin un dataset de entrenamiento previo, pero todos tienen también en común que el investigador ya sabe de antemano algo sobre la clasificación: en los primeros ejemplos que los datos forman una figura geométrica; en los sismos que la clasificación debe corresponder a una región del plano; en la detección de comunidades que hay bastantes enlaces intraclústers y pocos extraclústers. Todo esto será explicado con más detalle en la última parte. Ahora pasaremos a ver los algoritmos y métodos de clustering más comunes antes de pasar a ver los desarrollos de una teoría general del clustering.

Capítulo 3

Algoritmos más importantes del clustering

Conocidas las principales aplicaciones reales del clustering en la tecnología, se revisarán en este capítulo los algoritmos más utilizados e importantes y algunas de sus propiedades relevantes.

Dentro de los métodos de clustering que existen, se destacan los siguientes:

1. K-medias y métodos basados en centros.
2. Árboles de decisión.
3. Clustering jerárquico.
4. Clustering espectral.
5. Clustering basado en modelos probabilísticos.

Este listado resume el contenido principal de las principales fuentes revisadas para esta tesis. Si bien hay otros algoritmos y variaciones, en general son variaciones de los principales algoritmos expuestos en este capítulo, por lo que simplemente se detallará lo necesario de cada algoritmo ignorando una parte considerable de sus variaciones, ya que no son de interés para esta investigación.

Este capítulo tiene como fuentes principales los libros de "Handbook of Cluster Analysis" [22], "Cluster Analysis and Applications" [2] y "Foundations of Data Science" [9]. Aunque no se limita solo a los contenidos de esos documentos; en cada caso citaré las fuentes utilizadas.

3.1. Clustering basado en centros

Según Boris Mirkin en el capítulo 3 de [22], los métodos de clustering basados en centros son actualmente los métodos más populares de clustering. Es un método que se volvió conocido a finales de los sesenta luego de que un resultado teórico fue probado por MacQueen [43]. Este método se puede encontrar en prácticamente todos los libros o materiales que hablen del clustering incluidos textos clásicos como Hartigan [44] y Jain y Dubes [45]. Primero se empezará comentando sobre los representativos, luego se pasará a explicar el método de k-medias y las k-particiones y por último una adaptación de Mahalanobis. Todo esto está inspirado en los primeros capítulos de [2] y en su capítulo 6.

Para trabajar con clustering basado en centros uno debe asumir que los clústers que uno busca están basados en centros. Esto viene a significar que los clústers se pueden caracterizar por ciertos centros $c_1, \ldots, c_k$ en los cuales cada dato es asignado al más cercano [9].

3.1.1. Representativos

Cuando se hace estadística y se trabaja con datos, muchas veces es conveniente resumirlos en un solo dato que sea representativo de todos ellos. Por ejemplo, al calcular una media aritmética se está obteniendo un nuevo dato, a partir de los anteriores, que es una especie de resumen de todos los datos. Cuando la varianza es pequeña, se dice que la media es representativa de los datos, cuando la varianza es grande, la media no da mucha información. Otro dato que es posible tomar como representativo, o resumen, es la mediana, que indica el medio de los datos. A partir de ese dato, sumado el rango intercuantílico, o el rango, es posible determinar que tan representativa es la mediana. Incluso a nivel matemático esto ocurre, cuando se tiene una variable aleatoria X, la esperanza de X es una integral que viene a interpretarse como una especie de resumen de la variable aleatoria. Por eso mismo, en esta sección pasaremos a analizar qué es un representativo.

Definición 3.1.1. (**función pseudométrica**)
Una función $d : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}^+$ es una pseudométrica si satisface:

1. $d(x, y) = 0$ si y solamente si $x = y$.
2. $x \mapsto d(x, y)$ es continuo para todo y fijo.
3. $\lim_{||x|| \to +\infty} d(x, y) = +\infty$ para todo y fijo.

El término utilizado aquí para pseudométrica no corresponde al mismo concepto utilizado en otros ámbitos de la matemática, sino simplemente es un nombre escogido para nombrar una función con dichas propiedades.

Observaciones:

1. Todas las métricas l_p con $p \geq 1$ son pseudométricas. En particular una importante es la conocida con el nombre de mínimos cuadrados: $d_{LS}(x, y) = ||x - y||^2$; con $||\ \ ||$ la norma euclídea.

2. En general, una pseudométrica no tiene ni por qué ser simétrica ni satisfacer la desigualdad triangular.

3. El siguiente teorema es bastante ilustrativo de una de las propiedades que más interesa de las pseudométricas:

Teorema 3.1. *Sea* $\mathbf{A} = \{a_i : i = 1, \ldots, m\} \subset \mathbb{R}^n$ *conjunto de datos con pesos* $w_1, \ldots, w_m > 0$. *Sea* d *una pseudométrica y* $F : \mathbb{R}^n \to \mathbb{R}^+$ *definida de la siguiente forma:*

$$F(x) = \sum_{i=1}^{m} w_i d(x, a_i),$$

entonces existe un punto $c^* \in \mathbb{R}^n$ *tal que:*

$$F(c^*) = \min_{x \in \mathbb{R}^n} F(x).$$

Demostración. 1. $F(x) \geq 0$ para todo $x \in \mathbb{R}^n$.

2. Como todo conjunto de reales acotados inferiormente tiene ínfimo, se sigue que existe: $F^* = \text{ínf}_{x \in \mathbb{R}^n} F(x)$.

3. $F(x)$ es continua por la segunda propiedad de las pseudométricas y $\lim_{||x|| \to +\infty} F(x) = +\infty$ por la tercera propiedad de las pseudométricas.

4. Es un ejercicio de cálculo en varias variables probar que de 2 y 3 se sigue que existe un punto c^* donde se da el ínfimo, que es en este caso un mínimo.

□

Con esto es posible dar una definición de lo que es un representativo de un conjunto de datos:

Definición 3.1.2. (Representativos) Sea $d : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}^+$ una pseudométrica y $\mathbf{A} = \{a_i : i = 1, \ldots, m\} \subset \mathbb{R}^n$ conjunto de datos con pesos $w_1, \ldots, w_m > 0$. Un punto $c \in \mathbb{R}^n$ se dice que es representativo de esos datos según la pseudométrica d si cumple:

$$c \in \arg\min_{x \in \mathbb{R}^n} \sum_{i=1}^{m} w_i d(x, a_i). \tag{3.1}$$

Observaciones:

1. El representativo no tiene por qué ser único.
2. A continuación se verán dos de los representativos más comunes: la media y mediana.

Representativos en una dimensión sin pesos

Por simplicidad, se iniciará el análisis de los representativos para el caso de datos que sean unidimensionales y se verá cuáles son los valores representativos para el caso de las siguientes dos pseudométricas en $\mathbb{R}$:

1. $d_{LS}(x, y) = (x - y)^2$.
2. $d_1(x, y) = |x - y|$ métrica euclídea. Se decidió designarla por d_1, porque en más dimensiones esta métrica deja de ser la euclídea y pasa a ser la l_1.

Observación: d_{LS} no es una métrica en $\mathbb{R}$:

Para cada una de estas pseudométricas y un conjunto finito de datos $\mathbf{A}$ se tienen las correspondientes funciones: F_{LS} y F_1 definidas de la siguiente forma:

$$F_{LS}(x) = \sum_{i=1}^{m} (x - a_i)^2. \tag{3.2}$$

$$F_1(x) = \sum_{i=1}^{m} |x - a_i|. \tag{3.3}$$

Teorema 3.2. *(Media como representativo) Dado un conjunto finito de datos $\mathbf{A} \subset \mathbb{R}$ y la pseudométrica d_{LS} entonces hay un único representativo y es la media de los datos.*

Demostración. 1. La derivada segunda de la función $F_{LS}(x)$ es

$$F''_{LS}(x) = 2m > 0$$

luego es una función estrictamente convexa.

2. Por lo tanto, su mínimo lo alcanza en un único punto. De aquí se concluye la unicidad de la solución.

3. Para hallar el mínimo basta derivar:

$$F'_{LS}(x) = \sum_{i=1}^{m} 2(x - a_i) = 0,$$

$$\sum_{i=1}^{m} (x - a_i) = 0,$$

de aquí se sigue, completando la cuenta, que x debe ser la media aritmética.

□

Teorema 3.3. *(Mediana como representativo) Dado un conjunto finito de datos distintos entre sí* $\mathbf{A} \subset \mathbb{R}$ *y la métrica* d_1 *entonces los representativos son la mediana de los datos.*

Demostración. Para probar este teorema se dividirá en dos casos:

1. Cantidad impar de datos: habrá una única mediana.

2. Cantidad par de datos: habrá un intervalo de medianas.

Como la indexación de los datos es convencional, se puede asumir, sin pérdida de generalidad, el siguiente orden: $a_1 < a_2 < \ldots < a_m$. Si $x \in (a_k, a_{k+1})$ se tiene:

$$F_1(x) = \sum_{i=1}^{k} (x - a_i) - \sum_{i=k+1}^{m} (x - a_i) = (2k - m)x - \sum_{i=1}^{k} a_i + \sum_{i=k+1}^{m} a_i. \quad (3.4)$$

Cabe notar que si se conviene $a_0 = -\infty$ y $a_{m+1} = +\infty$ la fórmula sigue siendo válida. Derivando dicha fórmula se obtiene:

$$F'_1(x) = 2k - m. \quad (3.5)$$

Analizando el signo de la función, es negativo si $k < \frac{m}{2}$, positivo si $k > \frac{m}{2}$ y cero cuando $k = \frac{m}{2}$. Supongamos que m impar. En este primer caso en todos los puntos donde la función es derivable la función siempre es, o creciente o decreciente. El único punto de cambio en el crecimiento ocurre en el dato que ocupa la posición media, es decir, la posición: $\llcorner \frac{m}{2} \lrcorner + 1$; por lo tanto, el mínimo es único y se da en: $x = a_{\llcorner \frac{m}{2} \lrcorner + 1} = mediana$. Para el segundo caso ($m$ par) ocurre que todos los $x \in (a_{\frac{m}{2}}, a_{\frac{m}{2}+1})$ tienen derivada nula, por lo que todos son mínimos de esa función. □

Una observación interesante es recordar que la función $F_1(x)$ es la suma de las distancias en $\mathbb{R}$ de x respecto a cada dato, por lo que es muy fácil de convencerse vía una representación gráfica de la verdad de este teorema.

Corolario 3.3.1. *La verdad de las siguientes afirmaciones es directa a partir de los teoremas anteriores:*

1. *La función $F_{LS}(x)$ se puede interpretar como el error cuadrático medio muestral para el representativo x, por lo que la media es el único valor que minimiza el error cuadrático medio muestral de los datos con respecto a ella misma (la media).*

2. *La función $F_1(x)$ se puede interpretar como el costo muestral (lineal) total con respecto a x como representativo. La mediana por lo tanto minimiza este costo.*

Representativos en una dimensión con pesos

Se puede sustituir el análisis anterior para el caso unidimensional con pesos. Este análisis se encuentra en [2]. Aquí simplemente se expondrán los resultados.

Teorema 3.4. *Sea $\mathbf{A} = \{a_1, a_2, \ldots, a_m\}$ un conjunto finito de datos con pesos $w_1, \ldots, w_m > 0$. Entonces, para la función de costo:*

$$F_{LS}(x) = \sum_{i=1}^{m} w_i (x - a_i)^2,$$

existe un único representativo y es:

$$c_{LS} = \frac{1}{W} \sum_{i=1}^{m} w_i a_i \,, \;\; W = \sum_{i=1}^{m} w_i. \tag{3.6}$$

Dicho de otro modo, la media ponderada por los pesos.

Para el siguiente teorema concerniente al caso de la función de costo pesada $F_1(x) = \sum_{i=1}^{m} w_i|x - a_i|$ es necesario definir la mediana ponderada por los pesos:

Definición 3.1.3. (Mediana ponderada) Sea $\mathbf{A} = \{a_1, a_2, \ldots, a_m\}$ un conjunto de datos finitos con pesos $w_1, \ldots, w_m > 0$. Sea $I = \{1, \ldots, m\}$ el conjunto de los índices. Se define el conjunto J como:

$$J = \{\upsilon \in I \ : \sum_{i=1}^{\upsilon} w_i \leq \sum_{i=\upsilon+1}^{m} w_i\}. \tag{3.7}$$

Cuando $J \neq \emptyset$, se denota por $\upsilon_0 = \max J$. Entonces, se define la mediana ponderada en función de los tres posibles casos:

1. Si $J = \emptyset$ entonces el peso w_1 es estrictamente mayor que la suma de todos los restantes. Luego, la mediana ponderada se define como

$$Med_i(w_i, a_i) = a_1.$$

2. Si $J \neq \emptyset$ y $\sum_{i=1}^{\upsilon_0} w_i < \sum_{i=\upsilon_0+1}^{m} w_i$ entonces

$$Med_i(w_i, a_i) = a_{\upsilon_0+1}.$$

3. Si $J \neq \emptyset$ y $\sum_{i=1}^{\upsilon_0} w_i = \sum_{i=\upsilon_0+1}^{m} w_i$ entonces son medianas ponderadas todos los $x \in [a_{\upsilon_0}, a_{\upsilon_0+1}]$.

Ahora es posible enunciar el siguiente teorema:

Teorema 3.5. *Sea* $\mathbf{A} = \{a_1, a_2, \ldots, a_m\}$ *un conjunto de datos finito con pesos* $w_1, \ldots, w_m > 0$. *Entonces, para la función de costo pesada* $F_1(x)$ *se tiene que la mediana ponderada es el mejor representativo.*

Demostración. La prueba completa se encuentra en [2]. □

Observación: Esta teoría de los representativos puede dar explicación para otras medidas conocidas en la literatura estadística, como, por ejemplo, la media armónica o la media geométrica. El análisis de estos casos se encuentra en [2], en la sección 2.1.4, donde explica sobre la divergencia de Bregman; una forma de construir pseudométricas.

Representativos en más de una dimensión y sin pesos

En este apartado se comentará el caso de los representativos en más de una dimensión para las siguientes tres pseudométricas: d_{LS}, d_2 y d_1, en donde d_2 es la métrica euclídea y d_1 es la distancia Manhattan.

$$d_1(x,y) = \sum_{i=1}^{n} |x_i - y_i|.$$

Teorema 3.6. *(Media como representativo de d_{LS}) Sea $\mathbf{A} = \{a_1, a_2, \ldots, a_m\}$ un conjunto de datos en $\mathbb{R}^n$, la pseudométrica $d_{LS}(x,y) = ||x-y||^2$ y F_{LS} la función de costo correspondiente. Entonces, existe un único representativo y es la media aritmética de los datos.*

Demostración.

$$F_{LS}(x) = \sum_{i=1}^{m}\sum_{k=1}^{n}(x^k - a_i^k)^2 \ \textit{con}\ x = (x^1, \ldots, x^n)\ a_i = (a_i^1, \ldots, a_i^n). \quad (3.8)$$

Además se cumple por el Teorema 3.2 que:

$$\sum_{i=1}^{m}(x^k - a_i^k)^2 \geq \sum_{i=1}^{m}(\bar{x}^k - a_i^k)^2. \quad (3.9)$$

Luego, como una suma de términos positivos se minimiza minimizando cada término, se tiene:

$$F_{LS}(x) = \sum_{i=1}^{m}\sum_{k=1}^{n}(x^k - a_i^k)^2 = \sum_{k=1}^{n}\sum_{i=1}^{m}(x^k - a_i^k)^2 \geq \sum_{k=1}^{n}\sum_{i=1}^{m}(\bar{x}^k - a_i^k)^2. \quad (3.10)$$

□

Teorema 3.7. *(Mediana como representativo de d_1) Sea $\mathbf{A} = \{a_1, a_2, \ldots, a_m\}$ un conjunto de datos en $\mathbb{R}^n$, la métrica $d_1(x,y) = \sum_{i=1}^{n} |x_i - y_i|$ con $x = (x_1, \ldots, x_n)$, $y = (y_1, \ldots, y_n)$ vectores en $\mathbb{R}^n$ y F_1 la función de costo correspondiente. Entonces, los representativos son aquellos que en cada coordenada tienen la mediana de los datos en dicha coordenada. En el caso de que dos o más datos tengan el mismo valor en alguna de las coordenadas, se debe tomar la mediana ponderada con el peso igual a la cantidad de datos coincidentes.*

Demostración. Es análoga a la anterior; haciendo las adaptaciones necesarias. □

Observación: Para el caso de la métrica euclídea no hay una fórmula explícita para la solución de este problema, debido a que:

$$F_2(x) = \sum_{i=1}^{m} ||x - a_i|| = \sum_{i=1}^{m} \sqrt{\sum_{k=1}^{n} (x^k - a_i^k)^2}. \tag{3.11}$$

Y por lo tanto no se pueden dar vueltas los sumandos y acotar por algo conocido. No sé si al día de hoy hay o no una fórmula explícita para solucionar este problema. Por lo que, mi afirmación de que no hay solución explícita está enteramente basada en lo dicho por los autores en [2] en la sección 2.2.4. Este representativo existe y no tiene por qué ser único; se le conoce con el nombre de **mediana geométrica**.

Conclusión de los representativos

El objeto de este apartado no fue realizar una exposición detallada sobre los representativos, sino más bien mostrar los principales para avanzar a la siguiente sección sobre clustering basado en centros. En definitiva, ¿qué será un clustering basado en centros? Uno basado en representativos de los datos según una métrica o pseudométrica. Por lo tanto, para cada pseudométrica distinta tendremos un método distinto de realizar un clustering, ya que cambian los representativos. Esto se detallará a continuación. El lector interesado en ver otras técnicas de representativos puede consultar el capítulo octavo de [2]. Allí encontrará otros problemas no abordados aquí, como el caso de que los datos estén en una circunferencia (por ejemplo, un registro de sismos, etc).

3.1.2. K-particiones y K-medias

En esta sección se definirá el método de clustering basado en centros conocido como k-medias:

Definición 3.1.4. (**K-medias**) Dado un conjunto finito de datos $\mathbf{A} \subset \mathbb{R}^n$ y d una pseudométrica en $\mathbb{R}^n$ y k la cantidad de clústers a particionar; se define el criterio de partición en clústers k-medias de la siguiente forma:

1. Sean $\{c_1, \ldots, c_k\} \subset \mathbb{R}^n$ centros de los clústers tales que minimizan la siguiente función:

$$\mathcal{F}(c_1, \ldots, c_k) = \frac{1}{n} \sum_{i=1}^{n} \min_{1 \leq j \leq k} d(c_j, a_i). \tag{3.12}$$

2. Entonces, para cada c_j le corresponde un clúster C_j resultante de asignar a dicho conjunto los datos a_i más cercanos a c_j.

Observación: Este criterio, a priori, no garantiza la existencia de k-clusterings, porque podría ocurrir que alguno de los conjuntos C_j quede vacío. Se puede adaptar este algoritmo para resolver este problema obligando a que los centros sean datos.

No es el objeto de esta tesis estudiar la algorítmica detrás de estos criterios para lograr efectivamente en la práctica obtener dichos centros. Estos métodos algorítmicos se pueden encontrar en [2][9]. Lo importante es entender la geometría utilizada y que, en definitiva, el método se resume en conseguir encontrar los k mejores centros para dicha pseudométrica.

Existe otra forma de abordar el mismo problema:

Definición 3.1.5. (**K-partición**) Sea $\mathbf{A}$ un subconjunto finito de datos con más de un dato. Una partición de $\mathbf{A}$ en k-subconjuntos disjuntos y no vacíos es lo que se conoce como una k-partición de $\mathbf{A}$. A dicha k-partición la denotaremos $\mathbf{C} = \{C_1, \ldots, C_k\}$ y a los elementos C_j se les llamará **clústers** de $\mathbf{A}$.

Ahora bien, es posible hacer el procedimiento reverso al realizado con k-medias:

1. Para cada C_j se obtiene $c_j \in \arg\min_{x\in\mathbb{R}} \sum_{a_i\in C_j} d(x, a_i)$; y esto para todo clúster C_j.

2. Por lo tanto, se puede definir la siguiente función de costo para cada k-partición:
$$\mathcal{F}^*(\mathbf{C}) = \sum_{j=1}^{k} \sum_{a_i\in C_j} d(c_j, a_i). \tag{3.13}$$

3. La k-partición óptima para este método es la k-partición que minimiza esta función de costo; que existe porque la cantidad posible de k-particiones es finita. Es decir, si se designa por $\mathbf{C}^*$ a la k-partición óptima, esta cumple:
$$\mathbf{C}^* \in \arg\min_{\mathbf{C}} \mathcal{F}^*(\mathbf{C}). \tag{3.14}$$

Conclusiones:

1. En el segundo método, para cada clúster C_j el correspondiente centro c_j es el representativo de ese clúster con respecto a la pseudométrica d.

2. En el segundo método, el k-clustering óptimo $\mathbf{C}^*$ corresponde al clustering en el que la media del error entre los datos en un clúster y su correspondiente representativo es mínima; en efecto, como:

$$\mathbf{C}^* \in \arg\min_{\mathbf{C}} \mathcal{F}^*(\mathbf{C}) = \arg\min_{\mathbf{C}} \sum_{j=1}^{k} \sum_{a_i \in C_j} d(c_j, a_i).$$

$$= \arg\min_{\mathbf{C}} \frac{1}{k} \sum_{j=1}^{k} \sum_{a_i \in C_j} d(c_j, a_i). \tag{3.15}$$

3. Una digresión sobre este tema: se puede también definir lo que se conoce como **k-centros** (véase en [9]) de la siguiente forma: el k-clustering resultante de aplicar k-centros es el que cumple el siguiente criterio:

$$\mathbf{C} \in \arg\min_{\mathbf{C}} \left(\max_{j=1}^{k} \max_{a_i \in C_j} d(c_j, a_i) \right). \tag{3.16}$$

   Expliquemos esto para que se entienda: para cada clúster C_j el error consiste en la máxima diferencia dentro de ese clúster, es decir:

$$\max_{a_i \in C_j} d(c_j, a_i).$$

   Y el error total de todo el clustering es el máximo error en dicho clustering. A modo de analogía visual: si se piensa en cada clúster como una circunferencia que contiene datos, el error es el diámetro de la circunferencia, el error total es el máximo diámetro existente; minimizar esto es encontrar los centros que hacen el mayor diámetro el menor posible.

Conclusión: El lector ya puede reconocer la gran flexibilidad que se cuenta a la hora de definir un criterio basado en centros. En esencia, un método basado en centros consiste en lo siguiente: encontrar la k-partición que minimice una función de costo basada en k-centros. De aquí se deduce que se puede generalizar este método para incluso datos que no estén en el espacio euclídeo sino en un espacio general con una pseudométrica general. Esto es lo que se hace en [1] y [12].

3.1.3. K-centros para la distancia LS

Un caso particular de k-medias es el asociado a la pseudométrica:

$$d_{LS}(x, y) = ||x - y||^2.$$

Los centros de esta pseudométrica son centroides.

Para este caso particular, si tenemos una partición $C = \{C_1, \ldots, C_k\}$ se cumple:

$$c_j = \arg\min_{x \in \mathbb{R}^n} \sum_{a \in C_j} ||x - a||^2 = \frac{1}{|C_j|} \sum_{a \in C_j} a. \tag{3.17}$$

En este caso particular cada centroide es la media de los datos en un clúster. Y por el Teorema 3.9 estas son las k-medias óptimas que particionan el espacio en los clústers que minimizan el error. Es posible definir una función de costo asociada a un clustering de los datos:

$$\mathcal{F}_{LS}(C) = \sum_{j=1}^{k} \sum_{a \in C_j} ||c_j - a||^2. \tag{3.18}$$

De esta pseudométrica, un resultado interesante que se puede obtener es el resultado conocido como: **problema dual**. Para cada clustering de los datos se puede definir la siguiente función:

$$\mathcal{G}(C) = \sum_{j=1}^{k} |C_j| ||c_j - c||^2 \quad c = \frac{1}{n} \sum_{i=1}^{n} a_i. \tag{3.19}$$

Cada término de $\mathcal{G}$ viene a ser la distancia pesada del centroide c_j al centro de todos los datos; y este último es independiente de la elección de la partición. Para entender mejor qué significa esta suma es necesario conocer el resultado siguiente:

Teorema 3.8. *(Problema dual de optimización)*
Bajo las hipótesis habituales y las definiciones dadas anteriormente, se cumple para todo clustering C:

$$\sum_{i=1}^{n} ||c - a_i||^2 = \mathcal{F}_{LS}(C) + \mathcal{G}(C). \tag{3.20}$$

Es posible interpretar el lado izquierdo como la varianza muestral (aunque le faltaría dividir por $n-1$). De este teorema se deduce que el aumento de una función disminuye la otra y viceversa. Por lo que, si se pretende minimizar el error ($\mathcal{F}_{LS}$) es equivalente a maximizar la otra función. Por lo que es un corolario lo siguiente:

Corolario 3.8.1. *(Equivalencia de funciones objetivo) Es equivalente dar un clustering que minimice $\mathcal{F}_{LS}$ a dar uno que maximice $\mathcal{G}$.*

Se puede dar un intento de explicación de qué significa esta función $\mathcal{G}$, sin llegar esta a ser exacta o perfecta. Esta función viene a ser una especie de medida de separación de los centros entre sí, o de cuanta separación entre clústers hay. Por eso la necesidad de maximizarla. Veamos la demostración del teorema:

Demostración.

$$\sum_{a_i \in C_j} ||x - a_i||^2 = \sum_{a_i \in C_j} ||(x - c_j) + (c_j - a_i)||^2$$

$$= \sum_{a_i \in C_j} ||x - c_j||^2 + 2 \sum_{a_i \in C_j} < x - c_j, c_j - a_i > + \sum_{a_i \in C_j} ||c_j - a_i||^2$$

$$= \sum_{a_i \in C_j} ||x - c_j||^2 + \sum_{a_i \in C_j} ||c_j - a_i||^2$$

$$= |C_j|||x - c_j||^2 + \sum_{a_i \in C_j} ||c_j - a_i||^2.$$

Poniendo c en lugar de x y sumando se cumple la tesis. Para pasar del segundo renglón al tercero basta notar lo siguiente:

$$2 \sum_{a_i \in C_j} < x - c_j, c_j - a_i >= 2 < x - c_j, \sum_{a_i \in C_j} (c_j - a_i) > = 0.$$

Ya que: $\sum_{a_i \in C_j} (c_j - a_i) = 0$ por ser c_j la media de los datos en ese clúster. □

Conclusión: si interpretamos la función $\mathcal{F}$ como la dispersión intracluster y a $\mathcal{G}$ como la separación entre clústers; en el caso de la pseudométrica LS, sucede que el mínimo de la dispersión ocurre a la misma vez del máximo de la separación entre clústers.

3.1.4. Mahalanobis data clustering

Hasta ahora se ha revisado lo que se podría llamar: teoría clásica del k-clustering basado en centros. Ocurre que hay problemas de la vida real, como los vistos en el capítulo anterior, que se desea agrupar los datos en elipsoides. La propia naturaleza del problema muchas veces indica que la forma de los clústers debe ser un elipsoide. Véase el capítulo anterior. Para esto definiremos la pseudométrica denominada Mahalanobis.

Mahalanobis en el plano

La elipse $E(O, a, b, \theta)$ de centro en el origen, semiradios a,b y rotada un ángulo θ es el lugar geométrico del plano que cumple esta ecuación cartesiana [2]:

$$\frac{(x\cos\theta + y\sin\theta)^2}{a^2} + \frac{(-x\sin\theta + y\cos\theta)^2}{b^2} = 1. \tag{3.21}$$

Esto es una consecuencia directa de aplicar álgebra lineal. Es un resultado típico de cursos de álgebra lineal que la imagen de una circunferencia por una transformación lineal simétrica y definida positiva es una elipse. Esta ecuación se puede obtener de la siguiente forma:

sea Σ la siguiente matriz:

$$\Sigma = U \begin{bmatrix} \frac{1}{a^2} & 0 \\ 0 & \frac{1}{b^2} \end{bmatrix} U^T \quad U = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}. \tag{3.22}$$

Por lo que la ecuación 3.21 se obtiene de la siguiente:

$$[x, y]\Sigma[x, y]^T = 1. \tag{3.23}$$

Como puede verse, la matriz Σ es una matriz simétrica y definida positiva, por lo que las soluciones a dicha ecuación son elipses con las condiciones descritas al comienzo.

Para definir la pseudométrica Mahalanobis se utiliza el siguiente resultado propio del álgebra lineal:

Lemma 3.9. *Existe una biyección entre el conjunto de todas las elipses del plano centradas en el origen y el conjunto de todas las matrices cuadradas* 2×2 *simétricas y definidas positivas.*

Demostración. Ver la sección 2.2.2 del capítulo 2 de [23]. □

Ahora es posible definir la pseudométrica Mahalanobis. Esta consiste en la siguiente idea: dada una elipse $E(O, a, b, \theta)$ se quiere que todos los puntos en ella tengan la misma distancia con el origen. Para lograr esto basta con definir la pseudométrica de la siguiente forma:

Definición 3.1.6. (**Pseudométrica de Mahalanobis**) Se define la distancia de Mahalanobis para $\mathbb{R}^2$ asociada a la elipse: $E(O, a, b, \theta)$ de la siguiente forma:

$$d_m(x, y; \Sigma) = (x - y)\Sigma(x - y)^T. \tag{3.24}$$

Con Σ la matriz asociada a dicha elipse.

Cuando se tiene un conjunto de datos $\mathbf{A} = \{a_i = (a_i^1, \ldots, a_i^n) \ : \ i = 1, \ldots, m\} \subset \mathbb{R}^n$ con ciertos pesos $w_1, \ldots, w_m > 0$ se tiene siempre asociada una matriz cuadrada, conocida en la literatura estadística como: matriz de covarianza muestral. Sea:

$$\bar{a} = (\bar{a}_1, \ldots, \bar{a}_n) = \frac{1}{W} \sum_{i=1}^{m} w_i a_i \ , \ \ W = \sum_{i=1}^{m} w_i \ ,$$

$$\Sigma = \begin{bmatrix} \Sigma w_i(\bar{a}_1 - a_i^1)^2 & \Sigma w_i(\bar{a}_1 - a_i^1)(\bar{a}_2 - a_i^2) & \cdots & \Sigma w_i(\bar{a}_1 - a_i^1)(\bar{a}_n - a_i^n) \\ \Sigma w_i(\bar{a}_2 - a_i^2)(\bar{a}_1 - a_i^1) & \Sigma w_i(\bar{a}_2 - a_i^2)^2 & \cdots & \Sigma w_i(\bar{a}_2 - a_i^2)(\bar{a}_n - a_i^n) \\ \vdots & \vdots & \ddots & \vdots \\ \Sigma w_i(\bar{a}_n - a_i^n)(\bar{a}_1 - a_i^1) & \Sigma w_i(\bar{a}_n - a_i^n)(\bar{a}_2 - a_i^2) & \cdots & \Sigma w_i(\bar{a}_n - a_i^n)^2 \end{bmatrix}.$$

El símbolo Σ de la izquierda de la igualdad representa a la matriz; los de la derecha son sumatorias

Designando a la matriz de covarianza muestral con el signo Σ, **se define, entonces, la pseudométrica de Mahalanobis** asociada a ese conjunto de datos como:

$$d_m(x, y; \Sigma^{-1}) = (x - y)\Sigma^{-1}(x - y)^T. \tag{3.25}$$

Para este caso se utiliza la matriz inversa, ya que de esta forma los autovalores quedan inversos y por lo tanto el elipsoide asociado a esta ecuación: $d_m(x, 0; \Sigma^{-1})$ es el que tiene los radios del tamaño de los autovalores de Σ. A partir de esto, se puede definir naturalmente la función de costo. Dado un clustering $C_1, \ldots, C_k$ de $\mathbf{A}$, para cada C_i se tiene asociada una matriz de covarianza de los datos y, por lo tanto, una pseudométrica de Mahalanobis $d_m^{(i)}$; a su vez dicha pseudométrica permite tener un representativo asociado a ella: c_i; luego, el costo, o error, total asociado a dicha partición es:

$$\mathcal{F}(C) = \sum_{i=1}^{k} \sum_{a \in C_i} d_m^{(i)}(a, c_i; \Sigma_i^{-1}). \tag{3.26}$$

Por lo tanto, la mejor k-partición será aquella que minimice esta función de costo.

Conclusión: Esta pseudométrica es la utilizada para la detección de elipsoides en las imágenes vistas en el capítulo anterior en la sección 2.2.2 de detección de elipses. De la misma forma que se utiliza una pseudométrica ajustada para la detección de cierta figura geométrica (como en este caso elipsoides) se puede ajustar para otras: como círculos y líneas. El método en concreto de cómo ajustarlo para el caso de círculos y líneas se puede encontrar en: [2].

3.1.5. K-medias y probabilidad

El objeto de esta parte es mostrar cómo la teoría de k-medias se puede combinar con herramientas probabilísticas. Para esto se supondrá lo siguiente:

Hipótesis necesaria: Los datos son la realización de una sucesión de variables aleatorias independientes e idénticamente distribuidas que toman valores en $\mathbb{R}^n$, tal que:

$$\mathbb{E}(||X||^2) < \infty.$$

Observación: siempre que se quiere combinar estadística matemática con la teoría de la probabilidad es necesario asumir que los datos son el resultado de una variable aleatoria, ya que esta última es el principio de todo análisis probabilístico.

Definición 3.1.7. (**distribuciones poblacional y muestral**) La distribución de la variable aleatoria X se denotará como μ y a la distribución empírica como μ_n. Se recuerda que la distribución μ es la medida de probabilidad asociada a la variable aleatoria X, esto es:

$$\mathbb{P}(X \in A) = \mu(A). \tag{3.27}$$

Y la distribución empírica es otra variable aleatoria que cuenta el promedio de datos que caen en un conjunto:

$$\mu_n(A) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_A(X_i). \tag{3.28}$$

En donde $\mathbf{1}_A$ es la función indicatriz y $X_1, \ldots, X_n$ forman una sucesión de datos i.i.d. con la distribución poblacional μ.

Dicho esto, es posible definir la función de costo que tanto caracteriza este método:

Definición 3.1.8. (**función de costo poblacional**) Dado un conjunto de centros $c = (c_1, \ldots, c_k)$, se define el error poblacional de X para esos centros como:

$$W(c, \mu) = \mathbb{E}_\mu(\min_{1 \leq j \leq k} ||X - c_j||^2). \tag{3.29}$$

Y el riesgo óptimo:

$$W^*(\mu) = \inf_c W(c, \mu). \tag{3.30}$$

También se definen las versiones muestrales, aunque serán variables aleatorias y no números:

Definición 3.1.9. (**función de costo muestral**) Dado un conjunto de centros $c = (c_1, \ldots, c_k)$, el error muestral de una muestra aleatoria simple $X_1, \ldots, X_n$ es:

$$W(c, \mu_n) = \frac{1}{n} \sum_{i=1}^{n} \min_{1 \leq j \leq k} ||X_i - c_j||^2. \tag{3.31}$$

Equivalentemente, se define el mínimo error muestral posible (recordar que sigue siendo una variable aleatoria) como:

$$W^*(\mu_n) = \inf_c W(c, \mu_n). \tag{3.32}$$

Definido esto, se expondrá el teorema de consistencia para este método, el cual no se demostrará, porque no es el objeto de esta tesis profundizar en estos métodos.

Definición 3.1.10. (δ_n-**minimización**) Sea $\delta_n \geq 0$; decimos que un conjunto de centros $c_n = (c_{n,1}, \ldots, c_{n,k})$ es un δ_n-minimizador del error si:

1. $W(c_n, \mu_n) \leq W^*(\mu_n) + \delta_n$.

Teorema 3.10. *(Consistencia de k-medias) Sea X una variable aleatoria con segundo momento finito y sea c_n una sucesión de δ_n-minimizadores tales que $\lim_{n\to\infty} \delta_n = 0$; entonces se cumple:*

1. $\lim_{n\to+\infty} W(c_n, \mu) = W^*(\mu)$ *casi seguramente.*
2. $lim_{n\to+\infty} \mathbb{E}(W(c_n, \mu)) = W^*(\mu)$.

Este resultado muestra lo siguiente:

1. $W^*(\mu_n)$ viene a ser el menor error que es posible conseguir al realizar un k-clustering con los datos que ya tengo; luego, un δ_n-minimizador es un conjunto de centros que están a menos de δ_n de este óptimo.

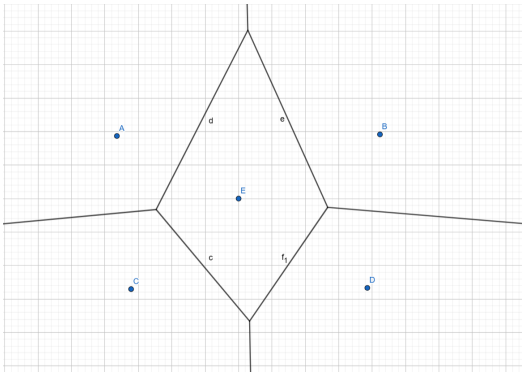
2. Por lo tanto, este resultado afirma que conforme se aumente la cantidad de datos (n es cada vez mayor) y se obtengan centroides cada vez más cercanos al óptimo muestral, entonces, casi seguramente, se acerque el error al óptimo poblacional; que viene a ser lo mejor posible en términos de error.

3. El óptimo muestral converge al óptimo poblacional, pero este teorema garantiza que si se mantiene cada vez más cerca de ese óptimo muestral entonces también se estará cerca del poblacional.

Para las demostraciones y para más información revisar los siguientes artículos: [29, 30, 31, 32].

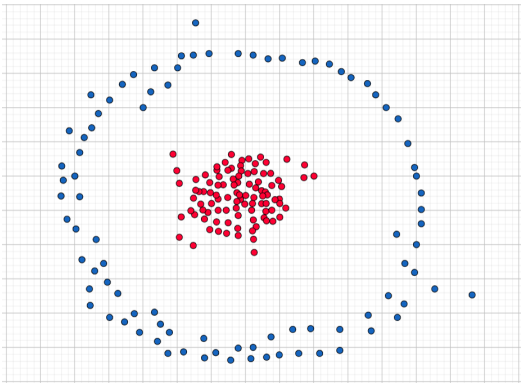
3.1.6. Geometría de k-medias

Un aspecto de estos algoritmos que es importante tener en consideración es su poca flexibilidad para separar en clústers ciertos grupos de datos. Esto ocurre porque se realiza la separación en base a ser más cercano a un centro. Los centros, por lo tanto, dividen al espacio en regiones de aspecto poligonal. Cada región se caracteriza por ser el conjunto de puntos más cercanos a ese centro. Véase la siguiente figura para mayor claridad:



Los puntos azules representan los centros. Las regiones son el lugar geométrico de los puntos más cercanos a cada centro. Se está utilizando la métrica usual.

Por esta razón, el algoritmo de k-medias nunca será capaz de clasificar, como sería deseado, el siguiente conjunto de datos:



Los colores corresponden al clúster que uno desearía obtener de un algoritmo. En caso este documento esté impreso en blanco y negro, los datos centrales corresponderían a un clúster y los restantes a otro.

3.2. Clústers jerárquicos

La presente sección se basa en el capítulo 6 del libro "Handbook of Cluster Analysis" [22]. Dentro de los métodos estadísticos para agrupar datos que no presuponen que provengan de una cierta distribución, los métodos jerárquicos son los que a partir de una cierta noción de similitud o disimilitud, ordenan los datos (de ahí nace la jerarquía) para luego posteriormente clusterizarlos (esto se desarrolla con más profundidad en el capítulo 6). Una noción de disimilitud es casi una distancia d entre pares de puntos, solamente que no tiene por qué cumplir la desigualdad triangular; una noción de similitud es el inverso de una función de disimilitud $\frac{1}{d}$ solo que además puede valer cero. Por lo tanto, para realizar un clustering jerárquico es necesario o bien una **función de similitud** o bien una de **disimilitud**. Se reducirá el análisis contemplando exclusivamente métodos de disimilitud ya que, en general, ambos métodos son equivalentes.

Observación: Toda medida de disimilitud establece un orden entre los datos y permite naturalmente agruparlos en función de su semejanza o desagruparlos en función de su desemejanza. Se le llama jerárquico porque se utiliza este orden para realizar la agrupación en clústers.

Definición 3.2.1. (Métodos aglomerativos y disociativos) Se dice que un método jerárquico es aglomerativo cuando se principia por los datos y sus semejanzas, o desemejanzas, para luego agruparlos en clústers. Un método es disociativo si se parte de un único clúster (el conjunto de los datos) y se empieza a dividir en clústers en función de la semejanza y desemejanza. Estos

dos métodos no son equivalentes; eso se verá en el capítulo 6 en las secciones 6.4.2 y 6.4.3.

3.2.1. Métodos aglomerativos

Para entender estos métodos y verlos aplicados se comenzará la exposición por los aglomerativos. Se denotará d como una función de disimilitud de los datos.

1. **Single linkage**: Para aplicarlo es necesario definir la disimilitud entre dos conjuntos finitos de datos C_1, C_2:

$$d(C_1, C_2) = \min_{x \in C_1 \; y \in C_2} d(x, y). \tag{3.33}$$

   Con esto es posible definir el método de la siguiente forma:

   a) Se empieza con los datos siendo cada uno un clúster.

   b) Luego se agrupan los dos clústers que tengan menor disimilitud.

   c) Se repite este proceso hasta que algún criterio de parada se cumpla. Existen varios criterios, algunos de estos son:

      1) Se alcanzan k clústers.
      2) Si $r > 0$ es una distancia prefijada, entonces se aglomeran hasta que todo par de clústers tenga una distancia entre sí superior a ese valor.
      3) Si $0 < \rho < 1$ es un porcentaje, se aplica el criterio anterior con el siguiente número: $r = \rho \max_{x,y} d(x, y)$.

2. **Complete linkage:** Las siguientes variaciones siguen el mismo esquema pero cambian la noción de distancia entre conjuntos finitos. En este caso es:

$$d(C_1, C_2) = \max_{x \in C_1 \; y \in C_2} d(x, y). \tag{3.34}$$

   Es decir, la disimilitud de un conjunto está dada por la máxima disimilitud encontrable al comparar las disimilitudes entre datos de uno y otro. En el caso anterior era opuesto, ya que la disimilitud entre dos conjuntos era su distancia; aquí es como el radio de la unión.

3. **Average linkage:** En este caso la distancia queda:

$$d(C_1, C_2) = \frac{\sum_{x \in C_1} \sum_{y \in C_2} d(x, y)}{|C_1||C_2|}. \tag{3.35}$$

aquí se mide el promedio de disimilitudes.

Como se pueden observar son esencialmente lo mismo en su estructura algorítmica, variando en la medición de la disimilitud entre clústers. Se puede notar que es posible crear infinitas variaciones; correspondiendo al investigador utilizar la más conveniente al problema.

3.2.2. Métodos disociativos

Los métodos disociativos, en general, son bastantes distintos a los aglomerativos. Esta diferencia se verá con más detalle en el capítulo 6, secciones 6.4.2 y 6.4.3. La principal diferencia radica en que los métodos disociativos son sensibles al contexto; esto significa que toman su decisión de partir los datos en base a la totalidad de ellos, mientras que los métodos aglomerativos son locales (ver sección 6.2.1), que viene a ser lo opuesto a la sensibilidad al contexto. En general siguen la siguiente estructura:

1. Empezamos con un único clúster conteniendo todos los datos.
2. Para cada clúster existente analizamos la división posible que maximice la diferencia entre los clústers. Una vez analizado esto para todos los clústers se inicia el proceso de división por el que tiene mayor diferencia, hasta o bien dividir todos los clústers existentes, o bien que se cumpla un criterio de parada.
3. Repetir el punto anterior hasta cumplir con algún criterio de parada.

Observación: Los criterios de parada se pueden definir análogamente a los tres vistos anteriormente:

- Para el caso de k clústers el criterio permanece intacto.
- Para el caso de un valor r que sirve para agrupar, podemos usarlo para separar, dividiendo el clúster si presenta una división que maximice la disimilitud y esta a su vez supere este valor.
- Idéntico para el caso de proporción.

3.2.3. Geometría del single-linkage

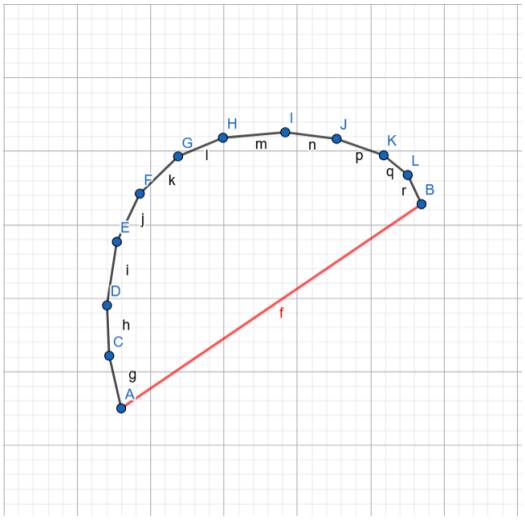
A diferencia de los algoritmos de k-medias y basados en centros, estos algoritmos son mucho más flexibles en la clasificación de los datos que están agrupados en figuras geométricas más complicadas. Estos algoritmos pueden clasificar agrupaciones de datos en sus más diversas figuras. Vuélvase a ver la subsección 3.1.6 sobre la geometría de k-medias. Allí se presenta un ejemplo que el algoritmo de k-medias no puede clasificar como se desearía. En cambio, single linkage, y otros linkage-based, sí son capaces; porque los clústers se generan uniendo puntos cercanos (especialmente en el caso single-linkage). Luego, no dependen de la figura geométrica como k-medias depende, porque este para dividir los datos utiliza el criterio de puntos más cercano a un conjunto de centros.

3.3. Métodos basados en procesos espaciales

3.3.1. algoritmo generalizado de Single-Linkage

Cuando se busca flexibilidad en la figura del clúster, en la identificación de clústers en un conjunto de datos en el espacio, single-linkage parece ser una buena opción. Como se ha visto en las dos secciones 3.2.3 y 3.1.6, el algoritmo de k-medias es poco flexible con las figuras de los clústers, mientras que single-linkage puede fácilmente identificar como clúster una larga cadena de datos. Ahora bien, cuando se busca identificar diversos clústers que varíen en su densidad, esto es, que haya ciertos clústers donde sus puntos están muy agrupados y a la vez otros donde sus datos sean más dispersos; se hace difícil utilizar los algoritmos de single-linkage y sus variantes. Esto también se debe a que se utiliza un único parámetro como condición de parada. Para contrarrestar esto se creó una versión generalizada del single-linkage para datos espaciales (Para más información, ver capítulo 14, secciones 14.1 y 14.2 de [22]).

Definición 3.3.1. (**Árbol mínimo generado**) Sea $\phi \subset \mathbb{R}^n$ localmente finito (ningún punto de ϕ es de acumulación de ϕ) sin que haya dos pares de puntos con idéntica distancia entre ellos. El árbol mínimo generado por estos puntos es el grafo que tiene por nodos a los mismos puntos, y dos nodos x, y están conectados si y solamente si no existe un número natural positivo n y una secuencia de nodos $x = x_1, x_2, \ldots, x_n = y$ tales que $||x_i - x_{i+1}|| < ||x - y||$ para todo $i = 1, \ldots, n$.



Notar que entre los puntos A y B no podría haber una conexión, porque existe un camino pasando por otros puntos de tal forma que entre un nodo y el sucesor la distancia es menor a la distancia entre A y B.

Observación: No es difícil de verificar que efectivamente es un árbol. Si hubiese un ciclo, supongamos $x_0, x_1, x_2, \ldots, x_{n-1}, x_n = x_0$ con $\{x_i, x_{i+1}\} \in E_\phi$ (E_ϕ son los links del grafo ϕ), entonces habría un par de puntos x_k, x_{k+1} con distancia máxima; luego por la definición de árbol mínimo generado, no podrían estar conectados.

Definición 3.3.2. (**Regla de corte**) Sea $\mathcal{G}^*$ el conjunto de todos los grafos localmente finitos y con raíz en $\mathbb{R}^n$ ("rooted", en inglés). Una regla de corte es una función:

$$f : \mathcal{G}^* \times \mathcal{G}^* \times [0, +\infty) \to \{0, 1\}, \tag{3.36}$$

que satisface:

$$f(g_1, g_2, x) = f(g_2, g_1, x). \tag{3.37}$$

La idea intuitiva de la regla de corte es la siguiente: tenemos un grafo mínimo generado ϕ con links E_ϕ. Como es un árbol, si elimino conexiones entre nodos el resultado es un grafo disconexo. La regla de corte sirve para decidir cuándo desconectar dos nodos. En definitiva, se poda el árbol y el resultado final son componentes conexas que serán los clústers.

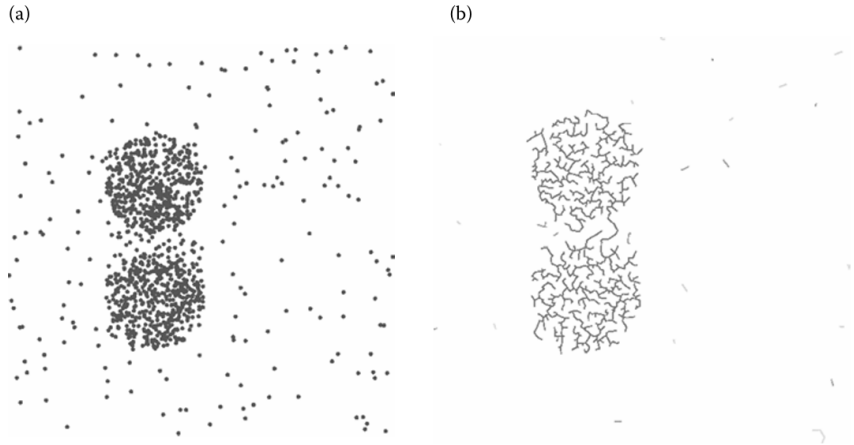
Definición 3.3.3. (**Single-linkage generalizado**) Sea $\phi \subset \mathbb{R}^n$ finito y f una regla de corte. A su vez sea también $T = (\phi, E_\phi)$ el árbol mínimo generado por ϕ. Entonces los clústers generados por la regla de corte son las componentes conexas del siguiente grafo:

$$(\phi, \{\{x_1, x_2\} \in E_\phi \ : \ f(T_{x_1}, T_{x_2}, ||x_1 - x_2||) = 1\}). \tag{3.38}$$

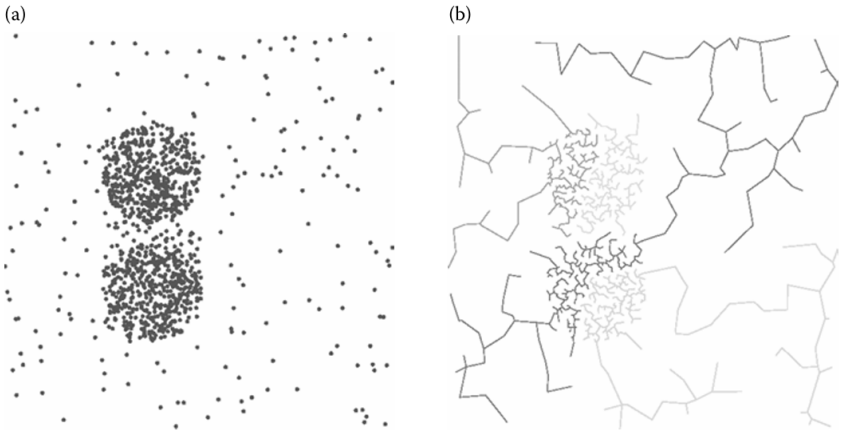
Aquí T_{x_i} significa el árbol que resulta de eliminar la conexión $\{x_1, x_2\}$ con x_i como raíz.

Veamos algunos ejemplos:

1. $f(g_1, g_2, l) = \mathbf{1}_{\{x:x\leq\alpha\}}(l)$ (**Single-linkage**).
2. $f(g_1, g_2, l) = 1 - \mathbf{1}_{\{x:x\geq n\}}(|g_{1,l}|)\mathbf{1}_{\{x:x\geq n\}}(|g_{2,l}|)$ donde n es un entero mayor a uno y $g_{i,l}$ es la componente conexa de g_i que contiene al nodo raíz y que sólo contiene los nodos que no distan más de l.



Clustering inducido por la primera regla de corte. Imagen retirada de [22].

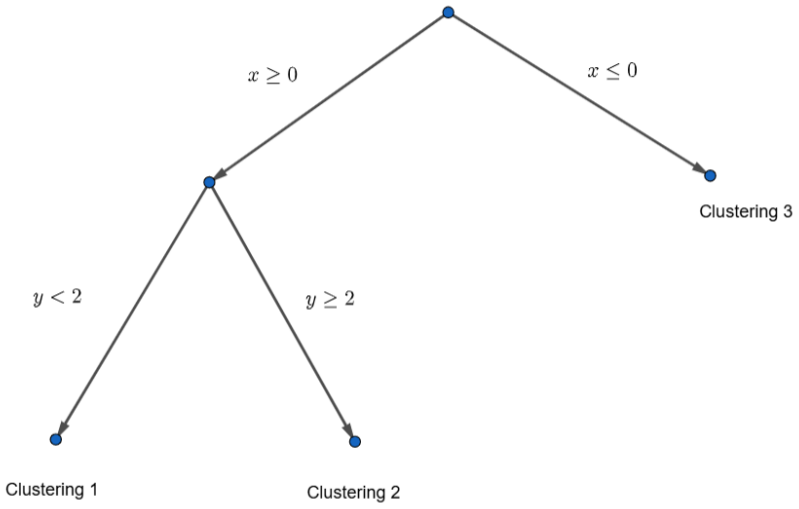


Clustering inducido por la segunda regla de corte. Imagen retirada de [22].

Sobre el último ejemplo, se puede encontrar una explicación más detallada en el capítulo 14, secciones 14.1 y 14.2 de [22].

3.3.2. Árboles de clasificación

Otro método de clasificación no supervisada, comentado por Ricardo Fraiman, Badih Ghattas y Marcela Svarc en [46], es crear un árbol de decisiones. Un ejemplo de decisión es, para datos que tienen dos coordenadas numéricas (x, y), ser $x \geq 0$ o $x < 0$. Cuando se tienen datos en un espacio euclídeo hay varias formas de generar un árbol de decisiones. En general, las decisiones se toman sobre las coordenadas de los datos, como en el ejemplo anterior y en la siguiente imagen:



¿Qué decide en cuál número se hace el corte? ¿Por qué, en la imagen anterior, cortar en cero y luego en dos, y no en tres y luego en cinco? Se requiere, por lo tanto, un criterio. Este criterio es la **medida de impureza** de un nodo del árbol. La estrategia es análoga al método disociativo visto en el caso de clustering jerárquico. La idea es que un nodo es "puro" si contiene un único clúster, luego es "impuro" si hay mixtura de clústers. Cuán impuro es un nodo es lo que mide el indicador de impureza. En términos matemáticos, un indicador de impureza, idealmente, es una función i que evalúa un vértice de decisión del árbol de decisiones y devuelve:

1. $i(v) = 0$ si todos los datos que llegan hasta el nodo v son de la misma categoría.

2. $i(v)$ es máximo si todas las categorías están igualmente presentes.

Una vez establecida una función que tenga características análogas a estas se puede establecer un criterio de cómo partir el nodo en dos decisiones. Para

un vértice v, se denotarán v_l y v_r a los dos subnodos correspondientes a las dos decisiones contradictorias posibles. La diferencia de impureza se define:

$$\Delta(v, v_l, v_r) = i(v) - i(v_l) - i(v_r). \tag{3.39}$$

El objetivo es dividir el nodo de tal forma que se consiga aislar en cada corte, lo máximo posible, las categorías de los datos, separando unas de otras. Cuando el corte es bueno, esto es: en cada nuevo nodo hay mucha menos mixtura de categorías, entonces $i(v_l)+i(v_r) \ll i(v)$ y por lo tanto $\Delta(v, v_l, v_r)$ va a ser muy parecido a $i(v)$. En cambio, si el corte es muy pobre, significa que en cada nodo se mantiene, más o menos, la misma mixtura inicial, y por lo tanto $i(v_l) + i(v_r) \approx i(v)$, por ende $\Delta(v, v_l, v_r) \approx 0$.

La mejor división será aquella en la cual esta diferencia es máxima. Ya que implicaría que toda la heterogeneidad de clases presentes en el nodo v se perdió en la división, al menos en gran manera. Hay muchas maneras de definir un criterio de parada para un algoritmo de esta especie, podría ser hasta que cada nodo disminuya hasta un umbral mínimo, o que haya un número máximo de divisiones, etc.

Un ejemplo de indicador de impureza es el siguiente:

$$i(v) = \frac{1}{n} \sum_{x_i \in v} ||x_i - \bar{x}||^2. \tag{3.40}$$

Aquí:

1. n es la cantidad de datos total.
2. $x_i \in v$ son los datos que pertenecen al nodo v, entiéndanse los que al aplicar las condiciones del árbol llegan a ese nodo.
3. $\bar{x}$ es el promedio de los datos de ese nodo.

Pruning y joining

Algunas veces un algoritmo basado en dividir según una medida de impureza genera un árbol con demasiadas clases y puede ser de interés aplicar a posteriori algoritmos que agrupen clases para disminuir la cantidad de clústers. La diferencia entre la etapa de pruning y la de joining consiste en que en pruning se busca juntar nodos con un ancestro en común, mientras que en joining se busca juntar nodos con distinto ancestro común. Para este apartado me limitaré a ver un ejemplo de cómo aplicar pruning y de cómo aplicar joining.

Ejemplo de pruning: Si v es un nodo con una división v_l y v_r, se definirá una medida de disimilitud para pares de nodos con ancestro común. Si la disimilitud resulta pequeña (según un parámetro ϵ) se agrupan en un único nodo. Si es mayor a ese parámetro se mantendrán separados. Para definir la disimilitud es necesario definir previamente:

$$\forall x_i \in v_l \, , \, \forall x_j \in v_r \quad d_i = \min_{x \in v_r} ||x - x_i|| \;\; d_j = \min_{x \in v_l} ||x - x_j||. \tag{3.41}$$

Estos valores representan la mínima distancia de un dato perteneciente a un clúster con respecto al otro clúster hijo de la misma raíz.

Se puede considerar, sin perjuicio del análisis, que los datos d_i y d_j están ordenados de menor a mayor. Dado $\delta \in (0, 1)$ se definen:

$$\bar{d_l}(\delta) = \frac{1}{[\delta n_i]} \sum_{i=1}^{[\delta n_i]} d_i \quad \bar{d_r}(\delta) = \frac{1}{[\delta n_j]} \sum_{j=1}^{[\delta n_j]} d_j. \tag{3.42}$$

Que son las distancias promedio de un porcentaje δ de datos más cercanos al otro clúster. Por lo tanto, se define como medida de disimilitud $\max(\bar{d_l}(\delta), \bar{d_r}(\delta))$, que es una medida de disimilitud razonable.

Ejemplo de joining: Para esta etapa se puede utilizar una estrategia similar a la anterior. La idea es utilizar la misma noción de disimilitud entre nodos terminales y empezar a agruparlos en un mismo nodo partiendo del par que tiene mínima disimilitud. Existen las siguientes reglas de parada: 1. Alcanzar una cantidad de clústers deseada, 2. Tener un parámetro umbral que se utiliza para definir si juntar, o no, dos nodos. Si dos nodos terminales no superan ese umbral numérico, entonces se juntan, si lo superan ya no se agrupan.

3.4. Spectral clustering

Spectral clustering es una familia de métodos que se basa en obtener clústers a partir de las propiedades algebraicas de la matriz de similitud asociada a los datos; principalmente a través de un análisis de los autovalores y autovectores. Por lo que es un método del mismo género que el anterior (basado en similitud o disimilitud), pero con la diferencia específica de que utiliza la matriz de similitud y sus propiedades algebraicas. El beneficio que trae este tipo de análisis, en comparación con los anteriores, es la flexibilidad

que tiene en las figuras resultantes de los clústers, pudiendo generar figuras que con otros métodos, como k-medias, no se podrían (ver sección 3.1.6).

3.4.1. Álgebra de grafos y matriz de similitud

Para iniciar esta sección se empezará primero con lo necesario de teoría algebraica de grafos. Un grafo G es pesado cuando cada link tiene asociado un peso $w_{ij} \geq 0$. Por lo tanto, para todo grafo pesado hay siempre asociada una matriz de pesos $W = ((w_{ij})) \in \mathbb{R}^{n \times n}$. Cuando un peso es cero esto significa que no hay link entre esos vértices. El grafo puede o no ser dirigido. En el caso no dirigido la matriz es simétrica. En esta sección se considerará solamente el caso de un grafo no dirigido y sin autolinks; por lo que la matriz será simétrica y con diagonal cero. Algunas definiciones importantes:

1. **Grado de un vértice:** $d_i = \sum_{j=1}^{n} w_{ij}$.
2. Si $K \subset G$ es un subgrafo, se denotará $|K|$ a su cantidad de nodos.
3. **Volumen de un subgrafo:** $\text{Vol}(K) = \sum_{i \in K} d_i$, es decir, la suma de todos los grados de los vértices que pertenecen al subgrafo.
4. **Matriz de grados:** La matriz de grados, que se denotará D, es la matriz diagonal que tiene en su diagonal los grados de cada nodo.
5. $\forall K, H \subset \text{Vertices}(G)$, se denotará $W(K, H) = \sum_{i \in K, \, j \in H} w_{ij}$.
6. Un grafo se dice **conexo** si para todo par de puntos existe un camino que los une. Un subgrafo se dice conexo cuando el camino está contenido en el subgrafo. Un subconjunto de vértices se dice conexo si para todo par de vértices existe un camino que los une que pasa exclusivamente por dichos vértices.

Para entender mejor algunas de estas definiciones, revisaremos el caso de un grafo no pesado ($w_{ij} = 1$ cuando hay link, o cero cuando no lo hay):

1. En este caso el grado de un vértice pasa a ser la cantidad de conexiones que tiene.
2. El volumen de un subgrafo pasa a ser el doble de la cantidad de links.
3. La matriz de grados es la matriz que contiene en su diagonal la cantidad de links de cada vértice.

4. El valor $W(K, H)$ es la cantidad de conexiones entre dichos subconjuntos de vértices.

Como es posible ver, la versión pesada es similar a esta última, pero contemplando los pesos. Pasemos ahora a definir una de las matrices más importantes en teoría algebraica de grafos: la matriz Laplaciana.

Definición 3.4.1. Sea G un grafo no dirigido con matriz de pesos W, la matriz Laplaciana, o también conocida como el Laplaciano no normalizado, se define como:

$$L = D - W, \tag{3.43}$$

donde se recuerda que D es la matriz de grados.

Esta definición, aunque un tanto extraña a primera vista, tiene su fundamentación en la teoría algebraica de grafos debido a ciertas propiedades geométricas de los grafos, el lector interesado puede consultar en [24]. Lo importante de esta matriz para este análisis se encuentra en las siguientes proposiciones:

Teorema 3.11. *La matriz Laplaciana de un grafo G con matriz de pesos W verifica:*

1. *$\forall v \in \mathbb{R}^n$ se cumple:*

$$v^t L v = \frac{1}{2} \sum_{i,j=1}^{n} w_{ij}(v_i - v_j)^2. \tag{3.44}$$

2. *L es simétrica y semidefinida positiva.*

3. *El menor autovalor de L es cero y el vector $(1, 1, \ldots, 1)$ es su autovector.*

4. *L es diagonalizable y sus autovalores son mayores o iguales a cero.*

Demostración. La primera afirmación se demuestra simplemente con el siguiente desarrollo:

$$v^t L v = v^t D v - v^t W v = \sum_{i=1}^{n} d_i v_i^2 - \sum_{i,j=1}^{n} v_i v_j w_{ij} \tag{3.45}$$

$$= \frac{1}{2} \left[\sum_{i=1}^{n} d_i v_i^2 - 2 \sum_{i,j=1}^{n} v_i v_j w_{ij} + \sum_{j=1}^{n} d_j v_j^2 \right] \tag{3.46}$$

$$= \frac{1}{2}\left[\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}v_i^2 - 2\sum_{i,j=1}^{n} v_i v_j w_{ij} + \sum_{j=1}^{n}\sum_{i=1}^{n} w_{ij}v_j^2\right] = \frac{1}{2}\sum_{i,j}^{n} w_{ij}(v_i - v_j)^2. \tag{3.47}$$

De este hecho también se sigue directamente el hecho de que sea semidefinida positiva. Su simetría resulta de que es la resta de dos matrices simétricas. Los puntos tercero y cuarto son resultados directos del teorema espectral, de ser semidefinida positiva y de que las filas sumen cero. □

Teorema 3.12. *(Laplaciano y componentes conexas de un grafo) Sea G un grafo no dirigido con pesos no negativos y sin autolinks. Entonces se cumple:*

1. *La multiplicidad del autovalor cero de su matriz Laplaciana L corresponde a la cantidad de componentes conexas del grafo G.*
2. *El subespacio propio asociado al autovalor cero está generado por los vectores: $\mathbf{1}_{A_1}, \ldots, \mathbf{1}_{A_k}$; donde $\mathbf{1}_{A_i}$ corresponde al vector que tiene unos en las coordenadas correspondientes a los nodos de la i-ésima componente conexa y ceros en las otras coordenadas.*

Demostración. Sea v un autovector del cero, luego:

$$v^t L v = \sum_{i,j=1}^{n} w_{ij}(v_i - v_j)^2 = 0. \tag{3.48}$$

Como los pesos son no negativos esto ocurre si y solamente si:

$$\forall_{i,j} \;\; w_{ij}(v_i - v_j)^2 = 0. \tag{3.49}$$

Si el grafo es conexo implica que $\forall_{i,j} \;\; v_i = v_j$, por lo cual el único autovector posible es el generado por $(1, \ldots, 1)$; y este, por el lema anterior, siempre es un autovector. Ahora bien, si el grafo tiene k componentes conexas, entonces, sin pérdida de generalidad, se puede suponer que los nodos se ordenan según la componente conexa a la que pertenecen; ya que el espectro es invariante por permutaciones:

$$L = \begin{bmatrix} L_1 & 0 & 0 & \ldots & 0 \\ 0 & L_2 & 0 & \ldots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \ldots & L_k \end{bmatrix}. \tag{3.50}$$

Aquí L_i es el cuadrante de la matriz que contiene los pesos correspondientes a la i-ésima componente conexa.

Por lo tanto, aplicando el argumento anterior a cada componente conexa queda clara la tesis. □

3.4.2. Algoritmo de clustering espectral

Este algoritmo está inspirado en el teorema anterior.

1. **Input:** matriz de similitud o pesos, y número de clústers.
2. Calcular la matriz Laplaciana: $L = D - W$.
3. Calcular los k autovectores de L asociados a los k autovalores más pequeños. Sean estos $\overrightarrow{u_1}, \ldots, \overrightarrow{u_k}$.
4. Se construye la matriz $U = (\overrightarrow{u_1}| \ldots |\overrightarrow{u_k})$, que consiste en disponer los vectores anteriores como columnas. Luego, se toman $Y_1, \ldots, Y_n$ los n vectores fila de dicha matriz.
5. Por último, se aplica algún algoritmo de k-medias a los vectores $Y_1, \ldots, Y_n$ obteniendo así k clústers para estos vectores. Sean estos llamados: $C_1, \ldots, C_k$.
6. **Output:** Los k clústers $A_i = \{x_j \ : \ Y_j \in C_i\}$. *Aquí x_j corresponde al dato j-ésimo. A_i es el conjunto de los datos que su vector fila asociado pertenece a C_i.*

Observaciones:

1. En primer lugar, antes de empezar el algoritmo se asume que el grafo contiene realmente k clústers y que se caracterizan por tener fuertes similitudes entre nodos de un mismo clúster y débiles entre nodos de distintos clústers. A esta hipótesis anterior se le suma otra: que el grafo consiste en un grafo de k componentes conexas pero perturbado, esto es, la matriz de pesos W corresponde a la suma de una matriz de pesos con k componentes conexas y otra matriz de perturbación (es decir, una matriz que tiene pequeñas cantidades numéricas en sus entradas).
2. Como se asume que la matriz de pesos consiste en una matriz conexa perturbada, se supone que los k autovectores correspondientes a los k autovalores más pequeños corresponden exactamente a versiones aproximadas de los k autovectores del Teorema 3.12. La razón de esto está en que se supone que al trabajar con una matriz aproximada al caso ideal de una conexa, los k autovalores más pequeños corresponderían a versiones aproximadas del autovalor cero. Sobre esto se comentará más adelante.

3. En el caso ideal de vectores $\mathbf{1}_{A_1}, \ldots, \mathbf{1}_{A_k}$, vistos en el Teorema 3.12, al seleccionar sus correspondientes vectores fila se obtienen k puntos en el espacio euclídeo; luego, si los vectores $\overrightarrow{u_1}, \ldots, \overrightarrow{u_k}$ son aproximados a los primeros, entonces los vectores filas correspondientes se aproximarán cada uno a sus centros, que dichos centros correspondrán a los vectores filas del caso ideal. Luego, aplicar k medias es razonable.

Observación: Es importante aclarar que se asume que autovectores de matrices que son parecidas son también parecidos, y esto no siempre es cierto. Existen resultados que regulan cuándo esto se cumple. Dos de estos resultados importantes son: el teorema de Weyl y el de Davis–Kahan. Ver primero [26] y luego [25].

3.4.3. Construcción de grafos de similitud

Un aspecto importante del clustering espectral es la construcción de grafos de similitud. Aquí se verán tres posibilidades:

1. **Grafo ϵ-entorno:** Se conectan todos los pares de puntos x_i, x_j si la similitud s_{ij} es superior a una cierta constante ϵ positiva. En este caso se obtiene un grafo sin pesos.

2. **Grafo $k - NN$ (k-Nearest Neighbors):** Se conectan x_i con x_j si, o bien este último se encuentra entre los k vecinos más cercanos del primero, o bien el primero se encuentra entre los k vecinos más cercanos del segundo. También se obtiene un grafo sin pesos. *Una variante:* conectar dos nodos si ambos se encuentran entre los k-vecinos más cercanos de cada uno.

3. **Grafo completamente conectado:** conectamos todos los nodos entre sí con la siguiente función de pesos:

$$w_{ij} = e^{\frac{-d^2(x_i,x_j)}{\sigma^2}}$$

, donde σ es una constante y d es la distancia entre dos nodos.

3.5. Clustering basado en modelos probabilísticos

Ya se han visto tres tipos de métodos de clustering, el primero basado en centros y los otros dos basados en similitud; uno utilizando el orden inducido por dicha relación de similitud y el otro utilizando propiedades algebraicas del grafo subyacente y de la matriz de pesos resultante. Transversalmente a todo esto, y como se vió en la sección 3.1.5 de k-medias y probabilidad, se pueden considerar, como en varias partes de la teoría de la estadística, los datos como provenientes de una cierta realización de variables aleatorias. Es decir, al estudiar matemáticamente el clustering es posible estudiar métodos que no suponen que los datos provengan de una realización de variables aleatorias, o estudiar métodos que sí lo suponen. Esto es lo que se verá en esta sección. Ya se ha visto en la sección 3.1.5 k-medias con teoría de la probabilidad, por lo que no se comentará esa posibilidad en este apartado. Se verán algunos ejemplos de métodos de clustering utilizando la teoría de la probabilidad. Estos métodos se pueden dividir en paramétricos y no paramétricos. En el primer caso se busca estimar los parámetros necesarios para conocer el modelo probabilístico, en el segundo caso no.

3.5.1. Mixtura de distribuciones

El modelo de mixtura de distribuciones (ver el capítulo 8 de [22]) consiste en suponer que los datos son la realización de una muestra aleatoria simple de vectores aleatorios con la siguiente densidad:

$$f(x|\Psi) = \sum_{i=1}^{k} \pi_i f_i(x|\theta_i), \tag{3.51}$$

donde $f_1, \ldots, f_k$ son densidades con parámetros $\theta_1, \ldots, \theta_k$ respectivamente, cada una en proporciones distintas $\pi_1, \ldots, \pi_k$ que son números mayores a cero y suman uno; f depende del parámetro $\Psi = (\pi_1, \ldots, \pi_k, \theta_1, \ldots, \theta_k)$. Y a su vez se supondrá que cada densidad corresponde a un clúster.

Si se tiene una muestra aleatoria simple $x_1, \ldots, x_n$ de esta distribución se puede considerar la probabilidad a posteriori de que x_j provenga del i-ésimo clúster como:

$$P(x_j \in C_i) = \frac{\pi_i f_i(x_j|\theta_i)}{f(x_j|\Psi)}. \tag{3.52}$$

Se puede observar que esta consideración nace de aplicar la formula de Bayes a este caso; aunque es una aplicación análoga.

Estimación por método de máxima verosimilitud

Como método paramétrico que es, el objetivo será estimar los parámetros de la distribución de probabilidad mixta. Un método es utilizar máxima verosimilitud. Para esto se buscará encontrar una solución a la siguiente ecuación:

$$\partial \log L(\Psi)/\partial \Psi = 0. \tag{3.53}$$

Donde $\Psi = (\pi_1, \ldots, \pi_k, \theta_1, \ldots, \theta_k)$. Y la función de máxima verosimilitud se define, para una muestra aleatoria simple $x_1, \ldots, x_n$ como:

$$L(\Psi) = \prod_{j=1}^{n} f(x_j|\Psi). \tag{3.54}$$

Según [22] se puede resolver la ecuación 3.54 utilizando un algoritmo llamado expectation-maximization (EM) de Dempster et al. (1977)[47] y McLachlan y Krishnan (1997)[48].

Por lo tanto, este problema se puede reducir en los distintos métodos y algoritmos para estimar los parámetros necesarios del modelo, y al estudio de sus consecuencias. Una vez estimados los parámetros se pueden clasificar los datos de la muestra. Una variante posible es considerar distribuciones específicas, como el caso de mixtura de normales o de la distribución t-multivariada, entre otras.

3.5.2. Grafos y detección de comunidades

Otro interesante modelo paramétrico es el Stochastic Block Model utilizado para la detección de comunidades [27]. La idea de este modelo es trabajar con un grafo aleatorio con una cierta distribución paramétrica de probabilidad que busca modelar el caso de varias comunidades. Entre nodos de una misma comunidad hay más probabilidad de que haya un link, comparado con nodos de distintas comunidades. El objetivo, por lo tanto, también será estimar ciertos parámetros del modelo que permitirán clasificar cada nodo.

Quienes introdujeron el modelo más sencillo fueron Paul Erdös y Alfred Rényi. Antes de pasar al problema de clustering es necesario definir el modelo probabilístico de Erdös-Rényi:

Definición 3.5.1. (Modelo Erdös-Rényi) Se dice que un grafo aleatorio G es generado por un modelo Erdös-Rényi, y se denotará $\mathcal{G}(n,p)$ al modelo, si n corresponde a la cantidad de nodos y p es la probabilidad que tienen los nodos de conectarse cada uno de forma independiente, sin autolinks. Por lo tanto, un grafo g particular de n nodos y q links tiene la siguiente probabilidad de ser generado por este modelo:

$$P(G=g) = p^q(1-p)^{\binom{n}{2}-q}. \tag{3.55}$$

Es fácil ver esto debido a que hay $\binom{n}{2}$ links posibles en todo grafo de n nodos.

Clustering y detección de comunidades

Para establecer un método de clustering basado en un modelo probabilístico se definirá el siguiente modelo. Sea n la cantidad de nodos y k la cantidad de subconjuntos disjuntos distintos que harán las veces de comunidades; el modelo es el siguiente:

1. Dentro de cada comunidad, denotada C_i, hay un modelo Erdös-Rényi $\mathcal{G}(n_i, p_i)$ con n_i el número de nodos de esa comunidad y p_i la probabilidad de que haya un link entre nodos de esa comunidad.

2. Dadas dos comunidades C_i y C_j distintas, hay una probabilidad p_{ij} de link entre pares de esas comunidades, con independencia entre cada par.

3. Se supone también que la mayor de las probabilidades entre distintas comunidades es significativamente menor que la menor de las probabilidades de link entre nodos de una misma comunidad. Esto es:

$$\max_{i \neq j}(p_{ij}) \ll \min_{i}(p_i).$$

   Es decir, se supone que hay más conectividad dentro de las comunidades que fuera de ellas.

De todo esto es posible retirar algunas conclusiones:

Conclusión I: para el modelo anterior se tiene asociada una matriz de probabilidades.
Se tiene la siguiente matriz $B \in \mathbb{R}^{k \times k}$ definida de la siguiente forma:

$$b_{ij} = p_{ij} \qquad b_{ii} = p_i.$$

A esta matriz se le llama matriz de conectividad.

Conclusión II: para el modelo anterior se tiene asociada una matriz de pertenencia a la comunidad.
Esto es, una matriz que se denotará Θ y que pertenece a $\mathbb{R}^{n\times k}$, y en la misma, cada fila corresponde a un nodo y cada columna a una comunidad. Por fila la matriz tiene un único 1 correspondiente a la comunidad a la que pertenece el nodo y ceros en las otras columnas.

Definición 3.5.2. (Stochastic Block Model) Un Stochastic Block Model, abreviado SBM, es el modelo anterior descrito y que queda totalmente parametrizado por el par de matrices (B, Θ). Al modelo se lo denotará SBM(B, Θ).

Dada una realización de este modelo, el método de clustering consiste en estimar Θ a partir de ese dato. Por lo tanto, la finalidad es dar un estimador de Θ que sea razonable. Esto significa que el clustering, en este caso, se reduce a dar primero una estimación. En lo siguiente se profundiza un poco más en este asunto:

Para cada número i asociado a un nodo del grafo, se denotará g_i a la etiqueta de su comunidad. Esto, en términos de filas y columnas de la matriz Θ, significa que g_i es la columna que contiene al 1 en la fila i.

Definición 3.5.3. (Matriz aleatoria de adyacencia del modelo) Dado un modelo SBM(B, Θ) la matriz aleatoria de adyacencia asociada a este modelo es la que en cada entrada tiene los siguientes valores:

$$A_{ij} = \begin{cases} Bernoulli(B_{g_i g_j}) & si \quad i < j \\ 0 & si \quad i = j \\ A_{ji} & si \quad i > j \end{cases} \tag{3.56}$$

1. $B_{g_i g_j}$ es el valor numérico en la fila g_i y columna g_j de la matriz B.
2. $Bernoulli(B_{g_i g_j})$, significa la realización de una variable aleatoria Bernoulli con el parámetro $B_{g_i g_j}$.
3. Las Bernoulli son variables aleatorias independientes.

Observaciones:

1. La idea es la siguiente, dados dos nodos distintos i y j, el que tengan o no un link entre ellos funciona como una variable aleatoria Bernoulli. Si pertenecen a la misma comunidad o a distintas comunidades está contemplado en $B_{g_i g_j}$.

2. El objetivo es estimar Θ, pero es importante aclarar que esta estimación es a menos de una permutación de sus columnas, ya que no importa la etiqueta de la comunidad a la que pertenece un nodo, sino las comunidades en sí. Por lo tanto, es necesario establecer un criterio de calidad adecuado para un estimador $\hat{\Theta}$ de Θ.

Definición 3.5.4. (Criterio de calidad) Dado un estimador $\hat{\Theta} \in \mathcal{M}_{n,k}$ el criterio de calidad es el siguiente:

$$L(\hat{\Theta}, \Theta) = \min_{J \in P_k} \frac{||\hat{\Theta} J - \Theta||_{\mathcal{F}}}{n} \tag{3.57}$$

- $\mathcal{M}_{n,k}$ es el conjunto de matrices $n \times k$ que por filas tienen cero en todas las columnas a excepción de una, que tiene un uno.
- $||M||_{\mathcal{F}} = \sqrt{\text{Traza}(M^t M)}$, es la norma de Frobenius.
- P_k es el conjunto de las matrices $k \times k$ de permutación de columnas.

Observación. Lo que mide este criterio de calidad es, a menos de una permutación de columnas, cuánto es el porcentaje de error al clusterizar. Este criterio tiene la malicia de que en el caso de comunidades grandes y pequeñas coexistiendo, al ser un porcentaje, las comunidades pequeñas tienen poco peso y por lo tanto no las toma en cuenta. Otro criterio que soluciona esto se puede ver en [27].

- Siendo A la matriz aleatoria de adyacencia, se puede probar que $\mathbb{E}(A)$ es la matriz $n \times n$ que en sus entradas tiene la probabilidad de que haya un link entre dichos nodos asociados.
- $\mathbb{E}(A) = P - \text{diag}(P)$ con $P = \Theta B \Theta^t$ siendo $\text{diag}(P)$ la matriz que tiene la misma diagonal que P y ceros en donde no es parte de la diagonal.

Clustering espectral aplicado a detección de comunidades

La idea para aproximar Θ es utilizar el algoritmo de clustering espectral comentado en el apartado anterior. La bondad de trabajar con un modelo probabilístico radica en que se puede probar que el método de clustering espectral para este caso es consistente; es decir, el estimador converge a la matriz verdadera.

La matriz P tiene una estructura ideal para el clustering espectral, ya que la misma tiene sólo k filas distintas. Por razón de la estructura de $P = \Theta B \Theta^t$, y siendo B diagonalizable, es posible ver que admite la siguiente estructura:

$$P = UDU^t, \tag{3.58}$$

con:

- U una matriz $n \times k$ y D diagonal $k \times k$.
- $U^t U = Id_k$ debido a que B no sólo es diagonalizable, sino que lo es en una base ortogonal.

La siguiente proposición se puede obtener de lo anterior y su demostración se encuentra en [27]:

Lemma 3.13. *(Relación de U con Θ) Sea un modelo SBM(Θ, B) con k comunidades, donde B es de rango completo. Si UDU^t es la descomposición espectral de $P = \Theta B \Theta^t$, entonces se cumple lo siguiente:*

1. *$U = \Theta X$ con $X \in \mathbb{R}^{k \times k}$*
2. *Si X_l y X_s son dos filas de X que corresponden a los valores propios n_l y n_s de D entonces se cumple:*

$$||X_l - X_s|| \leq \sqrt{\frac{1}{n_l} + \frac{1}{n_s}}, \tag{3.59}$$

donde $||\cdot||$ es la norma euclidiana.

Corolario 3.13.1. *Como Θ es una matriz $n \times k$ que en cada fila contiene un solo 1 correspondiente a la comunidad a la que pertenece el nodo; se sigue que ΘX corresponderá a una matriz $n \times k$ donde de esas n filas habrán solamente k distintas, que son copias de las filas de la matriz X.*

A modo de ejemplo: si en la primera fila de Θ hay un 1 en la segunda posición, en ΘX la primera fila será una copia de la segunda fila de X.

Corolario 3.13.2. *Cuanto menores sean los autovalores de B mayor diferencia puede haber entre las filas correspondientes a diferentes valores propios. Cuando los valores propios son grandes la diferencia es pequeña.*

Corolario 3.13.3. *(Reversibilidad de U a Θ) Si se conoce U se puede conocer Θ a menos de permutación. Basta reconocer que filas iguales corresponden a una misma comunidad.*

Por lo tanto, si B es de rango completo y con autovalores pequeños (esto siempre se conseguirá por la naturaleza de cómo fue construida B), entonces si se estima U se puede estimar Θ, pero para estimar U solo se dispone de una realización de A, luego es necesario estimar U a partir de esa única realización de A. Para ello la idea es la siguiente, considérese primero la diferencia entre A y P:

$$A - P = (A - \mathbb{E}(A)) - \text{diag}(P). \tag{3.60}$$

Esta es una matriz simétrica. Se puede considerar $A - \mathbb{E}(A)$ una matriz de ruido. Por lo tanto, la diferencia de $A - P$ es una matriz diagonal más un ruido. Luego, es razonable pensar que los autovectores de P y de A son parecidos. Es decir, si $A = \hat{U}\hat{D}\hat{U}^t$ es la descomposición espectral de A, al modo de P en la ecuación 3.58, correspondiente a los k autovalores más grandes en valor absoluto, $\hat{U}$ tendrá, groseramente hablando, k filas distintas, que serán las de U, pero perturbadas.

Algoritmo

1. **Input:** A la matriz de adyacencia. k la cantidad de comunidades y ϵ un parámetro de aproximación.
2. **Paso 1:** Calcular $\hat{U} \in \mathbb{R}^{n\times k}$ cuyas columnas corresponde a los k autovectores asociados a los k autovalores mayores en valor absoluto de A.
3. **Paso 2:** Aplicar $(1+\epsilon)$-aproximación a k medias con k clústers a las filas de $\hat{U}$.
4. **Output:** $\hat{\Theta}$ estimación de Θ.

Comentario: Para que se entienda el paso dos basta aclarar que Lei y Rinaldo en 2015 [27] propusieron usar k-medias definida de la siguiente forma para este problema:

$$(\hat{\Theta}, \hat{X}) = \underset{\Theta\in\mathcal{M}_{n,k}\,,\, X\in\mathbb{R}^{k\times k}}{\arg\min} ||\Theta X - \hat{U}||_{\mathcal{F}}^2, \tag{3.61}$$

donde X es una matriz de permutación de columnas. El tema problemático está en que este problema es computacionalmente un problema NP-completo. Sin embargo, hay algoritmos que hallan ciertas soluciones en tiempos polinomiales, aunque estas soluciones verifican una condición más laxa:

$$||\hat{\Theta}\hat{X} - \hat{U}||_{\mathcal{F}}^2 \leq (1+\epsilon) \min_{\Theta \in \mathcal{M}_{n,k}\,,\, X \in \mathbb{R}^{k\times k}} ||\Theta X - \hat{U}||_{\mathcal{F}}^2. \tag{3.62}$$

Por esto se entiende como algoritmo de $(1+\epsilon)$-aproximación a k-medias con k clústers a las filas de $\hat{U}$.

Para terminar esta sección simplemente comentar que Lei y Rinaldo en su artículo [27] demuestran la consistencia de este método. Otra idea interesante es la planteada en este artículo [28], en donde exploran las conexiones de spectral clustering con paseos aleatorios en grafos.

3.5.3. Clustering basado en modas y estimación de conjuntos de nivel

Este apartado tiene como referencias [33], [35] y [34]; se recomienda leerlas en ese orden.

Una variante no paramétrica del método de mixtura de modelos probabilísticos es suponer que los datos provienen de una mixtura de modelos, cada uno asociado a una moda (esto es, su densidad presenta una moda), y el objeto del algoritmo es estimar la densidad para luego estimar las modas. La idea es que cada clúster es una región cerca de la moda. Hartigan en 1975 presenta una idea semejante diciendo: "Los clústers pueden pensarse como regiones de alta densidad separadas unas de otras por regiones de baja densidad". Estas regiones en inglés se llaman: "Upper level sets", que vendría a ser: conjuntos de niveles superiores; matemáticamente hablando, algo de la forma $\{f \geq c\}$ para cierta constante c y para cierta función de densidad f. Para hallar los clústers es necesario primero estimar la función de densidad para luego estimar dichas regiones. Una vez estimadas dichas regiones fácilmente puede realizarse el clustering de los datos. El problema práctico por lo tanto es: cómo estimar la densidad y cómo elegir la constante c. En este apartado solamente se comentará cómo estimar una función de densidad y luego unos resultados sobre la consistencia de este método.

Método para estimar densidades

Uno de los métodos más populares para estimar densidades no paramétricas es el método del Kernel (núcleo en inglés).

Lemma 3.14. *Sea K una función que verifica:*

$$\lim_{x\to\pm\infty} |xK(x)| = 0 \qquad \int K(x)dx = 1 \tag{3.63}$$

Sea $g(x)$ una función que verifica:

$$\int |g(x)|dx = 1 \tag{3.64}$$

Y sea una sucesión h_n que tiende a cero. Entonces se cumple:

$$g_n(x) = \frac{1}{h_n}\int K(\frac{x-y}{h_n})g(y)dy \;\; \to \;\; g(x) \quad n \to +\infty \tag{3.65}$$

Por lo tanto, podemos a partir de ahora considerar lo siguiente:

1. f será una función de densidad.
2. $f : \mathbb{E} \to \mathbb{R}$, donde $\mathbb{E}$ es un espacio métrico no necesariamente euclidiano.
3. $L_n(c) = \{f_n \geq c\}$, estimador plug-in del conjunto de nivel $L = \{f \geq c\}$ con f_n estimador de la densidad f.

Convergencia en la métrica Hausdorff

En este apartado se estudia la estimación del conjunto de nivel L con respecto a la métrica de Hausdorff. La convergencia en la métrica de Hausdorff viene a significar la convergencia en las figuras. Es decir, la figura del conjunto de nivel estimado se parecerá a la figura del conjunto de nivel a estimar. Empecemos con las definiciones:

Definición 3.5.5. (Distancia Hausdorff). Dados dos conjuntos compactos A, B en un espacio métrico $(\mathbb{E}, d)$, se define su distancia Hausdorff de la siguiente forma:

$$d_H(A, B) = \text{ínf}\{\epsilon > 0 \;\; : \;\; A \subset \mathbf{B}(B, \epsilon)\, y\, B \subset \mathbf{B}(A, \epsilon)\}, \tag{3.66}$$

en donde $\mathbf{B}(A, \epsilon) = \cup_{x\in A}\bar{\mathbf{B}}(x, \epsilon)$, siendo esta última la bola cerrada de centro x y radio $\epsilon > 0$. También es equivalente a:

$$d_H(A, B) = \text{máx}\left(\sup_{x\in A} d(x, B), \sup_{y\in B} d(y, A)\right). \tag{3.67}$$

Esta métrica de Hausdorff tiene el inconveniente de que no captura la forma del conjunto de una forma completamente exitosa. Piense en el siguiente caso: A es la bola unidad en $\mathbb{R}^2$ y B es un conjunto finito de muchos puntos uniformemente distribuidos dentro de A. En ese caso, si bien es claro que los conjuntos son muy diferentes en aspectos muy importantes aun así la distancia de Hausdorff es pequeña. Para subsanar este problema se puede estudiar la distancia de Hausdorff entre fronteras de los conjuntos. En el ejemplo anterior, si se toma la distancia de Hausdorff entre sus fronteras entonces será cercana a 1; mostrando que son figuras diferentes.

Teorema 3.15. *(Consistencia en métrica Hausdorff) Sea $(\mathbb{E}, d)$ un espacio métrico, donde las funciones f y f_n están definidas, que verifican:*

1. *$\exists r_0 > 0$ tal que si $r < r_0$ toda bola cerrada $B(x, r) \subset \mathbb{E}$ es conexa.*

y la función $f : \mathbb{E} \to \mathbb{R}$ verifica:

1. *Es continua.*

2. *$\forall x \in \mathbb{E}$ tal que $f(x) = c$ $\exists u_n, v_n$ sucesiones tales que:*

$$u_n \to x \ \ y \ \ f(u_n) > c, \tag{3.68}$$

$$v_n \to x \ \ y \ \ f(v_n) < c. \tag{3.69}$$

3. *$\partial\{f \geq c\} \neq \emptyset$ y $\exists c^- < c$ tal que $\{f \geq c^-\}$ es compacto.*

Sea $(\Omega, \mathcal{A}, \mathbf{P})$ un espacio de probabilidad y f_n una sucesión de funciones aleatorias con probabilidad uno de ser continuas que estiman a la función f cumpliendo lo siguiente:

$$||f - f_n||_\infty = \sup_{x \in \mathbb{E}} |f(x) - f_n(x)| \to 0 \ c.s. \tag{3.70}$$

Entonces se cumple:

$$\lim_{n \to +\infty} d_H(\partial L_n, \partial L) = 0 \ c.s., \tag{3.71}$$

donde $L_n(c) = \{f_n \geq c\}$, $L = \{f \geq c\}$.

Comentarios:

1. La hipótesis del espacio métrico es bastante razonable, vendría a exigir que el espacio métrico sea localmente conexo uniformemente. Se descartarían espacios métricos tipo Cantor, u otros espacios patológicos.

2. La segunda hipótesis sobre la función de densidad es que c no sea un extremo relativo.

3. La tercera hipótesis es muy razonable, ya que va de la mano con la intuición de lo que pensamos de un clúster.

Demostración. Ver [35]. □

Podemos además citar un resultado sobre la velocidad de esta convergencia agregando una hipótesis extra:

Teorema 3.16. *(Velocidad de convergencia) Sea $(\mathbb{E}, d)$ un espacio métrico y f, verificando las hipótesis del teorema anterior, y además:*

$$\exists \gamma > 0 \; y \; A > 0 \; tq \; si \; |t-c| \leq \gamma \; \rightarrow \; d_H(\{f=t\}, \{f=c\}) \leq A|t-c|. \quad (3.72)$$

Sea $(\Omega, \mathcal{A}, \mathbf{P})$ un espacio de probabilidad y f_n una sucesión de funciones aleatorias con probabilidad uno de ser continuas que estiman a la función f cumpliendo lo siguiente:

$$||f - f_n||_\infty = \sup_{x \in \mathbb{E}} |f(x) - f_n(x)| \rightarrow 0 \; c.s. \quad (3.73)$$

Entonces se verifica:

$$d_H(\partial L, \partial L_n) = O(||f_n - f||_\infty) \; c.s. \quad (3.74)$$

Convergencia en medida

Así como se puede estudiar las propiedades de convergencia de un conjunto de nivel estimado según la distancia de Hausdorff de sus fronteras. También es posible estudiar las propiedades de convergencia según el punto de vista de la medida, y esto es lo que se verá en este apartado.

Si se está en un espacio de medida con una medida μ, se puede definir la distancia (se usa esta palabra en su sentido amplio y no formal) de los subconjuntos L_n y L (los mismos que se definieron en el apartado anterior) de la siguiente forma:

$$d_\mu(L_n, L) = \mu(L_n \Delta L). \quad (3.75)$$

Aquí se denota por Δ la diferencia simétrica de dos conjuntos. Para empezar el análisis se postularán algunas hipótesis que son aceptables para los casos que se desea estudiar:

- $\mu(\{f = c\}) = 0$.
- $\mu(\{x \ : \ c - \epsilon_0 \leq f(x) \leq c + \epsilon_0\}) < \infty$ para algún ϵ_0 positivo.

Con estas hipótesis se cumple entonces el siguiente teorema:

Teorema 3.17. *Sea $(\mathbb{E}, \mathcal{F}, \mu)$ un espacio de probabilidad con $f : \mathbb{E} \to \mathbb{R}$ una función de densidad que verifica las dos hipótesis mencionadas anteriormente. Sea $(\Omega, \mathcal{A}, P)$ un espacio de probabilidad y f_n una sucesión de funciones aleatorias, esto es, para cada $\omega \in \Omega$ se tiene $f_n(\omega) : \mathbb{E} \to \mathbb{R}$ una función que estima f. Supongamos, además, que para todo n estas funciones son medibles y que $f_n, f \in L_p(\mu)$ para algún $1 \leq p \leq \infty$.*

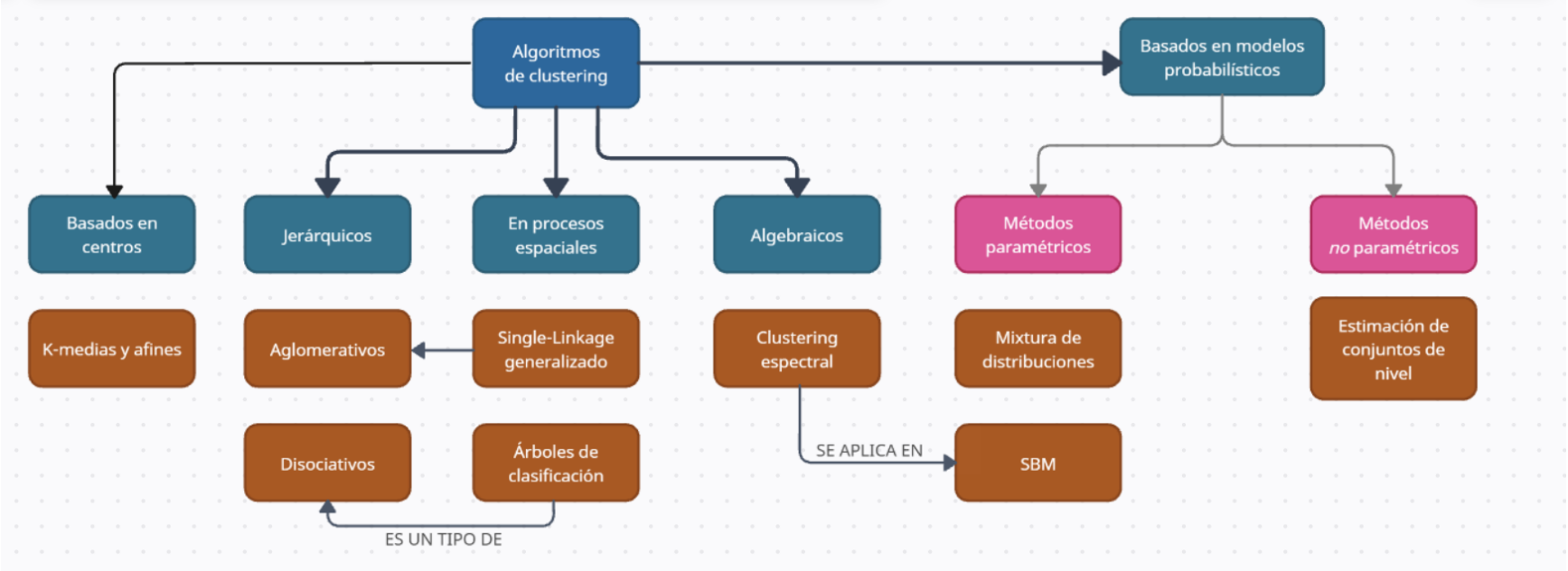
Luego si $||f_n - f||_p \to 0$ c.s. entonces se cumple:

$$d_\mu(L, L_n) = \mu(L \Delta L_n) \to 0 \ \ c.s. \tag{3.76}$$

Demostración. Ver [35]. □

Esto es suficiente para una introducción al tema de estudiar la estimación de conjuntos de nivel, ya sea en sus propiedades métricas o medibles, o cualquier otra que resulte conveniente, para utilizar estos conocimientos en orden al clustering de estimación de conjuntos de nivel. Para más información ver [33],[35] y [34].

Resumen gráfico del capítulo



Parte II

Desarrollos de una teoría general del clustering

Introducción y método investigativo

Habiendo estudiado los métodos de clustering aplicados al día de hoy con el objetivo de entender qué es lo que se utiliza, cómo funcionan, la diversidad y diferencias entre ellos, los diversos estudios, etc.; se pasará a analizar los diversos intentos de formalizar una teoría general del clustering hasta el día de hoy. Para esta empresa se utilizó como referencia la tesis doctoral de Margareta Ackerman [12], publicada en 2012. En la misma, en su segundo capítulo, detalla cuáles son los principales artículos que contribuyeron a este campo. Por lo que, en esta investigación se asumirá que solamente estos son los artículos anteriores a ella más importantes y relevantes para esta discusión. Para los estudios posteriores se investigó cuales son los artículos relevantes hasta la fecha y que sean posteriores a la tesis doctoral de Margareta Ackerman; eje central de este análisis. Estos estudios posteriores son pocos, y poco tienen que aportar sobre la solución al problema de una teoría general del clustering. Por eso se dará un breve comentario sobre ellos en el último capítulo de esta tesis.

Capítulo 4

Contribución anterior a Kleinberg

4.1. Contribución de William E. Wright

En su artículo "A Formalization of Cluster Analysis" [36], publicado en mil novecientos setenta y tres, Wright propone una fundamentación para el análisis matemático del clustering. En su introducción dice: "Thus it seems clear that there is a strong need for the development of a *theoretical* basis for this analytical method known as clustering, in contrast to the almost completely *heuristic* approach which has been so dominant up to this time. That is the purpose of this paper, to develop a theoretical clustering model, or to present a formalized notion of clustering analysis" ("Por lo tanto parece claro que existe una fuerte necesidad para el desarrollo de una base *teorética* para el análisis del método conocido como "clustering", en contraste con el método prácticamente *heurístico*, que ha sido el dominante hasta ahora. Este es el propósito de este artículo, desarrollar un modelo teórico del clustering, presentar una noción formalizada de lo que es el análisis del clustering"). El mismo autor a continuación aclara que no pretende con esto cerrar la discusión, sino simplemente dar una tentativa. ¿Cuáles son entonces estas cuestiones que no están bien definidas? El autor propondrá los siguientes problemas a resolver en su artículo:

1. El problema de la medición de los datos.
2. La naturaleza de los elementos.
3. La naturaleza de las particiones.
4. La naturaleza de los resultados.

5. La función de clustering.

6. Los axiomas que esta debe cumplir.

7. Algunos axiomas inadecuados.

El autor propone como principio de su análisis entender que el análisis del clustering consiste en el problema de:

Particionar un conjunto de elementos en clústers de tal forma que se maximice la similitud dentro de los clústers.

Las siguientes secciones reciben su nombre de las mismas secciones en las que está dividido su artículo.

4.1.1. El problema de la medición

Para Wright, el primer problema a resolver es el correspondiente a la noción de cómo medir la similitud entre elementos, a lo que él llama: el problema de la medición. Para resolver este problema, lo hará asumiendo que el conjunto de elementos que deben agruparse en clústers tienen que haberse previamente convertido en un conjunto de puntos en el espacio euclídeo, donde cada dimensión representa una medición de los datos, y es en función de esta medición que se tomará la decisión de clusterizar. Por lo tanto, si se tiene un conjunto de n datos y m mediciones para esos datos, se tiene asociada una matriz $n \times m$, donde cada fila representa las m mediciones de un elemento, y cada columna las distintas posiciones de los n elementos a lo largo de esa dimensión.

4.1.2. La naturaleza de los elementos

Los elementos, por lo tanto, serán representados como m-tuplas correspondientes a las mediciones de los elementos. Esto viene a ser su posición. Pero Wright agrega además una masa a dichos elementos. No solo tienen una posición sino que además tienen una masa, que es un número real positivo. La idea detrás viene de la posibilidad de varios datos coincidentes en posición, lo que agregaría naturalmente un peso a dicha posición. Esto permite la generalización al caso de masas no enteras. Por lo tanto, concluye afirmando que la naturaleza de los elementos es una m-tupla de números reales representando su posición más su masa asociada.

4.1.3. La naturaleza de las particiones

Una partición del conjunto debería ser, en principio, un conjunto de subconjuntos mutuamente excluyentes, totalmente exhaustivo y no vacíos. Pero como la noción de elemento fue extendida permitiendo una masa, Wright dirá que es deseable para el análisis teórico, extender la noción de partición, permitiendo que una partición no tenga por qué incluir un elemento entero sino un fragmento del mismo. Por lo tanto, si se parte el espacio en k-clústers, cada elemento tendrá asociada una k-tupla que dice cuánta masa de dicho elemento hay en cada clúster. En suma, una partición consiste en una matriz $n \times k$ cuyas columnas contienen la información de cuánta masa de dicho elemento hay en cada clúster.

4.1.4. La naturaleza del resultado

En esta sección, Wright plantea que el método para clusterizar un conjunto de datos está en definir una función de clustering que evaluará cuán bueno es un clustering. Es decir, el resultado será el clustering que maximice esta función de clustering que evalúa la bondad del clúster.

4.1.5. La función del clustering

Por lo dicho anteriormente, una función de clustering es una función que devuelve cuán bueno es el clúster obtenido a partir de los datos. Como vimos, todo dato tiene asociado un lugar y una masa, por lo tanto tiene asociada en la matriz una fila con su ubicación. Si tenemos n datos con m mediciones entonces tenemos asociada una matriz $n \times m$ cuyas filas son las ubicaciones de los datos y cada columna es una medición. A esta matriz la denotamos: $X_{n,m}$. A su vez, si se dividen los datos en k clústers, a cada dato se le puede asociar un vector fila donde cada entrada representa la cantidad de masa que pertenece a cada clúster. Esto da una matriz $n \times k$ donde cada fila es un dato y cada columna representa las masas de cada dato perteneciente al clúster correspondiente. A esta matriz la escribimos: $P_{n,k}$. Es posible unir estas dos matrices para formar una única matriz que escribiremos $(X_{n,m}, P_{n,k})$ que es $n \times (m + k)$. Mantiene las mismas filas pero agrega las columnas de $P_{n,k}$. A continuación se expone el ejemplo propuesto en dicho artículo para mayor claridad:

$$\begin{bmatrix} 1.5 & 3 & -2 & 2 & 0 \\ 2 & -2 & 1 & 1 & 2 \\ -3 & -2 & 2.5 & 0.5 & 0.5 \\ 0 & 1 & -0.5 & 0 & 1.5 \end{bmatrix}. \tag{4.1}$$

En este caso se suponen cuatro datos en una ubicación tridimensional: $(1.5, 3, -2)$, $(2, -2, 1)$, $(-3, -2, 2.5)$ y $(0, 1, -0.5)$ con masas 2,3,1 y 1.5 respectivamente.

Por lo tanto, una función de clustering f toma valores en estas matrices y devuelve un número, y la regla para clusterizar un conjunto de datos es buscar la matriz que maximiza esa función.

4.1.6. Los doce axiomas

Esta función de clustering vista anteriormente debe satisfacer, según Wright, estos doce axiomas:

1. f es invariante por reordenación de las filas.

2. f es invariante al permutar las columnas que refieren a los k clústers.

3. f es invariante por cualquier transformación lineal biyectiva de la ubicación de los datos.

4. Si en una matriz $(X_{n,m}, P_{n,k})$ hay una fila, digamos la i, que corresponde a un dato con masa cero, entonces la función de clustering evaluada en esa matriz es igual a evaluar en la matriz que corresponde a quitar esa fila. Es decir:

$$f((X_{n,m}, P_{n,k})) = f((X_{n-1,m}, P_{n-1,k})),$$

   donde $(X_{n-1,m}, P_{n-1,k})$ es la matriz menos la fila i.

5. Si una matriz $(X_{n,m}, P_{n,k})$ tiene dos datos con la misma ubicación, y $(X_{n-1,m}, P_{n-1,k})$ es la matriz correspondiente a eliminar una de las filas correspondiente a dichos datos con idéntica ubicación pero sumando las filas correspondientes a los clústers. Entonces

$$f((X_{n,m}, P_{n,k})) = f(X_{n-1,m}, P_{n-1,k}).$$

6. f es invariante por reescalar todas las masas por un factor positivo.

7. En una partición óptima de f, cada elemento está totalmente en un único clúster. Es decir, no puede tener parte de su masa en un clúster y otra parte en otro.

8. f es continua con respecto a la matriz de ubicación $X_{n,m}$.

9. f es continua con respecto a la matriz $P_{n,k}$.

10. Si tres elementos tienen ubicaciones colineales, entonces para todo clustering óptimo de f debe ocurrir que, los datos que caen en los extremos del segmento que los une, cada uno esté totalmente en un clúster, y que estén en clústers distintos.

11. Si dos elementos que están cerca además tienen masas pequeñas, entonces estos dos elementos estarán en el mismo clúster en el caso óptimo de f. De aquí se sigue que si tres elementos tienen la misma masa, los dos más cercanos entre sí estarán en el mismo clúster.

12. Si tres elementos son colineales, y uno de los que están en los extremos está más cerca que el otro del elemento mediano, y además tiene menos masa que el otro elemento extremal, entonces este estará en el mismo clúster que el del medio.

4.1.7. Algunos axiomas inadecuados

Wright en su artículo cita los siguientes axiomas como inadecuados:

1. Los dos elementos más cercanos entre sí están en el mismo clúster.

2. Los dos elementos más lejanos entre sí nunca están en el mismo clúster.

3. Siempre que tengo tres datos, los dos más cercanos están en el mismo clúster.

4.1.8. Ejemplo de función de clustering

Para verificar que estos axiomas son consistentes, Wright da un ejemplo de una función de clustering que cumple estos doce axiomas, pero no lo demuestra en su artículo, sino que cita su propia tesis doctoral como referencia. Si nuestra matriz de datos $(X_{n,m}, P_{n,k})$ es:

$$\begin{bmatrix} x_{11} & x_{12} & \dots & x_{1m} & p_{11} & p_{12} & \dots & p_{1k} \\ x_{21} & x_{22} & \dots & x_{2m} & p_{21} & p_{22} & \dots & p_{2k} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{nm} & p_{n1} & p_{n2} & \dots & p_{nk} \end{bmatrix}. \tag{4.2}$$

Entonces la función definida por Wright es:

$$f(X_{n,m}, P_{n,k}) = 1 - \frac{\sum_{r=1}^{k}\sum_{s=1}^{n} p_{sr}\sum_{u=1}^{m}\left(x_{su} - \frac{\sum_{t=1}^{n} p_{tr}x_{tu}}{\sum_{t=1}^{n} p_{tr}}\right)^2}{\sum_{s=1}^{n}\left(\sum_{r=1}^{k} p_{sr}\right)\sum_{u=1}^{m}\left(x_{su} - \frac{\sum_{t=1}^{n}\left(\sum_{r=1}^{k} p_{tr}\right)x_{tu}}{\sum_{t=1}^{n}\sum_{r=1}^{k} p_{tr}}\right)^2}.$$

4.2. Axiomatización de clustering basado en optimización por Puzicha, Hofmann y Buhmann

En el año 2000, en la revista "Pattern Recognition", Jan Puzicha, Thomas Hofmann y Joachim M. Buhmann publicaron un artículo denominado "A theory of proximity based clustering: structure detection by optimization" [37]("Una teoría del clustering basada en proximidad: detección de estructuras por optimización"). En él, desarrollan el principio de una teoría del clustering basado en proximidad (un sinónimo de similitud) que busca optimizar una función objetivo. El objeto de este capítulo es exponer las principales contribuciones de su artículo sin dar demostraciones; las mismas se pueden encontrar en el artículo referido.

4.2.1. Axiomatización de las funciones objetivo de clustering y primeros resultados

Se supondrá que se tienen datos y una medición de la disimilitud entre ellos. Por lo tanto, se tiene una matriz de disimilitud asociada: $\mathbf{D} \in \mathbb{R}^{N^2}$; con N la cantidad de datos y d_{ij} entradas que significan la disimilitud entre el dato i y el dato j; ***esta matriz puede no ser simétrica en este análisis***. Además se supondrá que el número de clústers es una cantidad fija K. Por lo tanto, se tiene para cada clúster posible una matriz asociada; cada dato está reflejado por una fila que contiene ceros y un único uno en una única columna; la columna que significa el clúster al que pertenece. Matemáticamente:

$$\mathbf{M} \in \{0,1\}^{N\times K}, \tag{4.3}$$

donde cada fila tiene un único uno y el resto cero; cada fila representa un dato y cada columna un clúster. Se denotará por $\mathcal{M}_{N,K}$ al conjunto de todas las matrices asociadas a un clustering de estos N datos en K clústers. A las entradas de dicha matriz se las denotará $\mathbf{M}_{ij}$, y pueden interpretarse como

una indicatriz; vale cero si el dato i no pertenece al clúster j y uno en el caso de pertenecer.

Axiomas elementales

En este apartado del artículo los autores van a definir los dos axiomas más elementales que según ellos debería cumplir toda función objetivo de clustering basado en similitud.

Definición 4.2.1. (**Criterio de clustering**) Una función de costo $\mathcal{H}_{N,K} : \mathcal{M}_{N,K} \times \mathbb{R}^{N^2} \to \mathbb{R}$ es un *criterio de clustering* si los siguientes axiomas se cumplen:

Axioma 1: Invarianza por permutación

$\mathcal{H}_{N,K}$ es invariante por permutaciones de los índices de los datos y de los índices de los clústers. Matemáticamente sería: si para toda $\mathbf{D} \in \mathbb{R}^{N^2}$, $\mathbf{M} \in \mathcal{M}_{N,K}$ y permutaciones π de $\{1, 2, \ldots, N\}$ y π^* de $\{1, 2, \ldots, K\}$ se tiene:

$$\mathcal{H}_{N,K}(\mathbf{M}, \mathbf{D}) = \mathcal{H}_{N,K}(\mathbf{M}^{\pi}_{\pi^*}, \mathbf{D}^{\pi}_{\pi}), \tag{4.4}$$

donde la notación $\mathbf{A}^{\pi}_{\pi^*}$ significa aplicar la permutación π a las filas y π^* a las columnas.

Axioma 2: Monotonía

Para toda $\mathbf{D} \in \mathbf{R}^{N^2}$, $\mathbf{M} \in \mathcal{M}_{N,K}$, $\Delta d \in \mathbb{R}^+$, y pares de índices i, j que corresponden a los respectivos datos, se cumple que:

1. Si pertenecen al mismo clúster:

$$\mathcal{H}_{N,K}(\mathbf{M}, \mathbf{D}) \leq \mathcal{H}_{N,K}(\mathbf{M}, \mathbf{D}^{ij}).$$

2. Si no pertenecen al mismo clúster:

$$\mathcal{H}_{N,K}(\mathbf{M}, \mathbf{D}) \geq \mathcal{H}_{N,K}(\mathbf{M}, \mathbf{D}^{ij}).$$

Aquí $\mathbf{D}^{ij}$ se obtiene de la matriz $\mathbf{D}$ por la modificación $\mathbf{D}_{ij} \to \mathbf{D}_{ij} + \Delta d$ y $\mathbf{D}_{ji} \to \mathbf{D}_{ji} + \Delta d$.

Observación:

1. El axioma uno implica que la función objetivo no hace acepción de índices; ya sea de los datos o de los clústers, sino que únicamente considera lo importante: sus relaciones de disimilitud y la partición de los clústers.

2. El segundo axioma significa que si, preservando el mismo clustering de los datos, se aumentan las disimilitudes entre datos de un mismo clúster (o, por el contrario, entre datos de distinto clúster), entonces el costo no puede disminuir (no puede aumentar).

Definición 4.2.2. (**Aditividad**) Una criterio de clustering $\mathcal{H}_{N,K}$ es *aditivo* si tiene la siguiente forma funcional:

$$\mathcal{H}_{N,K}(\mathbf{M}, \mathbf{D}) = \sum_{i=1}^{N} \sum_{j=1, j\neq i}^{N} \psi_{N,K}(i, j, D_{ij}, \mathbf{M}), \tag{4.5}$$

donde la función $\psi_{N,K} : \{1, \ldots, N\}^2 \times \mathbb{R} \times \mathcal{M}_{N,K} \to \mathbb{R}$ se llama *función de contribución*, y esta no depende de las etiquetas de cada clúster. Esto es, si a un clúster se lo asigna a la primer columna, y otro a la segunda, la función es indiferente si permutamos esas columnas; también es invariante por permutar filas.

Lemma 4.1. *Toda función de costo que sea un criterio de clustering aditivo se puede reescribir de la siguiente forma:*

$$\mathcal{H}_{N,K}(\mathbf{M}, \mathbf{D}) = \sum_{\nu=1}^{K} \sum_{i,j=1}^{N} M_{i\nu} \left[M_{j\nu} \psi^{(1)}_{N,K}(D_{ij}, n_\nu) - \sum_{\mu=1, \mu\neq\nu}^{K} M_{j\mu} \psi^{(2)}_{N,K}(D_{ij}, n_\nu, n_\mu) \right]. \tag{4.6}$$

Aquí las funciones $\psi^{(1)}$ y $\psi^{(2)}$ miden la contribución intracluster e intercluster respectivamente. A su vez, son monótonas con respecto a la variable D_{ij}. $n_\nu = \sum_{i=1}^{N} M_{i\nu}$ es la cantidad de elementos en el clústers ν.

Para ver la demostración, ver el apéndice del artículo [37].

Axiomas de invarianza

Para los autores, los axiomas anteriores son elementales, pero estos, denominados de invarianza, son realmente el núcleo de la axiomatización que proponen. A partir de ahora se supondrá N y K fijos, por lo que no es necesario explicitarlos en la notación.

Definición 4.2.3. (**Invarianza**) Una función objetivo $\mathcal{H}$ que es un criterio de clúster se le denomina *invariante*, si lo es con respecto a las siguientes transformaciones lineales:

Axioma 3: Invarianza por escala

Un criterio de clustering es invariante si para toda constante positiva c se cumple:

$$\mathcal{H}(\mathbf{M}, c \cdot \mathbf{D}) = c \cdot \mathcal{H}(\mathbf{M}, \mathbf{D}). \tag{4.7}$$

Axioma 4: Invarianza por traslación

Un criterio de clustering es invariante si para toda constante positiva $\Delta \in \mathbb{R}$ se cumple:

$$\mathcal{H}(\mathbf{M}, \mathbf{D} + \Delta) = \mathcal{H}(\mathbf{M}, \mathbf{D}) + N\Delta, \tag{4.8}$$

donde $\mathbf{D} + \Delta$ significa que a cada entrada de la matriz se le suma la constante.

Comentarios: Para empezar, me gustaría señalar lo que los propios autores exponen en su artículo al respecto de los axiomas expuestos:

1. Según ellos, la ventaja de tener axiomas de invarianza es no introducir un sesgo implícito en la escala con la que se trabaja.

2. El cuarto axioma es crucial para datos de intervalo y que no son de razón; esto es, para datos que solo importan sus diferencias entre sí (entiendo diferencia como la operación resta), y no sus proporciones; como sería el caso de la temperatura en grados Celsius, ya que su cero no es absoluto.

Lo que continúa del artículo es una serie de deducciones que suceden al combinar estos cuatro axiomas con la propiedad de la aditividad bajo ciertas hipótesis. A modo ilustrativo se expondrá solamente una de estas proposiciones.

Lemma 4.2. *Para todo criterio de clustering invariante y aditivo que su componente inter-cluster es igual a cero, esto es:* $\psi^{(2)} = 0$*, se cumple que su contribución intra-cluster debe ser:*

$$\psi^{(1)}(n_\nu) = \lambda \frac{1}{n_\nu - 1} + (1 - \lambda) \frac{1}{n_\nu(n_\nu - 1)}, \quad \lambda \in \mathbb{R}. \tag{4.9}$$

Este artículo es muy interesante, porque caracteriza ciertas funciones de costo a partir de propiedades más básicas, como puede observarse en el lema 4.2. Pienso que es muy enriquecedor una teoría de las funciones de costo para clustering, y que dichos resultados pueden ser muy útiles para los investigadores; ya que los mismos podrán conocer más a fondo las diferencias esenciales de las funciones de costo con las que están trabajando.

Capítulo 5

Teorema de imposibilidad de Kleinberg

Jon Kleinberg en su célebre artículo [1], publicado en "Advances in Neural Information Processing Systems 15" en 2002, titulado "An Impossibility Theorem for Clustering" describe en su "abstract" que si bien la teoría del clustering se centra en una idea muy intuitiva, a pesar de esto ha sido muy difícil desarrollar un paradigma común de razonamiento sobre el mismo a un nivel técnico. En dicho artículo, él propone que no existe una función de clustering que cumpla tres propiedades propuestas por él. En la sección introductoria del artículo, escribe lo siguiente: "Una aproximación estándar es representar una colección de objetos como un conjunto de puntos abstractos y definir una distancia entre ellos que representa la similitud - cuanto más cerca están, más similares son"(La traducción es propia). De lo que continúa diciendo que el objetivo del clustering está basado en la noción - vaga - de particionar en clústers que contengan puntos muy similares entre sí y disimilares entre clústers. Kleinberg sostiene que el estudio del clustering está basado solo en este nivel de descripción, y mientras por un lado hay una variedad enorme de técnicas y algoritmos (como se vio en el capítulo 3), por otro lado, hay poco trabajo que busque razonar sobre el clustering "independently of any particular algorithm, objective function, or generative data model" ("Independiente de cualquier algoritmo en particular, función objetivo, o modelo de generación de datos"). Pero, según Kleinberg, este no tiene por qué ser el caso, ya que se podría desarrollar una teoría general del clustering desde una estructura axiomática, de la misma forma que se realiza en matemática aplicada a la economía. La idea sería entonces enumerar una lista de axiomas que los algoritmos deben cumplir y, a partir de ellos, estudiar las propiedades derivadas. Por lo tanto, en su artículo se propone desarrollar una estructura axiomática para el clustering planteando tres axio-

mas que, según él, son intuitivos, llegar a la conclusión de que estos axiomas son contradictorios, luego mostrar que cada par de estos axiomas no son contradictorios, dando para cada par un ejemplo de algoritmo de clustering que cumple ambos axiomas, y finaliza estudiando los posibles outputs para un algoritmo que cumple dos de esos tres axiomas.

5.0.1. Introducción terminológica

Para el resto de este capítulo, se utilizará la misma simbología que la utilizada por Kleinberg en su artículo. En dicho documento, como no importa en qué ambiente o espacio están los datos, se utiliza $S = \{1, 2, \ldots, n\}$ para representar al conjunto de datos.

Definición 5.0.1. (**Función de distancia**) Una función de distancia es una función $d : S \times S \to \mathbb{R}$ tal que:

1. $d(i, j) = d(j, i), \quad \forall i, j \in S.$
2. $d(i, j) \geq 0, \quad \forall i, j \in S.$
3. $d(i, j) = 0$ si y solamente si $i = j$.

Bajo esta definición que usaremos no imponemos la desigualdad triangular.

Definición 5.0.2. (**Función de similitud**) Una función de similitud es una función $s : (S \times S) \setminus \{(i, i) : i \in S\} \to \mathbb{R}$ tal que:

1. $s(i, j) = s(j, i), \quad \forall i, j \in S.$
2. $s(i, j) \geq 0, \quad \forall i, j \in S.$

Observación: Una cuestión importante sobre la definición de la función de similitud es si debería definirse: $s(i, i)$. Sobre esto, no he encontrado una respuesta definitiva en las diversas fuentes revisadas; pero vislumbro tres posibilidades:

1. Mantener la definición anterior.
2. Definir $s(i, i) = 0$, por más contraintuitivo que parezca. De esta forma, el grafo de similitudes asociado no tendrá autolazos.
3. Exigir que $\forall i \in S$, $\max_{j \neq i} s(j, i) < s(i, i)$.

Definición 5.0.3. (**Función de clustering**) Una función de clustering es una función $f : \Omega_d \to \text{Part}(S)$, Ω_d es el conjunto de todas las distancias en el conjunto de datos S y $\text{Part}(S)$ es el conjunto de todas las particiones de S. A los elementos de cada partición se les llama *clústers*.

Observación: Es posible también hablar de funciones de clustering basadas en similitud, y no en distancias. Es simplemente cambiar Ω_d por Ω_s, el conjunto de todas las similitudes definidas en S.

5.0.2. Axiomatización

Kleinberg propone como primer axioma para una función de clustering que la misma sea invariante por escala.

Invarianza por escala: Una función de clustering f es invariante por escala si para cualquier $\alpha > 0$ se tiene que $f(\alpha \cdot d) = f(d)$.

Una función de clustering con esta condición no está construida en base a la escala de medición, sino en base a las proporciones entre las distancias de los datos entre sí.

El segundo axioma que propone es que sea capaz de devolver cualquier partición posible con tal de que la distancia sea conveniente. Esto se traduce así:

Abundancia: Una función de clustering f es abundante si es sobreyectiva, es decir, para toda partición Γ de S existe una distancia d de los datos tal que $f(d) = \Gamma$.

La palabra "abundancia" es una traducción de la palabra "richness" en inglés, intentando ser lo más fiel posible a la idea original que esta transmite. Esta es que las posibles devoluciones de la función de clustering sea lo suficientemente abundante (o rica) y no se restringa a solo algunas posibilidades. Habla de la abundancia de posibilidades de clústers que esta puede detectar. Recuerde lo comentado en la sección 3.1.6. Otra idea intuitiva es que si se tiene una partición Γ de S, debería la función f devolver esa misma partición si a los puntos de un mismo clúster están suficientemente cerca, y los clústers están suficientemente lejos entre sí.

Por último, agregará el axioma denominado *consistencia*. La idea es que si a puntos de un mismo clúster se acercan, y se alejan a los clústers entre sí, una función de clúster consistente debería, según Kleinberg, devolver la misma partición. Para entender la definición es necesario hacer una definición previa; se hará la misma que Kleinberg en su artículo pero con una adaptación terminológica.

Definición 5.0.4. (**d-transformación**) Dada una función de clustering f, una distancia d^* es una transformación de d consistente, o una d-transformación, si:

1. $d^*(i,j) \leq d(i,j)$ para todo par de datos i, j en un mismo clúster de $\Gamma = f(d)$.

2. $d^*(i,j) \geq d(i,j)$ para todo par de datos i, j en distintos clústers de $\Gamma = f(d)$.

Consistencia: Una función de clustering f es consistente si, para toda distancia d, es invariante por d-transformaciones. Es decir, es invariante por transformaciones consistentes. Simbólicamente: para toda d^* que sea una $d-$transformación entonces $f(d^*) = f(d)$.

5.0.3. Imposibilidad

Teorema 5.1. *(**Teorema de imposibilidad de Kleinberg**) Estos tres axiomas son contradictorios si nuestro conjunto de datos es mayor a uno. Dicho de otro modo: no existe una función de clustering cumpliendo estos tres axiomas para un conjunto de datos mayor que uno.*

Demostración. Si una función f satisface los tres axiomas entonces:

1. Por abundancia existen dos distancias d_1 y d_2 tales que $f(d_1)$ es el clustering donde cada punto es un clúster y $f(d_2)$ un clustering distinto al anterior.

2. Sea $r = \text{máx}\{d_1(x,y) : x, y \in S\}$.

3. Sea $c \geq 0$ tal que $\forall x \neq y \;\; cd_2(x,y) \geq r$.

4. Sea $d_3 = c \cdot d_2$.

5. Por invarianza por escala $f(d_3) = f(d_2)$.

6. Por consistencia $f(d_3) = f(d_1)$ lo que es una contradicción.

□

Observación: Este teorema usa muy fuertemente el hecho de que d es una distancia. Pero si se estudian funciones de clustering basadas en similitud, entonces este teorema resulta inválido y los tres axiomas consistentes. Esto ocurre porque una función de similitud admite que para datos distintos, $i \neq j$, se tenga: $s(i,j) = 0$; esto implica que una similitud no es exactamente el

inverso de una distancia (no se cumple $s = \frac{1}{d}$), a menos que se permitan distancias infinitas.

Para probar que existe una función de clustering basada en similitud y que cumple con los tres axiomas propuestos por Kleinberg (adaptados al caso de similitud) son necesarias algunas aclaraciones previas. En primer lugar, los axiomas de invarianza por escala y abundancia se mantienen prácticamente iguales en sus definiciones cambiando distancia por similitud. En segundo lugar, el axioma de consistencia se mantiene igual mientras se aclare que una similitud s^* es una s-transformación cuando ocurre lo opuesto al caso de una distancia, esto es, se aumentan las similitudes intraclústers y se disminuyen interclústers. En tercer lugar, todo conjunto de datos (S, s), donde s es una similitud, permite definir un grafo asociado; con S el conjunto de vértices y las aristas pesadas son las similitudes entre dichos vértices. En este grafo, es posible definir la siguiente relación de equivalencia: $i \sim j$ si y solamente si existe una sucesión de datos, $i = i_0, i_1, \ldots, i_k = j$, tales que $s(i_l, i_{l+1}) > 0 \forall l \in \{0, \ldots, k-1\}$. A las clases de equivalencia se las llaman *las componentes conexas del grafo*, y al conjunto de clases de equivalencia lo denotaremos Γ_s. Es importante observar que este conjunto es una partición de los datos y por lo tanto corresponde a un posible clustering de los mismos. Dicho todo esto, se puede enunciar el siguiente resultado que no figura en el artículo de Kleinberg, sino que se halló en [39]:

Proposición 1. *La función de clustering $f : \Omega_s \to Part(S)$ que se define de la siguiente forma:*

$$f(s) = \Gamma_s,$$

es invariante por escala, abundante y consistente.

Demostración.

1. La invarianza por escala es resultado de que para cualquier $\lambda > 0$ se cumple: $\Gamma_s = \Gamma_{\lambda \cdot s}$.

2. Dada una partición $\Gamma \in \text{Part}(S)$ se define $s(i, j) = 1$ si i y j están en el mismo clúster de Γ, y $s(i, j) = 0$ si no lo están. Entonces resulta que $\Gamma = \Gamma_s$, y por lo tanto es abundante.

3. Sea $s \in \Omega_s$, entonces $f(s) = \Gamma_s$. Sea s^* una s-transformación. Esto significa que si i y j están en el mismo clúster: $s^*(i, j) \geq s(i, j)$, y en el caso contrario $s^*(i, j) \leq s(i, j)$. Luego, si denotamos $\sim_{s^*}$ a la relación de equivalencia según s^* y $\sim_s$, según s, se cumple que $i \sim_{s^*} j$ si y solamente si $i \sim_s j$. Si $i \sim_s j$ significa que existe una sucesión de datos, $i = i_0, i_1, \ldots, i_k = j$, tales que $s(i_l, i_{l+1}) > 0 \, \forall l \in \{0, \ldots, k-1\}$, y por ser s^* una s-transformación, se cumple que: $s^*(i_l, i_{l+1}) > 0 \, \forall l \in \{0, \ldots, k-1\}$, ergo, $i \sim_{s^*} j$. Por otro lado, si fuese posible que $i \sim_{s^*} j$, pero $i \nsim_s j$,

entonces existiría una sucesión de datos, $i = i_0, i_1, \ldots, i_k = j$, tales que $s^*(i_l, i_{l+1}) > 0 \ \forall l \in \{0, \ldots, k-1\}$, y para algún l se cumpliría que $s(i_l, i_{l+1}) = 0$ y además $i_l \nsim_s i_{l+1}$. En ese caso, entonces, i_l e i_{l+1} no estarían en el mismo clúster de $f(s) = \Gamma_s$, lo que implicaría que $s^*(i_l, i_{l+1}) \leq s(i_l, i_{l+1}) = 0$ y por lo tanto $s^*(i_l, i_{l+1}) = 0$ resultaría en una contradicción. Luego si $i \sim_{s^*} j$ es necesario que $i \sim_s j$; y por lo tanto $\Gamma_{s^*} = \Gamma_s$, lo que prueba que f es consistente.

□

Teorema 5.2. *Para cada par de los tres axiomas existe un algoritmo de clustering que cumple esos dos axiomas.*

Daré un comentario sobre la demostración. Remito al lector al algoritmo de single linkage en el capítulo tres (ver la sección 3.2.1) para entender su funcionamiento. Para probar este teorema, usaremos tres versiones del mismo algoritmo, pero con tres condiciones de parada diferentes. Estas condiciones son:

1. Alcanzar un número k de clústers. Con k menor a la cantidad de datos.

2. Definir una distancia máxima entre dos clústers. Se define un número r positivo que sirve como criterio para que el algoritmo siga agrupando en clústers los datos. Cuando dos clústers tienen una distancia menor que r, son agrupables. El algoritmo finaliza cuando todos los clústers tienen una distancia superior a r entre sí.

3. Definir una distancia máxima entre dos clústers basada en una proporción α del diámetro total. Esto es aplicar el criterio anterior para $r = \alpha d_{max}$ donde d_{max} es la máxima distancia entre dos datos; también llamada diámetro.

Entonces, para el primer caso el algoritmo de single linkage queda invariante por escala y consistente. En el segundo caso, queda abundante y consistente. Para el último, cumple con la invarianza por escala y la abundancia. Veamos cada uno por separado:

1. **Primera condición de parada:** El hecho de ser invariante por escala es directo de la definición del single-linkage con condición de parada en k clústers. Ya que este algoritmo va a ir agrupando los clústers más cercanos hasta conseguir k. Por lo que modificar las distancias por cualquier función que respete el orden de magnitudes hará que devuelva los mismos clústers. La consistencia también es directa, ya que para una cierta distancia d el algoritmo devolvió un clustering Γ compuesto de k

clústers, significa que es el resultado final de empezar con los dos pares de datos que tienen la mínima distancia e ir agrupando los clústers con esta regla. No es, por lo tanto, difícil de ver que si se disminuyen las distancias intraclústers y se magnifican las distancias extraclústers entonces el resultado será indudablemente el mismo debido al criterio de parada.

2. **Segunda condición de parada:** Mostrar que es consistente es muy simple, ya que un clúster es un conjunto de puntos tales que todo punto tiene al menos un vecino a una distancia menor a r. Luego, si se disminuyen las distancias intraclústers y magnifican extra-clústers, ese mismo criterio se seguirá aplicando. Para probar la abundancia, si se desea obtener el clustering Γ es suficiente definir una distancia d que cumpla que es menor a r para puntos en un mismo clúster de Γ y mayor a r para puntos que no están en un mismo clúster.

3. **Tercera condición de parada:** Claramente es invariante por escala, porque la distancia r que se usa como criterio de parada no depende del diámetro total, sino de un porcentaje del mismo, y este r se reajusta por escala. La prueba de su abundancia es muy simple, dado un clustering Γ, a puntos en distintos clústers les asigno la distancia 1 y a puntos en un mismo clúster les asigno una distancia menor a α.

Observaciones:

1. El primer caso del teorema anterior no cumple el axioma de abundancia, ya que siempre devuelve k clústers, luego nunca podrá devolver distinto número de clústers.

2. En el segundo, es evidente por qué no es invariante por escala. Basta ver que si se multiplican las distancias por un número tan grande que todas superen r el algoritmo devolverá a cada dato como un clúster.

3. En el tercer caso, ocurre que si a una distancia se le aumentan las distancias entre puntos de distintos clústers eso modifica el diámetro y, por lo tanto, altera la distancia r aumentándola, eso puede generar que dos clústers que antes estaban separados se unan.

5.0.4. Rango de una función de clustering consistente e invariante por escala

La demostración vista del teorema de imposibilidad no es debida a Kleinberg. Él realiza el análisis que veremos a continuación que tendrá como corolario su teorema de imposibilidad. Se estudiará una propiedad de las posibles

salidas de una función de clustering consistente e invariante por escala. Se verá que el ser una *anticadena*, como se definirá a continuación, es incompatible con el axioma de abundancia y, por lo tanto, servirá como otra prueba del teorema de inconsistencia.

Definición 5.0.5. (**Refinamiento de un clustering**) Un clustering Γ^* es un refinamiento de otro clustering Γ si todo clúster del primero está incluido en algún clúster del segundo. Esto escrito simbólicamente queda:

$$\forall C^* \in \Gamma^* \ \exists C \in \Gamma \ : \ C^* \subset C. \tag{5.1}$$

Observación: Esta definición define un orden parcial en el conjunto de todas las particiones posibles de S.

Definición 5.0.6. (**Rango**) El rango de una función de clustering f, que se denotará Rango(f), es el conjunto de todas las particiones de S que son alcanzadas por alguna distancia; es decir, es la imagen de f:

$$\Gamma \in \text{Rango}(f) \iff \exists d : f(d) = \Gamma. \tag{5.2}$$

Para funciones de clustering basadas en similitud, la definición es análoga.

Definición 5.0.7. (**Anticadena**). Una colección de particiones de S es una *anticadena* si ninguna partición es un refinamiento de otra.

Observación: Para una cantidad de datos mayor a dos, el conjunto de todas las particiones posibles nunca forma una anticadena. Por lo tanto, ***el teorema de imposibilidad de Kleinberg se sigue también como corolario del siguiente***:

Teorema 5.3. *Todo rango de una función de clustering consistente e invariante por escala es una anticadena.*

La demostración es sencilla y se encuentra en el artículo de Kleinberg [1]. Se dará simplemente la idea general. Se supone por absurdo que pueda existir una partición que sea el refinamiento de otra, eligiendo adecuadamente una distancia d se llega a que $f(d)$ tiene que ser ambas particiones a la vez, lo que es imposible. En el artículo, se detalla la construcción de esta distancia cuidadosamente elegida para que a partir de las propiedades de consistencia e invarianza por escala ocurra la contradicción.

5.0.5. Clustering basado en centros y consistencia

Además de lo anterior, Kleinberg en su artículo demuestra que existe una tensión entre los algoritmos de clustering basados en centros y la propiedad de consistencia. Muestra que, en general, las funciones basadas en centros no cumplen esta propiedad.

Definición 5.0.8. Para todo número $k \geq 2$ y para toda función $g : \mathbb{R}^+ \to \mathbb{R}^+$ continua, no decreciente y no acotada, se define un algoritmo de clustering basado en (k, g)-centros de la siguiente forma:

1. Primero, se selecciona un conjunto de k centros $T \subset S$ para el cual la función objetivo $\Lambda_d^g(T) = \sum_{i \in S} g(d(i, T))$ es minimizada; donde:

$$d(i, T) = \min_{j \in T} d(i, j),$$

   es la distancia al centro más cercano del dato i.

2. Luego, el algoritmo devolverá una partición de S en k clústers correspondientes a asignar a cada centro el conjunto de puntos que tienen a dicho centro como el más cercano.

Puede observarse que si $g(x) = x$ se tiene un análogo al método de k-medianas, y si $g(x) = x^2$ se tiene un análogo al de k-medias.

Teorema 5.4. *Para todo $k \geq 2$ y para toda función g en las condiciones anteriores y para una cantidad de datos n suficientemente grande, el método de clustering basado en (k, g)-centros no satisface el axioma de consistencia.*

Demostración. Se dará un *sketch* de la prueba. La idea central de la misma consiste en que para probar que no se cumple la consistencia basta encontrar dos distancias, una d y otra d^*, que produzcan distintos clústers y que la última sea una d-transformación de la primera, haciendo que puntos de un mismo clúster en $\Gamma = f(d)$ se acerquen y entre distintos clústers se alejen. Si se supone que n, la cantidad de datos, es muy grande, es posible construir una distancia d de la siguiente forma: se divide S en dos subconjuntos X y Y, tal que la cantidad de datos en el primero es enorme en comparación con el segundo. En X, la distancia entre puntos se definirá como una constante fija $r > 0$, y en Y la distancia entre datos será un número muy pequeño en comparación con r; llámesele $\epsilon > 0$; y la distancia entre datos de distintos clústers será $r+\delta$, con $\delta > 0$, también pequeño. Bajo estas condiciones, $f(d) = \{X, Y\}$. Ahora, se definirá la distancia d^* que sea una d-transformación de la siguiente forma: se divide X en dos subconjuntos X_0 y X_1 de igual tamaño

y se reducirán las distancias entre puntos de esos mismos subconjuntos a un $r_1 < r$. Se dejarán todas las otras distancias iguales. La idea es que si r_1 es lo suficientemente pequeño, entonces la opción óptima de centros será que ambos centros estén uno en cada uno de esos dos conjuntos, ya que cada uno tiene mucho más peso en la cuenta que los pocos puntos de Y. Luego $f(d^*)$ corresponde a dos clústers que serán aproximadamente $\{X_0, X_1\}$, salvo agregarles los puntos de Y, y esto contradice el axioma de consistencia. □

5.0.6. Flexibilización de las propiedades

Para finalizar el artículo, Jon Kleinberg plantea el estudio de versiones más flexibles de los tres axiomas vistos al inicio. Estos son los ejemplos que él plantea:

Invarianza por escala flexibilizado: Cuando una función de clustering satisface para todo $\alpha \geq 1$ que $f(\alpha \cdot d)$ es un refinamiento de $f(d)$.

La función de single-linkage con la condición de parada de distancia r cumple abundancia, consistencia y esta versión de invarianza por escala.

También plantea versiones relajadas del axioma de consistencia:

1. **Versión 1:** Si $f(d) = \Gamma$ y d^* es una d-transformación, entonces $f(d^*)$ es un refinamiento de $f(d)$.

2. **Versión 2:** Si $f(d) = \Gamma$ y d^* es una d-transformación, entonces alguna de las dos, $f(d^*)$ o $f(d)$, es un refinamiento de la otra. Esta versión es más relajada que la anterior.

Una relajación posible para el axioma de abundancia consiste en lo que él denominó en inglés "Near-Richness" (lo denominaremos: cuasiabundancia), que consiste en que el rango de una función de clustering es el conjunto de todas las particiones posibles excepto la partición donde cada punto es un clúster.

A partir de las relajaciones anteriores, Kleinberg en su artículo comenta el siguiente resultado, cuya demostración se presenta aquí:

Teorema 5.5. *Hay un algoritmo que cumple el axioma de invarianza por escala, la primera versión relajada de consistencia y la cuasiabundancia.*

Demostración. Basta considerar el algoritmo single-linkage con la condición de parada de distancia máxima $r = \alpha \cdot \delta$ con $\alpha \geq 1$ y δ la mínima distancia entre dos puntos:

1. **Invarianza por escala:** es evidente ya que el δ cambia proporcionalmente con la distancia. Además, single-linkage solo le interesa el orden en la jerarquía de las distancias.

2. **Relajación de la consistencia:** Sea d una distancia en S y $\Gamma = f(d)$, y sea además d^* una d-transformación. Entonces de esto se sigue que $r^* = \alpha\delta^*$ es menor a $r = \alpha\delta$ con δ^* y δ las mínimas distancias entre datos según d^* y d respectivamente. Dos puntos están en el mismo clúster si se puede llegar de uno a otro pasando por puntos que están a una distancia menor o igual a la distancia de parada r. Luego, es evidente por este hecho y por la definición de d^* que $f(d^*)$ es un refinamiento de $f(d)$.

3. **Cuasiabundancia:** En primer lugar, necesariamente todo punto está a una distancia menor o igual a r de su punto más cercano, ya que α es mayor o igual a uno y δ es la mínima distancia entre dos puntos. En segundo lugar, para cualquier partición no trivial si se define la distancia entre puntos en distintos clústers como $r > \alpha$ y entre puntos de un mismo clúster como 1, este algoritmo devolverá esa misma partición.

□

Capítulo 6

Tesis doctoral de Margareta Ackerman

De toda la literatura existente sobre este tema, posiblemente la mayor contribución es debida a Margareta Ackerman, en su tesis doctoral titulada "Towards Theoretical Foundations of Clustering" (Hacia una fundamentación teórica del clustering) [12]. En esta tesis, Margareta se propone aportar al problema de estudiar matemáticamente el clustering desde el mayor punto de vista general posible realizando las siguientes contribuciones:

1. **Capítulo 3:** Discute sobre funciones de calidad para clusterings y propone una axiomatización al estudio de funciones de calidad para clusterings.

2. **Capítulo 5:** Aquí propone diversas propiedades y realiza una taxonomía de la multitud de funciones de clustering en función de cada una de estas propiedades. Es el capítulo más importante de la tesis, a mi entender.

3. **Capítulo 6:** Estudia la robustez de diversos algoritmos respecto a inputs atípicos.

4. **Capítulos 4 y 7:** Da una caracterización de los algoritmos Linkage-based dando las propiedades esenciales que, valga la redundancia, lo caracterizan.

Esta obra es la primera, principal y, hasta donde se ha podido encontrar, única respuesta al problema planteado: dar una fundamentación teórica al estudio general del clustering. En este capítulo se hará un análisis de cada capítulo, extrayendo de ellos los resultados más importantes para el desarrollo de la siguiente parte.

6.1. Funciones de calidad

Esta sección se divide en las siguientes partes: 1. Comentario sobre los axiomas de Kleinberg; 2. Su propuesta axiomática para las funciones de calidad; 3. Ejemplos de medidas de calidad y 4. Caso de dependencia de número de clústers.

6.1.1. Comentario de Ackerman a los axiomas de Kleinberg

El comentario principal que hace Margareta Ackerman al artículo de Kleinberg es el siguiente: ella afirma que la interpretación habitual del teorema de imposibilidad es la imposibilidad de establecer una teoría general del clustering. En esto ella discuerda; alega que el problema está en el formalismo utilizado por Kleinberg en vez de un problema inherente al clustering.

Ella en vez de intentar formalizar qué es una función de clustering va a dar una propuesta de formalización de funciones que midan la calidad de un clustering; inspirada en los tres axiomas de Kleinberg para funciones de clustering. Según ella, el marco de funciones de calidad es más flexible y, por lo tanto, permite definir axiomas muy similares a los de Kleinberg sin contradicción.

6.1.2. Propuesta axiomática

Una función de calidad de clustering, según Ackerman, es toda función que, recibiendo como input el dataset y el clustering, devuelve un número real positivo que tiene como objetivo medir qué tan bueno es el clustering, o qué tan significativo es. Como explica Ackerman, se puede pensar de dos formas: o bien cuanto mayor, mejor, o al revés, cuanto menor la medida de calidad: mejor. En principio, aquí se utilizará la primera. Las utilidades son diversas: poder comparar distintos clusterings realizados por distintos algoritmos, saber qué tan significativo es el clúster obtenido, etc. La idea de la propuesta axiomática está en dar los mínimos axiomas que toda función que mida la calidad de un clustering debería cumplir. Dar, como dice ella, una base teórica. Para probar la relevancia y consistencia de la base axiomática, muestra en este capítulo que las mejores medidas de calidad, estudiadas en un artículo de Milligan [40], cumplen los axiomas propuestos por ella. Veamos cuáles son estos axiomas. Para esto se utilizará la terminología empleada en el capítulo anterior.

Invarianza por escala: Una función de calidad m es invariante por escala si para todo clúster Γ de un dataset (X, d) y para toda constante positiva c se cumple: $m(\Gamma, (X, d)) = m(\Gamma, (X, c \cdot d))$.

Es la traducción del axioma homónimo de Kleinberg para funciones de clustering. No es difícil ver la justificación a este axioma. El siguiente axioma que propone es la traducción del axioma de consistencia.

Consistencia: Una función de calidad m es consistente si para todo clúster Γ de un dataset (X, d) y para toda distancia d^* que sea una transformación consistente con d con respecto al clúster Γ, se cumple: $m(\Gamma, (X, d^*)) \geq m(\Gamma, (X, d))$.

Aclaración: se dice que una distancia d^* es una transformación consistente con d con respecto al clúster Γ cuando $d^*(x, y) \leq d(x, y)$ en un mismo clúster de Γ y $d^*(x, y) \geq d(x, y)$ para datos en distintos clústers.

Esta definición, según la autora, no solo captura la idea central de consistencia sino que agrega la posibilidad de que un cambio consistente mejore la calidad de un clustering; aunque es necesario invertir la desigualdad en caso de que se utilice una medida de calidad que puntúa los mejores clústers con valores bajos y los peores con valores numéricos altos.

Para definir el próximo axioma, cabe la siguiente aclaración: un clustering de los datos Γ se dice, en este contexto, *trivial* si, o bien es el clustering correspondientes a que todos los datos estén en un mismo clúster, o bien aquel en el que cada dato forma su propio clúster.

Abundancia: Una función de calidad m es abundante si para todo clustering Γ no trivial de un dataset X existe una distancia d para ese dataset tal que:

$$\Gamma = \arg\max_{\Gamma'} \{m(\Gamma', (X, d)) \mid \Gamma' \ \textit{no es trivial}\}.$$

Estos tres axiomas son consistentes y verificarlo consiste en demostrar que la siguiente medida de calidad de un clustering cumple con los tres axiomas. Pero antes un precisión terminológica: $x \sim_\Gamma y$ significa que x y y están en el mismo clúster de Γ; $x \nsim_\Gamma y$ significa que x y y no están en el mismo clúster de Γ.

Lemma 6.1. *La siguiente medida de calidad, denominada Dunn's index, cumple con los tres axiomas anteriores.*

$$Dunn(\Gamma, (X, d)) = \frac{\text{mín}_{x \nsim_\Gamma y}\, d(x, y)}{\text{máx}_{x \sim_\Gamma y}\, d(x, y)}. \tag{6.1}$$

La demostración se encuentra en [12].

Observación: La medida de calidad anterior es paupérrima; principalmente para k-medias.

6.1.3. Ejemplos de medidas de calidad

En este apartado, se expondrán las dos principales medidas de calidad que Ackerman presenta en su tesis; las demostraciones de que cumplen los tres axiomas descritos anteriormente se pueden revisar allí mismo; no son difíciles.

1. **Índice Gamma**: Este índice fue propuesto por Baker y Hubert en [41], y en el estudio de Milligan [40] lo puntúa como el mejor índice. Para definirlo se requiere definir las siguientes cantidades:

$$d(+) = |\{\{x, y, x^*, y^*\} \subset X \mid x \sim_\Gamma y\ ,\ x^* \nsim_\Gamma y^*\ ,\ d(x, y) < d(x^*, y^*)\}|,$$

$$d(-) = |\{\{x, y, x^*, y^*\} \subset X \mid x \sim_\Gamma y\ ,\ x^* \nsim_\Gamma y^*\ ,\ d(x, y) \geq d(x^*, y^*)\}|.$$

   A partir de esto, el índice Gamma se define como:

$$\text{Gamma}(X, \Gamma) = \frac{d(+) - d(-)}{d(+) + d(-)}.$$

   Aquí $|\ |$ es el cardinal.

2. **C-Index**: Este es el segundo mejor índice según el estudio de Milligan. Fue introducido por Hubert y Levin en [42]. Nuevamente es necesario definir unas cantidades previas para luego definir el índice:

$$d_w(\Gamma, d) = \sum_{x \sim_\Gamma y} d(x, y).$$

   El w como subíndice del d proviene del inglés *within*, porque este número es la suma de todas las distancias de pares dentro de un mismo clúster; esto en inglés es: *within-cluster*. La misma explicación vale para la siguiente cantidad:

$$n_w = |\{\{x, y\} \mid x \sim_\Gamma y\}|.$$

Esto es: el número total de pares que están en un mismo clúster. Por eso $\frac{d_w}{n_w}$ es la distancia promedio entre puntos de un mismo clúster. Se denotará $S_{\text{mín}}$ al conjunto de n_w pares con menor distancia en X. Análogamente, $S_{\text{máx}}$ es el conjunto de n_w pares con mayor distancia en X. Por mayor claridad daré un ejemplo: sea $n_w = 5$, lo que acabo de decir es que $S_{\text{mín}}$ es un conjunto que contiene los cinco pares de puntos con menores distancias entre sí; en caso de ambigüedad elegir cualesquiera. Lo mismo para $S_{\text{máx}}$, sería el conjunto de los cinco pares con mayor distancia entre sí. A partir de esto se define $\text{mín}(n_w, d)$ como la suma de las distancias de los pares de puntos en $S_{\text{mín}}$ y análogamente se define $\text{máx}(n_w, d)$. A partir de aquí se define el C-Index como:

$$\text{C-Index}(\Gamma, d) = \frac{d_w(\Gamma, d) - \text{mín}(n_w, d)}{\text{máx}(n_w, d) - \text{mín}(n_w, d)}. \tag{6.2}$$

6.1.4. Dependencia del número de clústers

El objeto de esta sección, por parte de Margareta, es analizar las funciones de costo, en algoritmos como k-medias, como medidas de calidad. En general, estas funciones fallan en dos de los tres axiomas: invarianza por escala y abundancia; aunque la invarianza por escala se puede obtener al normalizar la función de costo. Aun así, estas funciones como medidas de calidad tienen una propiedad interesante: tienen preferencia por clústers que son refinados de otros clústers. Esto es, si tengo dos clusterings distintos de un mismo dataset pero que uno es un refinado del anterior, entonces la función de costo puntúa siempre mejor al refinado. Esta propiedad hace difícil utilizarla como medida de calidad para comparar clusterings con distintos números de clústers; piénsese en la función de costo de k-medias. Por lo tanto, si se quiere utilizar una medida de calidad que sea independiente del número de clústers, esta no puede tener preferencia por refinados; incluso un resultado de este apartado es que no se puede ser preferente por refinados y abundante a la misma vez, una medida con preferencia por refinados implica que no cumple el axioma de abundancia. Por otro lado, cuando se trabaja con funciones de clustering que tienen el número de clústers fijo, entonces evidentemente no es necesario que la medida de calidad cumpla el axioma de abundancia, y se podría utilizar la función de costo como medida de calidad.

6.2. Taxonomía y propiedades

Quizás el aporte más importante, o nuclear, de su tesis se encuentra en el estudio taxonómico de las distintas funciones de clustering y su diversidad de propiedades. Este estudio se encuentra principalmente en el capítulo cinco de la tesis; si bien algunas propiedades y definiciones importantes se encuentran en el capítulo cuatro.

Por lo tanto, en esta sección se revisarán: 1. Propiedades importantes descritas por Margareta; 2. Taxonomía de los principales algoritmos y potenciales propiedades que podrían ser axiomas; 3. Teoremas y resultados importantes.

6.2.1. Propiedades importantes

Es importante rememorar las principales tres propiedades propuestas por Kleinberg: *Invarianza por escala*, *Abundancia* y *Consistencia*. Fueron explicadas en el capítulo anterior; es importante comprender estas propiedades porque son la motivación para algunas de las propiedades que se verán a continuación.

Función de clustering e invarianza por representación

A continuación se harán redefiniciones de conceptos ya estudiados en el capítulo anterior para estar más acorde con la terminología empleada por la autora en su tesis.

Definición 6.2.1. Una función de clustering es una función que recibe como input el par (X, d), siendo X el conjunto de los datos y d una medida de distancia (o similitud s), y devuelve un clustering de los datos. Una función de k-clustering, además de recibir ese par como input, recibe un número natural k y devuelve un k-clustering, esto es: una partición de los datos en k subconjuntos disjuntos y no vacíos.

Definición 6.2.2. (**Isomorfismo de datasets**) Dos datasets (X, d) y (Y, d^*) se dicen isomorfos si existe una biyección ϕ que preserva las distancias (es decir, como una isometría), esto es: $d(x_i, x_j) = d^*(\phi(x_i), \phi(x_j)),\ \forall x_i, x_j \in X$.

Definición 6.2.3. (**Isomorfismo de clusterings**) Dos clusterings Γ y Γ', el primero asociado al dataset (X, d) y el segundo asociado al dataset (Y, d^*), se dicen isomorfos si existe un isomorfismo de datasets $\phi : X \to Y$ y además todo clúster $C'_i \in \Gamma'$ es la imagen de otro clúster $C_j \in \Gamma$; esto es: $C'_i = \phi(C_j)$.

Definición 6.2.4. (**Invarianza por representación**) Una función de clustering, o k-clustering, f es invariante por representación si siempre que dos dataset (X, d) y (Y, d^*) son isomorfos, entonces los clústers $f(X, d)$ y $f(Y, d^*)$ también son isomorfos.

Observación: esta definición será de suma importancia para el desarrollo del último capítulo. La intuición es que el output de la función no depende del dato en sí, sino exclusivamente de las relaciones de distancia d (o similitud s) entre los datos.

Localidad

Esta es una de las propiedades más interesantes y más importantes para caracterizar las funciones del tipo single-linkage. El concepto detrás es el siguiente: una función de k-clustering es local cuando al restringir el dominio a solamente algunos clústers del output y al volver a poner ese nuevo dominio como input, se vuelven a obtener esos mismos clústers.

Definición 6.2.5. (**Localidad**) Una función de k-clustering f es local si para todo clustering $\Gamma = f(X, d, k)$ que sea un output de f para algún dataset (X, d) y para todo subconjunto de clústers $\Gamma^* \subset \Gamma$ se tiene:

$$f(\bigcup \Gamma^*, d, |\Gamma^*|) = \Gamma^*. \tag{6.3}$$

Aquí $\bigcup \Gamma^*$ significa el conjunto formado por todos los datos del subclustering Γ. La función d debe tomarse como restringida a este nuevo conjunto.

Para mayor comprensión se expondrá el siguiente ejemplo: supongamos que se tiene un dataset (X, d) y se obtienen los siguientes k clústers: $f(X, d, k) = \Gamma = \{C_1, \ldots, C_k\}$; que la función f sea local significaría que si se seleccionan, por ejemplo, los primeros $k^* < k$ clústers de ese clustering: es decir, a $C_1, \ldots, C_{k^*}$, y se denotará a este subclustering Γ^*, y se tomara como nuevo dataset $Y = C_1 \cup C_2 \cup \ldots \cup C_{k^*}$ entonces $f(Y, d, k^*) = \Gamma^*$.

Esta propiedad es esencial para el caso de las funciones del tipo single-linkage; porque estas generan los clústers utilizando un método aglomerativo. Los métodos aglomerativos, en general, son locales.

Consistencia

La definición de consistencia ya se expuso en el capítulo sobre Kleinberg (véase el capítulo 5); se verán, por lo tanto, dos propiedades derivadas y relajadas de la misma.

Definición 6.2.6. (**Consistencia exterior**) Dado un clustering Γ de un dominio (X, d), decimos que una distancia d^* es exteriormente (o externamente) consistente con respecto a (Γ, d) si:

1. $d^*(x, y) = d(x, y)$ cada vez que $x \sim_\Gamma y$,
2. $d^*(x, y) \geq d(x, y)$ cuando $x \nsim_\Gamma y$.

Luego, una función de clustering f es exteriormente consistente (o externamente consistente) si para toda distancia en las condiciones anteriores se cumple: $f(X, d^*) = f(X, d)$.

Definición 6.2.7. (**Consistencia interior**) Dado un clustering Γ de un dominio (X, d), decimos que una distancia d^* es interiormente consistente (o internamente consistente) con respecto a (Γ, d) si:

1. $d^*(x, y) \leq d(x, y)$ cada vez que $x \sim_\Gamma y$,
2. $d^*(x, y) = d(x, y)$ cuando $x \nsim_\Gamma y$.

Luego una función de clustering f es interiormente consistente si para toda distancia en las condiciones anteriores se cumple: $f(X, d^*) = f(X, d)$.

Observación: ser consistente es ser exterior e interiormente consistente.

Abundancia

Las definiciones que se darán a continuación tienen el mismo espíritu de las anteriores, pero a diferencia de las anteriores que son relajaciones de la propiedad de consistencia, en este caso serán dos propiedades más restrictivas que la abundancia.

Definición 6.2.8. (**Abundancia exterior o exteriormente abundante**) Una función de k-clustering f es exteriormente abundante si para toda colección de datasets $(X_1, d_1), (X_2, d_2), \ldots, (X_n, d_n)$ existe una distancia d sobre $\cup_{i=1}^n X_i$ que extiende las distancias anteriores y además:

$$f(\cup_{i=1}^n X_i, d, n) = \{X_1, X_2, \ldots, X_n\}.$$

Definición 6.2.9. (**Abundancia interior o interiormente abundante**) Una función de k-clustering f es interiormente abundante si para todo dataset (X, d) y para todo k-clustering Γ de X existe una distancia d^* tal que $d^*(x, y) = d(x, y)$ cuando $x \nsim_\Gamma y$ y $f(X, d^*) = \Gamma$.

Definición 6.2.10. (**k-abundancia**) Una función de k-clustering es k-abundante si para toda partición del dominio X en k subconjuntos no vacíos $X_1, X_2, \ldots, X_k$ hay una distancia d sobre $\cup_{i=1}^{k} X_i$ tal que:

$$f(\cup_{i=1}^{k} X_i, d) = \{X_1, X_2, \ldots, X_k\}.$$

Definición 6.2.11. (**Umbralmente abundante**) Una función de k-clustering f es umbralmente abundante si para todo clustering Γ de X existen pares de números reales $a < b$ los cuales forman un umbral con la siguiente propiedad: para toda distancia d que cumpla $d(x, y) \leq a$ cuando $x \sim_\Gamma y$ y $d(x, y) \geq b$ cuando $x \nsim_\Gamma y$, entonces $f(X, d, |\Gamma|) = \Gamma$.

Otras propiedades:

Un clustering Γ_2 es un refinado de Γ_1 en X si todo clúster de Γ_1 es una unión de clústers de Γ_2. Se denotará $\Gamma_2 \preceq \Gamma_1$.

Definición 6.2.12. (**Refinante**) Una función de k-clustering F es ***refinante*** si para todos $1 \leq k_1 \leq k_2 \leq |X|$ se cumple:

$$F(X, d, k_2) \preceq F(X, d, k_1)$$

Esto quiere decir que al aumentar la cantidad de clústers (k) se obtienen refinados del clustering anterior.

Antes de la siguiente definición veamos una definición previa:

Definición 6.2.13. (**Equivalencia según el orden**) Dos distancias d_1 y d_2 se dicen equivalentes según el orden si:

$$\forall x, y, z, w \;\; d_1(x, y) < d_1(z, w) \iff d_2(x, y) < d_2(z, w)$$

Definición 6.2.14. (**Independiente del orden**) Una función F de k-clustering es independiente del orden si para todo par de distancias d_1 y d_2 equivalentes según el orden se tiene:

$$\forall k \;\; F(X, d_2, k) = F(X, d_1, k)$$

6.2.2. Taxonomía de los principales algoritmos

Function	outer consistent	inner consistent	local	refinement-preferring	order invariant	k-rich	outer rich	inner rich	threshold rich	scale invariant	rep. independence
Single Linkage	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Average Linkage	✓	*X*	✓	✓	*X*	✓	✓	✓	✓	✓	✓
Complete Linkage	✓	*X*	✓	✓	✓	✓	✓	✓	✓	✓	✓
k-medoids	✓	*X*	✓	*X*	*X*	✓	✓	✓	✓	✓	✓
k-means	✓	*X*	✓	*X*	*X*	✓	✓	✓	✓	✓	✓
Min sum	✓	✓	✓	*X*	*X*	✓	✓	✓	✓	✓	✓
Ratio cut	*X*	✓	*X*	*X*	*X*	✓	✓	✓	✓	✓	✓
Normalized cut	*X*	*X*	*X*	*X*	*X*	✓	✓	✓	✓	✓	✓

La imagen supra fue extraída directamente de la tesis de Margareta, la cual es una tabla con dos fines: 1. Ver las principales diferencias de los algoritmos según las propiedades vistas anteriormente; 2. Encontrar las propiedades que todos los algoritmos más importantes cumplen en orden a seleccionarlas como candidatas a axiomas. En esta imagen *refinement-preserving* es la propiedad de *refinante*; *k-rich* es *k-abundante*; *outer rich* es *externamente abundante*; *inner rich* es *internamente abundante*; *threshold rich* es *Umbralmente abundante* y por último *rep. independence* es *invariante por representación*.

En su tesis, Margareta demuestra estas propiedades para cada una de esas funciones de clustering. Alguna de estas funciones no se han definido anteriormente, por lo que se darán sus definiciones a continuación:

Definición 6.2.15. (**Min-sum**) Algoritmo de k-clustering basado en minimizar la siguiente función de costo:

$$\text{MinSum}(\Gamma, (X, d)) = \sum_{x \sim_\Gamma y} d(x, y)$$

Las dos siguientes son funciones de clustering basadas en similitud. Recordar que $s(x, y)$ es una función de similitud y la definición de función de similitud es la 5.0.2; se encuentra en el capítulo 5. Se denotará a un clúster del clustering Γ como Γ_i, y se define:

- $cut(\Gamma_i, \Gamma_j) = \sum_{x \in \Gamma_i \, , \, y \in \Gamma_j} s(x, y)$,

- $vol(\Gamma_i) = \sum_{x,y \in \Gamma_i} s(x,y)$.

Definición 6.2.16. (**Ratio cut**) Algoritmo de clustering que devuelve el clustering que maximiza la siguiente función:

$$\text{RatioCut}(\Gamma, (X,s)) = \sum_{\Gamma_i \in \Gamma} \frac{cut(\Gamma_i, \Gamma_i^c)}{|\Gamma_i|}.$$

Definición 6.2.17. (**Normalized cut**) Algoritmo de clustering que devuelve el clustering que maximiza la siguiente función:

$$\text{NormalizedCut}(\Gamma, (X,s)) = \sum_{\Gamma_i \in \Gamma} \frac{cut(\Gamma_i, \Gamma_i^c)}{vol(\Gamma_i)}.$$

Observación: en el último capítulo comentaré que, en mi opinión, estos algoritmos deberían buscar minimizar estas funciones, y no maximizarlas; porque cada término en dichas sumas mide una noción de similitud interclústers. He puesto la definición tal como está en la tesis de Ackerman.

Conclusión de la autora

En primer lugar, comenta que una lista de axiomas para funciones de clustering, idealmente, debería cumplir las siguientes dos propiedades, que son conceptos pertenecientes a la Lógica matemática:

- *Corrección.*
- *Completitud.*

Un conjunto de axiomas es *correcto* si toda proposición deducible a partir de los axiomas es válida en la interpretación semántica pretendida; en otras palabras, si todo lo que es posible deducir es verdadero según lo que se pretende formalizar. En este contexto, según la autora, sería que todos los elementos de la clase que dichos axiomas pretenden describir cumplen esos axiomas. Esto presupone una noción semántica de qué es una *función de clustering*; es decir, sería necesario tener previamente una definición conceptual para poder verificar que la formalización producida por los axiomas realmente captura esa realidad que se intenta formalizar. Por claridad compárese con el ejemplo de la definición de función continua. Previamente ya había una definición semántica de lo que es continuidad para una función, y la formalización introducida por Weierstrass captura formalmente dicha idea.

Un conjunto de axiomas es *completo* si todo lo que es semánticamente verdadero y expresable en el lenguaje es sintácticamente deducible. La autora lo expresa diciendo que para una clase de elementos todas las propiedades

comunes a todos ellos son deducibles a partir de dichos axiomas.

Según Ackerman, una de las mayores dificultades en este contexto está en que no existe una definición semántica de lo que es una función de clustering. Como esta definición no existe, se desearía, entonces, definir qué es una función de clustering a partir de ciertos axiomas; pero eso entrañaría la dificultad de que las propiedades de corrección y completitud perderían su sentido habitual, ya que ese sistema sería considerado correcto y completo. Ella opina que hay que ir a por menos. Hay ejemplos de lo que consideramos una función de clustering, y, a su vez, ejemplos de lo que no debería ser una función de clustering. Ella propone, por lo tanto, utilizar como axiomas aquellas propiedades compartidas por todos los ejemplos conocidos y comunes (llama a esto: corrección relajada) y además que funciones de partición que no son de clustering fallen en al menos uno de los axiomas (esta sería una relajación de la completitud, según ella).

La taxonomía expuesta supra revela que algunas propiedades aspiradas a axiomas no sirven para todas las funciones (*consistencia interna*, *localidad*,*consistencia externa*). Pero hay otras que sí cumplen todos esos ejemplos, como: invarianza por representación, invarianza por escala, y las propiedades de k-abundancia.

Según parece, esas propiedades son buenas candidatas a axiomas, aunque de algunas se podrían prescindir. Si se seleccionan solamente: *invarianza por representación*, *invarianza por escala* y *umbralmente abundante* es posible deducir las restantes: *k-abundante*, *internamente abundante* y *externamente abundante*. Esto está demostrado en su tesis y se expondrá la prueba enseguida.

Aun así, según ella, este conjunto de propiedades no cumple con la relajación de completitud, ya que aún existirían ciertas funciones que cumplen estos tres axiomas pero que no deberían calificarse como de clustering, aunque ella no expone cuáles. Por lo que propone que estas tres propiedades junto con algunas otras podrían formar un conjunto axiomático completo.

6.2.3. Teoremas y resultados importantes

Teorema 6.2. *Si una función de k-clustering es invariante por escala y umbralmente abundante entonces es externamente abundante e internamente abundante.*

Demostración. Primero veamos la abundancia externa. Consideremos los siguientes datasets con sus respectivas distancias: $(X_1, d_1), \ldots, (X_k, d_k)$. Sea $X = \cup_{i=1}^{k} X_i$. Como f es umbralmente abundante, para el clustering $\Gamma = \{X_1, X_2, \ldots, X_k\}$ existe un umbral numérico $a < b$ que genera ese clustering (bajo la consigna dada en la definición de umbralmente abundante). Tomemos una constante c tal que:

$$\max_{1 \leq i \leq k} \max_{x,y \in X_i} c \cdot d_i(x, y) < a. \tag{6.4}$$

Sea d^* una distancia sobre X tal que $d^*(x, y) = b$ si x, y no pertenecen a un mismo X_i, y si $x, y \in X_i$ entonces $d^*(x, y) = c \cdot d_i(x, y)$. Como f es umbralmente abundante se sigue:

$$f(X, d^*, k) = \Gamma = \{X_1, X_2, \ldots, X_k\}. \tag{6.5}$$

Tomando como nueva distancia: $d = \frac{1}{c} d^*$ se cumple por invarianza por escala:

$$f(X, d, k) = \Gamma = \{X_1, X_2, \ldots, X_k\}. \tag{6.6}$$

Y se puede verificar fácilmente que si $x, y \in X_i$ se tiene:

$$d(x, y) = \frac{1}{c} d^*(x, y) = \frac{c}{c} d_i(x, y) = d_i(x, y).$$

Luego, es externamente abundante. Veamos ahora la abundancia interna. Dado un dataset (X, d) y una partición Γ, queremos mostrar que existe una distancia d^*, que coincide con d para pares de datos en distintos clústers, y que $f(X, d^*, k) = \Gamma$. Como f es umbralmente abundante tenemos nuevamente las constantes umbrales $a < b$ de la definición. Construiremos una nueva distancia d_1 de forma similar al caso anterior. Primero tomemos una constante c tal que:

$$\min_{x \nsim_\Gamma y} c \cdot d(x, y) > b. \tag{6.7}$$

Definimos d_1 de la siguiente forma. Si $x \nsim_\Gamma y$ entonces $d_1(x, y) = c \cdot d(x, y)$. Si $x \sim_\Gamma y$, entonces $d_1(x, y) = a$. Por ser umbralmente abundante:

$$f(X, d_1, k) = \Gamma.$$

Tomando $d^* = \frac{1}{c} d_1$ se concluye la tesis.

□

Corolario 6.2.1. *Toda función de k-clustering umbralmente abundante e invariante por escala es k-abundante.*

Demostración. Se sigue de que tanto ser exteriormente abundante como interiormente abundante implican ser k-abundante. □

Con este teorema se probó lo afirmado anteriormente: si se pide como axioma ser independiente por representaciones, invariante por escala y umbralmente abundante entonces necesariamente se sigue la k-abundancia externa, interna y a secas.

Veamos ahora dos resultados interesantes que siguen la misma idea:

Teorema 6.3. *1. Ser invariante por escala, exteriormente abundante y exteriormente consistente implica ser interiormente abundante.*

2. Ser invariante por escala, interiormente abundante e interiormente consistente implica ser exteriormente abundante.

Demostración. Probemos la primera afirmación primero: sea (X, d) un dataset y $\Gamma = \{X_1, X_2, \ldots, X_k\}$ una k-partición. Sea además $d_i = d|_{X_i}$. Como f es exteriormente abundante existe una distancia d^1 que extiende cada una de las distancias $d_1, d_2, \ldots, d_k$ y $f(X, d^1, k) = \Gamma$. Tomemos la siguiente constante:

$$m = \max_{a \nsim_\Gamma b} \frac{d^1(a, b)}{d(a, b)}. \tag{6.8}$$

Sea la distancia d^2 definida de la siguiente forma: se mantiene idéntica dentro de los clústers, y para pares de puntos en clústers distintos: $d^2(x, y) = m \cdot d(x, y)$. Esta distancia es (Γ, d^1)- exteriormente consistente. Luego $f(X, d^2, k) = f(X, d^1, k) = \Gamma$. Tomando $d^3 = \frac{d^2}{m}$ por invarianza por escala tenemos: $f(X, d^3, k) = f(X, d^2, k)$. Observemos que si $x \nsim_\Gamma y$ se tiene:

$$d^3(x, y) = \frac{d^2(x, y)}{m} = \frac{m \cdot d(x, y)}{m} = d(x, y). \tag{6.9}$$

De lo que resulta la abundancia interna.

Veamos ahora la segunda afirmación. Consideremos los siguientes datasets: $(X_1, d_1), (X_2, d_2), \ldots, (X_k, d_k)$. Sea $X = \cup_{i=1}^{k} X_i$. Tomemos como partición de X a $\Gamma = \{X_1, X_2, \ldots, X_k\}$. Por abundancia interna existe una distancia d^1 para todo X tal que $f(X, d^1, k) = \Gamma$. Es importante notar que d^1 no necesariamente extiende ninguna de las d_i. Consideremos la siguiente constante:

$$m = \min_{x \sim_\Gamma y} d^1(x, y). \tag{6.10}$$

A partir de ella construiremos la siguiente distancia d^2. Si $x \nsim_\Gamma y$ entonces $d^2(x, y) = d^1(x, y)$ y caso contrario, si $x, y \in X_i$ definimos la distancia como $d^2(x, y) = c \cdot d_i(x, y) \leq m$, para alguna constante c positiva. Claramente d^2 es una transformación (Γ, d^1)- interiormente consistente. Luego, por consistencia interior $f(X, d^2, k) = f(X, d^1, k) = \Gamma$. Ahora bien, tomando $d^3 = \frac{d^2}{c}$ tenemos, por invarianza por escala, que $f(X, d^3, k) = f(X, d^2, k)$. Notemos que esta distancia es una extensión de las distancias originales: sean $x, y \in X_i$ se tiene:

$$d^3(x, y) = \frac{1}{c} \cdot d^2(x, y) = \frac{c}{c} d_i(x, y) = d_i(x, y). \tag{6.11}$$

De aquí se sigue que es exteriormente abundante. □

Observación personal: En la tesis de Ackerman no lo aclara, pero se puede relajar la hipótesis de ser interiormente abundante en la segunda afirmación anterior por ser k-abundante. Nótese que en realidad lo que se utiliza en la prueba es la k-abundancia y no la abundancia interior. Eso ya no ocurre con la primera afirmación, donde se utiliza necesariamente la abundancia exterior en la prueba. Por lo que puedo dar el siguiente corolario:

Corolario 6.3.1. *Toda función de k-clustering invariante por escala, k-abundante e interiormente consistente es exteriormente abundante.*

Otro corolario de lo anterior es:

Corolario 6.3.2. *Una función de k-clustering que es consistente e invariante por escala es exteriormente abundante si y solamente si es interiormente abundante.*

Teoremas de imposibilidad

Teorema 6.4. *No es posible que una función de clustering sea abundante, externamente consistente e invariante por escala.*

Demostración. Es similar a la vista en el capítulo sobre Kleinberg. Sea X un dominio que tenga más de un punto. Por abundancia hay dos distancias d_1 y d_2 tales que $f(X, d_1) = X$ y $f(X, d_2) \neq X$. Sea r el diámetro de X según d_1 y sea c una constante positiva que cumple que para todo par de puntos $x \neq y$ se tiene que $c \cdot d_2(x, y) \geq r$. Por lo tanto llegamos a una contradicción, ya que por invarianza por escala $f(X, d_2) = f(X, c \cdot d_2) \neq X$ pero por consistencia externa $f(X, d_1) = f(X, c \cdot d_2) = X$. □

Teorema 6.5. *No es posible que una función de clustering sea interiormente consistente, invariante por escala y abundante.*

Demostración. Por abundancia hay dos distancias d_1 y d_2 tales que $f(X, d_1) = \{X\}$ (Esto es, solo hay un clúster) y $f(X, d_2) \neq \{X\}$. Sea r la mínima distancia entre dos puntos distintos según d_1. Luego hay una constante positiva c tal que $c \cdot d_2(x, y) \leq r$ para cualquier par de puntos $x \neq y$. Con un argumento similar al anterior llegamos a una contradicción. □

Observación: El problema por el cual hay estas imposibilidades es doble. Una razón es que la propiedad de abundancia es muy fuerte. Ackerman observa que este problema desaparece cuando la cantidad de clústers es especificada; a saber, cuando hablamos de funciones de k-clustering. En estos casos hay funciones k-abundantes, consistentes e invariantes por escala. La segunda razón es porque se consideran funciones de clustering basadas en distancias (o disimilitud); cuando se consideran funciones de clustering basadas en similitud, al permitir que exista $s(x, y) = 0$ algunas de esas imposibilidades desaparecen, como se comentó en el capítulo de Kleinberg.

6.3. Robustez y oligarquías

El capítulo seis de su tesis es titulado: oligarquías en clusterings. Básicamente, es un estudio sobre la robustez de las funciones de clustering, esto es, qué tan sensibles son a pocos datos atípicos que estén muy lejos de los verdaderos datos típicos.

Para dar una síntesis de lo más importante de este capítulo, se dividirá en los siguientes apartados: 1. Definiciones previas 2. Principales resultados.

6.3.1. Definiciones previas

En general, los algoritmos de clustering son sensibles a pocos datos atípicos bajo ciertas condiciones. Por lo que no es una gran elección evaluar la

robustez de una función de clustering en función de qué tan suceptible es a pocos datos muy distintos a los originales. Esto se debe a que la propia distribución de los datos originales puede ser mala para clusterizar; piénsese en un montón de datos todos aglomerados, en ese caso es difícil discernir cuáles serían los k-clústers. Esta noción de datos que son aptos para ser clusterizados y datos que no lo son es importantísima para la noción de robustez. Una función de clustering será robusta si su output no es afectado significativamente ante datos atípicos cuando los datos no atípicos se presentan en condiciones ideales. Caso contrario, no será robusta.

Definición 6.3.1. (**Distancia de Hamming**). Dado dos clusterings Γ_1, Γ_2 de un dataset X, se define la distancia de Hamming de los siguientes clusterings como:

$$\Delta(\Gamma_1, \Gamma_2) = \frac{|\{\{x, y\} \subset X \mid (x \sim_{\Gamma_1} y) \oplus (x \sim_{\Gamma_2} y)\}|}{\binom{|X|}{2}}, \tag{6.12}$$

donde el símbolo $\oplus$ es el operador lógico XOR.

En el numerador se están contando los pares de datos que fueron agrupados en un mismo clúster en alguno de los clústers Γ_1, Γ_2, pero no en el otro. Es una medición de la divergencia en la clasificación. El denominador cuenta la cantidad total de pares $\{x, y\}$ posibles. En definitiva, $\Delta(\Gamma_1, \Gamma_2)$ devuelve la proporción total de divergencias; es un número entre 0 y 1.

Definición 6.3.2. (**Restricción a un clustering**) Sea $X \subset Z$ y $\Gamma = \{C_1, C_2, \ldots, C_k\}$ un clustering de Z. Entonces se define la restricción del clustering Γ al subconjunto X de la siguiente forma:

$$\Gamma|X = \{C_1 \cap X, C_2 \cap X, \ldots, C_k \cap X\}. \tag{6.13}$$

Definición 6.3.3. (δ-**robustez**) Sea f una función de clustering fija. Sean los siguientes conjuntos X,Y y O de tal forma que $X = Y \cup O$ y $Y \cap O = \emptyset$. En este contexto, O representa a los datos atípicos y Y los datos comunes. X es la reunión de ambos. Decimos que el dataset Y es δ-robusto con respecto a los datos atípicos O y con respecto a F si:

$$\Delta(f(Y), f(X)|Y) \leq \delta. \tag{6.14}$$

Cuando el algoritmo además requiere como input la cantidad de clústers k, entonces se dice en ese caso que Y es δ-robusto con respecto a O, f y k. Si la cantidad de clústers es clara por el contexto, se omite k en la nomenclatura.

Es importante observar que menores valores de δ significan que los datos O tienen menos influencia en el clustering final.

Definición 6.3.4. (α**-separable**). Un clustering Γ de X es α-separable si hay un $\alpha > 0$ tal que para todos los puntos: x_1, x_2, x_3, x_4 tales que $x_1 \sim_\Gamma x_2$ y $x_3 \nsim_\Gamma x_4$ se tiene:

$$\alpha \cdot d(x_1, x_2) < d(x_3, x_4).$$

Definición 6.3.5. (β**-balanceado**). Un clustering $\Gamma = \{C_1, C_2, \ldots, C_k\}$ de X está β-balanceado si:

$$\forall i \in \{1, 2, \ldots, k\} \ |C_i| \leq \beta |X|.$$

Notar que $\frac{1}{k} \leq \beta \leq 1$ *y si* $\beta = \frac{1}{k}$ *entonces el clustering está perfectamente balanceado.*

Definición 6.3.6. (**Bien clusterizable**). Un dataset X con una distancia entre datos d es *bien clusterizable* si existe un clustering Γ de X tal que sea α-separable para algún α grande, como mínimo $\alpha \geq 1$.

Teorema 6.6. *Para todo* $\alpha \geq 1$*, si en un dataset* X *existe un clustering* α*-separable este es único.*

6.3.2. Principales resultados

Teorema 6.7. *Consideremos* X, Y *y* O *tales que* $X = Y \cup O$ *y* $Y \cap O = \emptyset$*. Además, supongamos que* Y *admite un* k*-clustering* α*-separable y* β*-balanceado, con diámetro* $s \in (0, 1]$*. Entonces:*

1. *Si* f *es* k*-medias entonces* Y *es* δ*-robusto para:*

$$\delta \leq \frac{8}{\alpha^2}(1 + \frac{|O|}{|Y|s^2}) + 2k \cdot \beta^2.$$

2. *Si* f *es* k*-medianas o* k*-medoids entonces* Y *es* δ*-robusto para:*

$$\delta \leq \frac{4}{\alpha}(1 + \frac{|O|}{|Y|s}) + 2k \cdot \beta^2.$$

Aquí el diámetro s *es la máxima distancia entre dos puntos en* Y*.*

Definición 6.3.7. (**Detección de** α**-separables**) Una función de k-clustering f detecta clústers α-separables para algún $\alpha \geq 1$ si para todo dataset (X, d) que admite un k-clustering C α-separable entonces $f(X, d, |C|) = C$.

Un ejemplo es la función *single-linkage* que detecta clústers 1-separables.

Teorema 6.8. *Sea f una función de k-clustering que detecta clusterings α-separables para algún $\alpha \geq 1$. Entonces para todo $\beta \in [\frac{1}{k}, 1]$, $s \in [0, \frac{1}{\alpha+1})$ y para todo entero $m \geq k-1$, existe un dataset $X = Y \cup O$ tal que $|O| \leq m$, Y admite un k-clustering α-separable y β-balanceado con diámetro s pero que Y no es siquiera $\beta(k-1)$-robusto con respecto a O y a f.*

Aquí el diámetro s es la máxima distancia entre dos puntos de un mismo clúster.

Algunos comentarios sobre los teoremas 6.7 y 6.8 El teorema 6.7 significa que las funciones de k-medias, k-medianas y k-medoids son robustas; bajo ciertas condiciones que son las especificadas en los puntos 1 y 2. Mientras que el teorema 6.8 muestra que las funciones que detectan clusterings α-separables no son robustas; incluso son muy sensibles a los datos atípicos.

Para el teorema 6.7, suponga que se dispone de un conjunto de datos Y perfectamente balanceado, esto es, $\beta \approx \frac{1}{k}$. Suponga además que la cantidad de datos atípicos es muy pequeña en comparación con la cantidad de datos de Y, o sea $\frac{|O|}{|Y|} \approx 0$. Entonces quedaría aproximadamente:

$$\delta \leq \frac{8}{\alpha^2} + \frac{2}{k}. \tag{6.15}$$

Por lo tanto, para un α mediano ($\alpha \geq 8$) y un k también mediano ($k \geq 8$) se cumple que $\delta \leq \frac{1}{2}$; como se ve, este es un resultado asintótico. k-medias es robusto cuando es muy separable el clustering y son bastantes los clústers que se desean obtener; al menos, según este resultado.

Sobre el teorema 6.8, la hipótesis lo que aclara es que para toda función de clustering que detecta clusterings α-separables aunque se esté en la situación óptima de balance ($\beta = \frac{1}{k}$), diámetro pequeño ($s \approx 0$) y pocos datos atípicos ($|O| \leq k-1$), aún así Y no es siquiera $\frac{k-1}{k}$-robusto con respecto a O y a F. Esto quiere decir que aún en las condiciones óptimas, algoritmos como single-linkage son muy sensibles a pocos datos atípicos.

A primera vista, ambos resultados parecen entrar en tensión cuando se aplican al algoritmo de k-medias. En efecto, Ackerman demuestra que k-medias detecta clusterings 2-separables, por lo que el Teorema 6.8 parecería aplicársele. Sin embargo, el Teorema 6.7 afirma que, bajo condiciones ideales de separabilidad y balance, el algoritmo es robusto frente a conjuntos pequeños de datos atípicos. No he logrado identificar con claridad el punto

exacto donde ambas afirmaciones dejan de ser comparables, por lo que dejo planteada esta observación y como referencias el capítulo 6 de su tesis, especialmente las secciones 6.2 y 6.4.1.

6.4. Funciones Linkage-based y su caracterización

Por último, ella realiza una caracterización para cierto tipo de funciones denominadas *linkage-based* (este concepto será definido con precisión más adelante). Alguna de estas ya fueron presentadas anteriormente: single-linkage, average-linkage, complete-linkage, etc. Ella, en los capítulos 4 y 7, responde a dos preguntas: qué propiedades caracterizan a las funciones linkage-based *en tanto funciones de k-clustering* (capítulo 4) y la segunda es: qué propiedades caracterizan a las funciones linkage-based *en tanto funciones jerárquicas* (capítulo 7); un concepto que también se verá más adelante. Por lo tanto, esta sección está dividida en dos: 1. Caracterización de funciones linkage-based en general. 2. Caracterización de funciones linkage-based como funciones jerárquicas.

6.4.1. Caracterización de funciones linkage-based en general

Antes de ver el resultado principal, se debe definir **qué significa ser una función de k-clustering linkage-based**. Para esto, es necesario definir el siguiente concepto:

Definición 6.4.1. (**Función de ligado**) Una función de ligado es una función:

$$\ell : \{(X_1, X_2, d) \mid d \text{ es una función de distancia en } X_1 \cup X_2\} \to \mathbb{R}^+, \quad (6.16)$$

tal que:

1. ℓ es **independiente por representación**; esto es, si $(\{X_1, X_2\}, d) \cong (\{X_1', X_2'\}, d')$ entonces $\ell(X_1, X_2, d) = \ell(X_1', X_2', d')$. Aquí $\cong$ significa que son isomorfos como clústers.

2. ℓ es **monótona al aumentar la distancia extraconjuntos**. Para todo (X_1, X_2, d) si d' es una distancia sobre $X_1 \cup X_2$ tal que si $x \sim_{\{X_1, X_2\}} y$ entonces $d(x, y) = d'(x, y)$ y si $x \nsim_{\{X_1, X_2\}} y$ entonces $d(x, y) \leq d'(x, y)$; luego se tiene que $\ell(X_1, X_2, d') \geq \ell(X_1, X_2, d)$.

3. ℓ permite **desligar dos conjuntos todo lo posible**. Para todo par de datasets $(X_1, d_1), (X_2, d_2)$ y para cualquier número r en el rango de ℓ, existe una función de distancia d que extiende a ambas distancias d_1 y d_2 tal que $\ell(X_1, X_2, d) > r$.

Definición 6.4.2. (**Ser linkage-based**) Una función F de k-clustering es linkage-based si existe una función de ligado ℓ tal que:

1. $F(X, d, |X|) = \{\{x\} \, : \, x \in X\}$.
2. Para $1 \leq k < |X|$ se tiene que $F(X, d, k)$ es el resultado de unir los dos clústers en $F(X, d, k+1)$ que minimizan ℓ. Esto es,

$$F(X, d, k) = \{C_i \, : \, C_i \in F(X, d, k+1),\, C_i \neq C_1,\, C_i \neq C_2\} \cup \{C_1 \cup C_2\}, \tag{6.17}$$

siendo $\{C_1, C_2\} = \arg\min_{\{C_1, C_2\} \subset F(X,d,k+1)} \ell(C_1, C_2, d)$.

Observación: Notar que las funciones linkage-based que se comentaron en el capítulo 3, en 3.2.1, cumplen esta definición con las siguientes funciones de ligado:

1. *Single linkage:* $\ell_{SL}(A, B, d) = \text{mín}_{a \in A, b \in B}\, d(a, b)$.
2. *Average linkage:* $\ell_{AL}(A, B, d) = \frac{\sum_{a \in A, b \in B} d(a,b)}{|A| \cdot |B|}$.
3. *Complete linkage:* $\ell_{CL}(A, B, d) = \text{máx}_{a \in A, b \in B}\, d(a, b)$.

Ahora, es posible enunciar el resultado principal:

Teorema 6.9. *Toda función de k-clustering invariante por escala es linkage-based si y solamente si es refinante, exteriormente consistente, local y exteriormente abundante.*

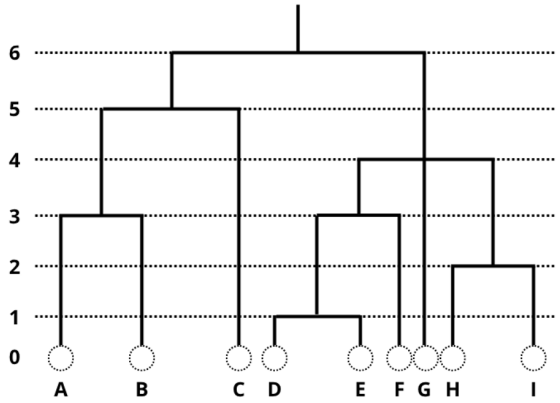
La demostración se encuentra en [12].

6.4.2. Caracterización de funciones linkage-based como funciones jerárquicas

Dentro del capítulo siete se profundiza en las funciones jerárquicas. Se ha dado una breve definición de ellas en el capítulo tres, pero Ackerman da una definición más explícita de qué es una función de clustering jerárquica. La idea está en que toda función de clustering jerárquica es capaz de generar un dendrograma en función de la cantidad de clústers k que se desea.

Definición 6.4.3. (**Dendrograma**) Un dendrograma sobre (X, d) es la tríada (T, M, η) donde T es un árbol binario, M : hojas$(T) \to X$ una biyección y $\eta : \mathrm{V}(T) \to \{0, 1, \ldots, h\}$ sobreyectivo, con h un entero positivo. Aquí $\mathrm{V}(T)$ son los nodos del grafo y $\mathrm{E}(T)$ sus links. Y se debe cumplir además:

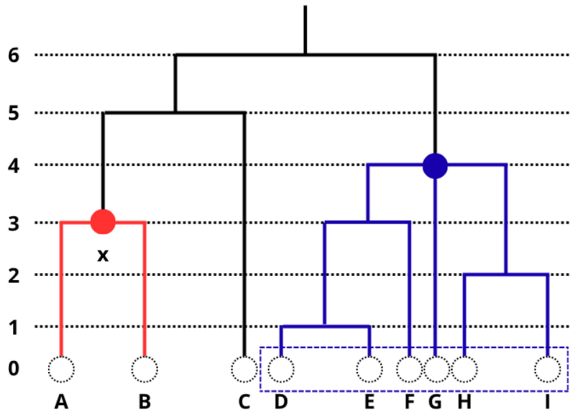
1. Para toda hoja $x \in \mathrm{V}(T)$ se tiene $\eta(x) = 0$.
2. Si $(x, y) \in \mathrm{E}(T)$ entonces $\eta(x) > \eta(y)$.



Aquí las letras A, B, C, etc; representan los datos y los números los niveles de η.

Definición 6.4.4. (**Clúster asociado a un nodo**) Para todo nodo $x \in \mathrm{V}(T)$ se denota $C(x) = \{M(y) : y \text{ es una hoja descendiente de } x\}$; este es el clúster asociado al nodo x.

Definición 6.4.5. (**Clúster en un dendrograma**) Se dice que A es un clúster en (T, M, η) si hay un nodo $x \in \mathrm{V}(T)$ tal que $A = C(x)$.



1. Se puede ver que el clúster asociado al nodo x es $C(x) = \{A, B\}$. 2. $\{D, E, F, G, H, I\}$ es un clúster, porque hay un nodo del cual todos son descendientes comunes.

Observación: como un dendrograma es un árbol, entonces se sigue que si A es un clúster en (T, M, η) hay un *único* x que cumple la propiedad anterior. Por lo que es posible hacer la siguiente definición:

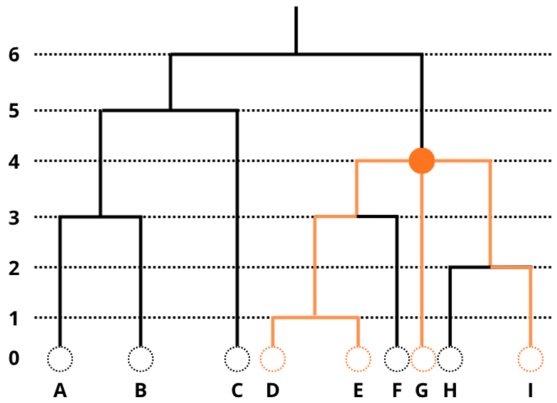
Definición 6.4.6. (**Nodo asociado a un clúster.**) Sea A un clúster en (T, M, η), y x el único nodo que cumple: $C(x) = A$, entonces este x es el nodo asociado a A y se denotará $\nu(A) = x$.

Definición 6.4.7. Un clustering $\Gamma = \{C_1, \ldots, C_k\}$ de $X' \subset X$ está en (T, M, η) si $\forall i\, C_i$ está en (T, M, η).

Es importante notar del hecho $\forall i \neq j\ C_i \cap C_j = \emptyset$ se sigue que ni $\nu(C_i)$ ni $\nu(C_j)$ son uno descendientes del otro.
En la imagen anterior $\{\{D, E, F, G, H, I\}, \{A, B\}\}$ es un clustering que está en el dendrograma.

Definición 6.4.8. (**Subdendrograma**) Un subdendrograma de (T, M, η) es una triada (T', M', η') tal que:

1. T' es un subárbol de T con raíz en $x \in V(T)$.
2. $\forall y \in$ hojas(T') se tiene $M(y) = M'(y)$, es decir M' es la misma función restricta al subconjunto hojas(T).
3. $\forall y, z \in \mathrm{V}(T')$ se tiene que $\eta'(y) < \eta'(z)$ si y solamente si $\eta(y) < \eta(z)$.



En color naranja se muestra un ejemplo de subdendrograma.

Definición 6.4.9. (**Isomorfismo de dendrogramas**) Dos dendrongramas (T_1, M_1, η_1) de (X, d) y (T_2, M_2, η_2) de (X', d') son isomorfos, y se denotará $(T_1, M_1, \eta_1) \cong_{\mathcal{D}} (T_2, M_2, \eta_2)$, si hay dos biyecciones: $\phi : X \to X'$ y $H : \mathrm{V}(T_1) \to \mathrm{V}(T_2)$ tales que:

1. $\forall x, y \in \mathrm{V}(T_1)\ (x, y) \in \mathrm{E}(T_1) \iff (H(x), H(y)) \in \mathrm{E}(T_2)$.
2. $\forall x \in \mathrm{V}(T_1),\ \eta_1(x) = \eta_2(H(x))$ (En realidad, aquí Ackerman pide una condición que es relajable: se podría pedir que simplemente se respetase el orden, esto es: $\forall x, y \in \mathrm{V}(T_1),\ \eta_1(x) < \eta_1(y) \iff \eta_2(H(x)) < \eta_2(H(y))$).
3. $\forall x \in$ hojas$(T_1),\ \phi(M_1(x)) = M_2(H(x))$.

Ahora es posible ver la definición de Ackerman para algoritmos jerárquicos:

Definición 6.4.10. (**Algoritmo jerárquico**) Una función de clustering jerárquico F es una función que recibe como input el par (X, d) y devuelve un dendrograma (T, M, η) y además:

1. Es independiente por representación: Es decir, si $(X, d) \cong (X', d')$ se tiene que $F(X, d) \cong_{\mathcal{D}} F(X', d')$.

2. Invariante por escala: $F(X, d) = F(X, c{\cdot}d)$ para toda constante positiva c.

3. Abundancia: para datasets $\{(X_1, d_1), \ldots, (X_k, d_k)\}$ disjuntos dos a dos, hay una distancia d' sobre $\cup_{i=1}^{k} X_i$ que extiende a todas las distancias tal que $\{X_1, \ldots, X_k\}$ es un clustering de $F(\cup_{i=1}^{k} X_i, d')$.

Definido algoritmo jerárquico, se pasará a ver la definición de algoritmo jerárquico linkage-based, para luego ver el resultado principal y terminar con un ejemplo de un algoritmo jerárquico que no es linkage-based.

Definición 6.4.11. (**Algoritmo jerárquico linkage-based**) Una función de clustering jerárquico F es linkage-based si existe una función de ligado tal que:

$$\forall (X, d), \quad F(X, d) = (T, M, \eta),$$

en donde

$$\eta(\text{parent}(A, B)) = m$$
$$\Updownarrow$$
$$\ell(A, B) \text{ es mínimo en } \left\{ \ell(S, T) \;\middle|\; \begin{array}{l} S \cap T = \emptyset, \\ \eta(\nu(S)) < m, \quad \eta(\nu(T)) < m, \\ \eta(\text{parent}(S)) \geq m, \quad \eta(\text{parent}(T)) \geq m \end{array} \right\}.$$

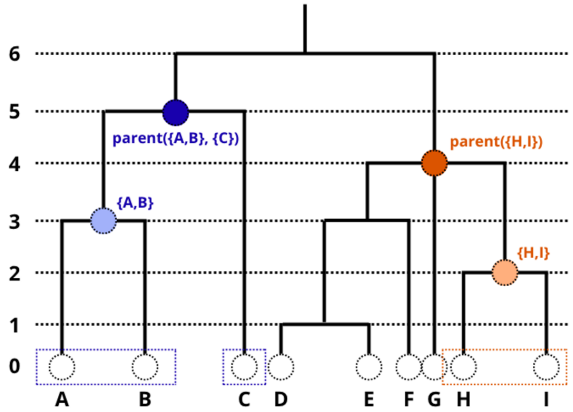
Algunas observaciones:

1. Para la función de ligado ir a la definición 8.4.1; según la autora, en este caso puede relajarse la definición exigiendo solamente las primeras dos propiedades.

2. Dado un clúster A, parent(A) es el único nodo tal que $(\text{parent}(A), \nu(A)) \in \text{E}(T)$; visualmente, en el gráfico sería el nodo superior al nodo asociado al clúster.

3. Si dados dos clústers A y B tales que parent$(A) =$ parent(B) entonces a ese nodo se le denotará parent(A, B). En dicho caso, el significado es que ambos clústers serán agrupados en un mismo clúster, juntos e inmediatamente; en ningún momento anterior a su agrupamiento alguno de ellos se agrupará con a tercero.

4. Caso contrario al anterior se denota parent$(A, B) = \emptyset$.



(1)En azul puede observarse un ejemplo de parent *para dos clústers,* $\{A, B\}$ *y* $\{C\}$*. (2) En naranja se muestra un ejemplo de* parent *para un clúster,* $\{H, I\}$*. (3) En este ejemplo* parent$(\{A, B\}, \{H, I\}) = \emptyset$.

5. La definición de algoritmo jerárquico linkage-based significa lo siguiente: el algoritmo en un paso t junta dos clústers A y B en un único clúster en un paso $t + 1$ si y solamente si $\ell(A, B)$ es el mínimo valor de todos los posibles para los pares de clúster en el paso t.

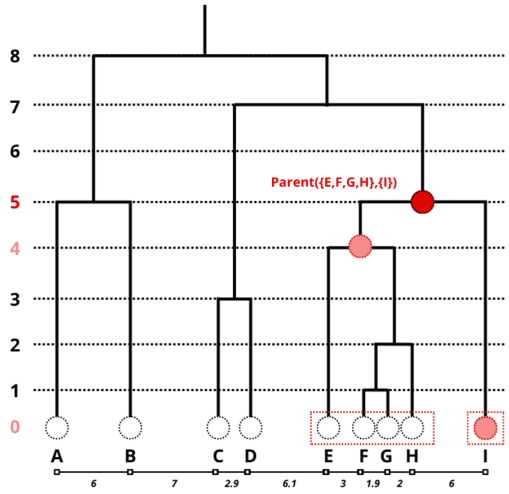


Imagen ilustrativa de un dendrograma correspondiente a un algoritmo single-linkge. Los números debajo de los datos representan la distancia entre ellos; en el caso unidimensional. Note que se cumple el punto 4 y la definición.

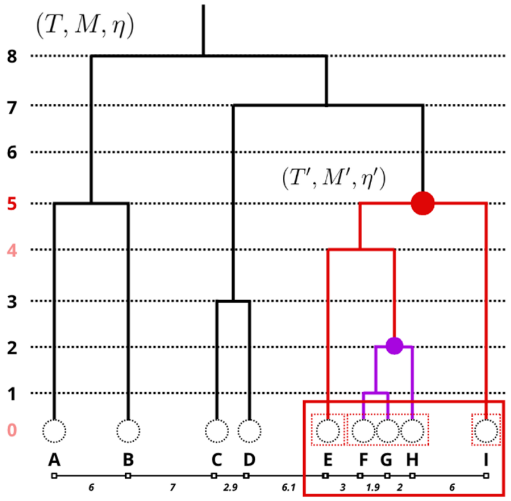
6. Notar que el mínimo puede no ser único y la definición no imposibilita al algoritmo de mezclar dos pares de clústers a la vez en un solo paso (ver la imagen anterior). Aunque, si consideramos el caso de que todas las distancias sean distintas (como ocurriría para datos con distribución continua casi seguramente), entonces, ahí sí, el algoritmo mezcla solamente un par de clústers en cada paso.

Antes de ver el resultado principal es necesario revisar algunas propiedades ya vistas pero adaptadas al contexto de clustering jerárquico.

La primera de ellas es la *localidad*. Localidad, en este contexto, viene a significar que si se selecciona un clustering dentro del dendrograma y se reinicia el algoritmo únicamente con los datos de ese clustering entonces se obtendría un resultado consistente con el dendrograma original.

Definición 6.4.12. (**Localidad**) Una función de clustering jerárquica F es local si para todo dataset (X, d) y $X' \subset X$, siempre que $\Gamma = \{C_1, C_2, \ldots, C_k\}$ sea un clustering de X' que está en $F(X, d) = (T, M, \eta)$ entonces se cumple para todo $i \in \{1, 2, \ldots, k\}$ lo siguiente:

1. El clúster C_i está en $F(X', d_{|_{X'}}) = (T', M', \eta')$.
2. El subdendrograma de $F(X, d)$ con raíz en $\nu(C_i)$ es también un subdendrograma de $F(X', d_{|_{X'}})$ con raíz en $\nu(C_i)$.
3. Para todos $x, y \in X'$ $\eta'(x) < \eta'(y) \iff \eta(x) < \eta(y)$.



Ejemplo de localidad. *(1)* $X' = \{E, F, G, H, I\}$ *y un clustering de esos datos es* $\{\{E\}, \{F, G, H\}, \{I\}\}$. *(2) Notar que dicho clustering está en* (T, M, η).*(3)Para la*

primera propiedad, vea que cada uno de esos clústers está en (T', M', η')*.(4) Un ejemplo de la segunda propiedad puede verse en el subdendrograma de color violeta; correspondiente a los datos* $\{F, G, H\}$*.*

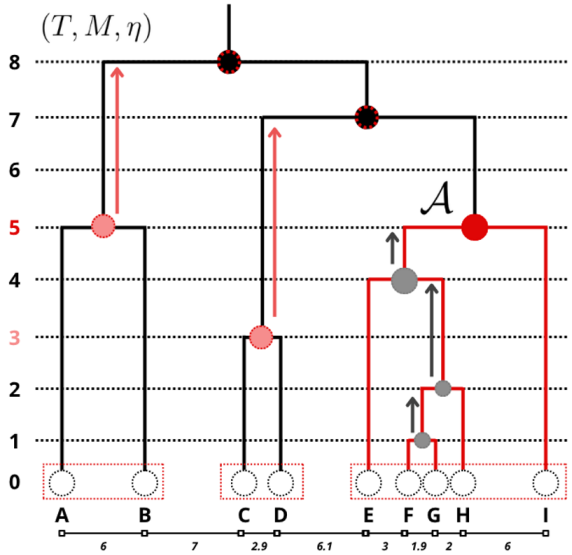
Según la autora, esta propiedad en algunas situaciones es muy ideal, principalmente en contextos donde se cumple en la propia realidad esta localidad; por ejemplo, cuando se estudian mutaciones de ciertos organismos vivos.

Antes de ver el resultado principal, es necesario revisar primero la versión adaptada de la propiedad de consistencia externa. Para ello se requiere la siguiente definición previa:

Definición 6.4.13. ($\mathcal{A}$-corte) Dado un clúster $\mathcal{A}$ en un dendrograma (T, M, η), el $\mathcal{A}$-corte del dendrograma es el siguiente conjunto de clústers:

$$\{C(x) \mid x \in \mathrm{V}(T),\ \eta(x) \leq \eta(\nu(\mathcal{A})) < \eta(\mathrm{parent}(x))\}.$$

A este conjunto se lo denota $cut_{\mathcal{A}}((T, M, \eta))$.



Ejemplo de $\mathcal{A}$-corte. *(1) En el pie de la imagen puede verse el $\mathcal{A}$-corte con respecto a $\mathcal{A} = \{E, F, G, H, I\}$, que corresponde al clustering: $cut_{\mathcal{A}}((T, M, \eta)) = \{\{A, B\}, \{C, D\}, \{E, F, G, H, I\}\}$. (2) En rosado pueden verse los vértices que cumplen que su nivel es menor o igual al de $\mathcal{A}$ y su* parent *tiene un nivel estrictamente superior. (3) En gris se muestran ejemplos de vértices que no cumplen dicha propiedad.*

Observaciones:

1. Notar que este clustering corresponde a los clústers que están al mismo nivel de $\mathcal{A}$ o directamente más abajo que $\mathcal{A}$. Es decir, ***son los clústers que hay en el momento en el que $\mathcal{A}$ fue formado.***

2. Esto es un clustering del dataset. En efecto, si existiesen x, y tales que $C(x) \cap C(y) \neq \emptyset$ entonces o $\eta(x) < \eta(y)$ o $\eta(x) = \eta(y)$ o $\eta(x) > \eta(y)$. En el primer caso $\eta(parent(x)) \leq \eta(y) \leq \eta(\nu(A)) < \eta(parent(x))$ que es una contradicción. La misma contradicción acontece en el tercer caso. El segundo caso es imposible, porque desde cualquier dato de $C(x) \cap C(y)$ es posible ir subiendo en el árbol y necesariamente hay de pasar por x y por y y siempre se pasa antes por uno que por el otro, y eso implica, por la definición de dendrograma, que no pueden estar al mismo nivel. Que todo dato está en algún clúster se prueba de la siguiente forma: sea x_0 un dato del dataset, luego $\eta(x_0) = 0$. Sea x_r el nodo raíz del dendrograma, luego hay una sucesión:

   $$x_0, x_1 = parent(x_0), x_2 = parent(x_1), \ldots, x_r = parent(x_n),$$

   para algún n.

   Y siempre $\eta(x_k) < \eta(x_{k+1})$. Por lo que existe un único momento, sea este l, tal que $\eta(x_l) \leq \eta(\nu(A)) < \eta(x_{l+1})$. Luego $x_0 \in C(x_l)$.

Definición 6.4.14. (**Consistencia externa**) Una función de clustering jerárquico F es exteriormente consistente si para todo dataset (X, d) y para cualquier clúster $\mathcal{A}$ en $F(X, d)$ se cumple lo siguiente: si d' es una distancia $(cut_{\mathcal{A}}(F(X, d)), d)$-exteriormente consistente, entonces:

$$cut_{\mathcal{A}}(F(X, d)) = cut_{\mathcal{A}}(F(X, d')). \tag{6.18}$$

Explicación de la definición:
Considere un clúster $\mathcal{A}$ que esté en $F(X, d)$, $cut_{\mathcal{A}}(F(X, d))$ significa los clústers que hay en el momento de su formación. d' es una nueva distancia que aleja a dichos clústers entre sí preservando las distancias originales intraclústers, en otras palabras, es una distancia $(cut_{\mathcal{A}}(F(X, d)), d)$-consistente. F exteriormente consistente significa que si el algoritmo vuelve a iniciarse con esos mismos datos pero con la nueva distancia d' entre ellos entonces $F(X, d')$ es un dendrograma en el que:

1. $\mathcal{A}$ es un clustering en $F(X, d')$.

2. Los clústers que hay al momento de la formación de $\mathcal{A}$ en $F(X, d')$ son los mismos que había antes en $F(X, d)$. Esto es lo que significa la ecuación 6.18.

Resultado principal

Teorema 6.10. *(**Caracterización de funciones linkage-based**)*
Una función de clustering jerárquica F es linkage-based si y solamente si F es exteriormente consistente y local.

La demostración de este resultado se puede encontrar en [12].

6.4.3. Funciones jerárquicas disociativas o divisivas

La diferencia principal de los algoritmos linkage-based con los disociativos es que los primeros toman solamente consideraciones locales mientras los segundos, inevitablemente, consideran más la totalidad de los datos. Esta diferencia esencial puede ayudar a la hora de seleccionar de entre los algoritmos jerárquicos cuál elegir. Estas diferencias fueron descritas por la autora:

Definición 6.4.15. (**Función de 2-clustering**) Una función $\mathcal{F}$ es de 2-clustering si para cada dataset (X, d) no unitario devuelve un clustering con dos clústers. Es decir, divide en dos partes los datos.

Definición 6.4.16. (**Algoritmo disociativo**) Un algoritmo jerárquico de clustering F es $\mathcal{F}$-disociativo con respecto a una función de 2-clustering $\mathcal{F}$ si, para todo dataset (X, d), $F(X, d) = (T, M, \eta)$ cumple que para todo nodo $x \in V(T)$, que no sea una hoja, sus hijos x_1, x_2 cumplen $\mathcal{F}(C(x)) = \{C(x_1), C(x_2)\}$.

Esta definición, nota la autora, no impone una restricción muy severa sobre la función de nivel η, por lo tanto no fuerza un orden en la división de los clústers.

Definición 6.4.17. (**Sensible al contexto**) Una función $\mathcal{F}$ de 2-clustering es sensible al contexto si:

$$\mathcal{F}(\{x, y, z\}, d) = \{\{x\}, \{y, z\}\} \quad \text{y} \quad \mathcal{F}(\{x, y, z, w\}, d') = \{\{x, y\}, \{z, w\}\},$$

para algún par de datasets $(\{x, y, z\}, d)$ y $(\{x, y, z, w\}, d')$ tal que d' sea una extensión de la distancia d.

Una de las diferencias entre las funciones jerárquicas asociativas, estilo single-linkage, y las disociativas está en la localidad. Algunas funciones disociativas por su naturaleza sensible al contexto no son locales; esto es lo que dice el siguiente teorema:

Teorema 6.11. *Si $\mathcal{F}$ es una función de 2-clustering sensible al contexto entonces toda función de clustering jerárquico $\mathcal{F}$-disociativa no es local.*

Demostración. Como $\mathcal{F}$ es sensible al contexto existen dos datasets como en la definición anterior: $(\{x, y, z\}, d)$ y $(\{x, y, z, w\}, d')$ tal que d' es una extensión de la distancia d; de modo que se cumple:

$$\mathcal{F}(\{x, y, z\}, d) = \{\{x\}, \{y, z\}\} \quad \text{y} \quad \mathcal{F}(\{x, y, z, w\}, d') = \{\{x, y\}, \{z, w\}\}.$$

Sea F una función de clustering jerárquica $\mathcal{F}$-disociativa. Entonces el siguiente clustering $\{\{x, y\}, \{z\}\}$ es un clustering de $\{x, y, z\}$ en el dendrograma generado por $F(\{x, y, z, w\}, d')$ (recordar la definición 8.4.7), pero no es un clustering en $F(\{x, y, z\}, d)$, luego el algoritmo no es local. □

Corolario 6.11.1. *Ninguna función de clustering jerárquica $\mathcal{F}$-disociativa con $\mathcal{F}$ sensible al contexto es linkage-based.*

Definición 6.4.18. (**Divergencia fuerte**) Dos algoritmos de clustering jerárquicos F_0 y F_1 divergen fuertemente si existe un dataset (X, d) y un clustering Γ de X tal que Γ está en $F_i(X, d)$ pero no en $F_{1-i}(X, d)$, con $i \in \{0, 1\}$.

Teorema 6.12. *(**Divergencia entre algoritmos asociativos linkage-based y disociativos**) Todo algoritmo jerárquico disociativo sensible al contexto diverge fuertemente de cualquier algoritmo jerárquico linkage-based.*

Demostración. Supóngase que F_0 es $\mathcal{F}$-disociativa sensible al contexto y F_1 linkage-based. Como F_0 es sensible al contexto existen dos datasets como en la definición anterior: $(\{x, y, z\}, d)$ y $(\{x, y, z, w\}, d')$ tales que d' es una extensión de la distancia d; de modo que se cumple:

$$\mathcal{F}(\{x, y, z\}, d) = \{\{x\}, \{y, z\}\} \quad \text{y} \quad \mathcal{F}(\{x, y, z, w\}, d') = \{\{x, y\}, \{z, w\}\}.$$

Tómese como clustering de $\{x, y, z, w\}$ a $\Gamma = \{\{x, y\}, \{z, w\}\}$; hay, por lo tanto, dos opciones: o no está en $F_1(\{x, y, z, w\}, d')$ y queda demostrado el teorema (recordar que basta encontrar un dataset y un clustering del mismo) o lo está. Si está sabemos que $\{\{x, y\}, \{z\}\}$ es un clustering que está en

$F_0(\{x, y, z, w\}, d')$, si Γ también estuviese en $F_1(\{x, y, z, w\}, d')$ tendría que ocurrir que $\{\{x, y\}, \{z\}\}$ estuviese también en $F_1(\{x, y, z, w\}, d')$. Como F_1 es un algoritmo linkage-based, es local, luego $\{\{x, y\}, \{z\}\}$ también está en $F_1(\{x, y, z\}, d)$, pero no está en $F_0(\{x, y, z\}, d)$. □

Corolario 6.12.1. *Todo algoritmo jerárquico disociativo basado en las siguientes funciones de* 2*-clustering diverge fuertemente con cualquier algoritmo jerárquico linkage-based: k-medias, min-sum,min-diameter.*

Demostración. Es simplemente mostrar que son funciones de 2-clustering sensibles al contexto. Para $x = 1$, $y = 3$, $z = 4$ y $w = 6$ se puede verificar con facilidad en $\mathbb{R}$ con la distancia usual. □

Parte III

Conclusiones

Capítulo 7

Propuesta de fundamentación

7.1. Clasificación en general

A lo largo de este documento se han visto diversos algoritmos y aplicaciones del clustering (parte primera), y luego diversos intentos de establecer una teoría general del clustering que dé unidad al área de estudio; en especial ha habido diversos intentos de establecer una lista de axiomas que concreticen matemáticamente qué es una función de clustering. Este intento sigue siendo, según Ackerman en su tesis, un problema abierto.

El objetivo de este capítulo es abordar dicho problema. No se pretende dar una solución definitiva, pero sí traer luz a la temática. Estoy convencido de la importancia de partir de principios e ideas claras, eliminando términos equívocos que den lugar a ambigüedades. Por eso, antes de abordar la axiomatización del clustering, y problemática afín, se principiará el análisis desde el fundamento para, a partir de él, construir el edificio lógico. Este fundamento inicia con la siguiente pregunta: ¿Qué es la clasificación como ciencia estadística y matemática?

7.1.1. Objeto material y formal

En general

Para responder la pregunta anterior se harán algunas definiciones que pertenecen al campo de la lógica filosófica, especialmente de la lógica material. Sobre esta temática se puede consultar el primer capítulo de la cuarta parte de [38]. Estas definiciones son de gran ayuda para distinguir a las ciencias entre sí y qué estudian, por ejemplo: la diferencia entre la antropología, la medicina y la zoología; la diferencia entre la física, la química y la astronomía; etc.

Definición 7.1.1. (Objeto material de una ciencia) El objeto material de una ciencia es el ente estudiado. La cosa sobre la que recae el estudio.

Volviendo a los ejemplos anteriores: tanto la antropología como la medicina tienen el mismo objeto material: el hombre; en cambio el de la zoología son todos los animales. La física y la química tienen el mismo objeto material: los entes materiales, pero la astronomía solo contempla a los astros del espacio.

Como puede verse, el objeto material no es el criterio que hace específica cada ciencia. ¿Qué es lo que las diferencia? Su objeto formal.

Definición 7.1.2. (Objeto formal de una ciencia) El objeto formal de una ciencia es el aspecto bajo el cual se considera el objeto material, o el punto de vista específico de una ciencia. El aspecto de la cosa sobre la que recae el estudio.

Por ejemplo, el objeto formal de la medicina es la salud del hombre, o el hombre en cuanto tiene salud o está enfermo. El objeto formal de la antropología es el hombre en la totalidad de sus aspectos, en los primeros principios y causas últimas del hombre, el hombre en tanto hombre. La física estudia los seres materiales en tanto pasibles de cambios accidentales: como traslación, rotación, temperatura, etc; mientras la química estudia sus cambios sustanciales: como los procesos para transformar agua en hidrógeno y oxígeno, o cómo se forman ciertas sustancias a partir de otras, etc. Ya la zoología tiene una diferencia, no en el objeto formal, sino en el material con la antropología y la medicina; así como la astronomía se diferencia en el objeto material con la física y la química.

De la clasificación estadística

Si se llama clasificación estadística y matemática, o simplemente clasificación, a todos los métodos de clasificación, ya sean supervisados o no supervisados, es decir, al problema de agrupar datos cuantitativos en general; se ve, a la luz de lo anterior, cuáles son el objeto material y formal. El **objeto material** de la clasificación son *los datos cuantitativos* y el **objeto formal** es *en tanto son agrupables o clasificables*. Por lo tanto, la clasificación como ciencia dentro de la estadística matemática estudia los métodos de clasificación de los datos y las propiedades de dichos datos en tanto capaces de ser clasificados, ya sea porque hay un dataset de ejemplo o porque hay una matriz de similitudes de dichos datos. Esto lleva a la siguiente conclusión:

No existe la clasificación no supervisada per se en el sentido de una clasificación desprovista de toda noción previa de agrupabilidad. Esto es, no es posible estudiar cómo clasificar datos numéricos -ni establecer principios metodológicos o axiomáticos- sin ningún tipo de supervisación; es decir, sin ningún dato o información que de cierta forma sea la razón de su clasificación, porque toda clasificación presupone algún principio formal de agrupabilidad. Puede observarse en todos los algoritmos y métodos vistos en la primera y segunda parte que, siempre y sin excepción, se parte de algún dato que se lo considera un criterio para clasificar. **En el capítulo 2**, se presentaron tres tipos de aplicaciones distintas: reconocimiento de imágenes, detección de zonas sísmicas y detección de comunidades. En reconocimiento de imágenes, el método de clasificación tiene como *criterio para clasificar* una propiedad geométrica: que los datos formen un círculo, elipse o recta. Cuando se buscaron zonas sísmicas, *la razón que hacía a los datos agrupables* era su cercanía, más alguna consideración geométrica. Para el caso de detección de comunidades, *la razón* era una cierta noción de similitud, ya sea que eran amigos en una red social, o cierta regularidad entre equipos de fútbol. **En el capítulo 3** puede observarse que los métodos basados en centros toman en consideración una cierta noción de distancia (o disimilitud), y en algunos casos, como Mahalanobis, se toman en consideración aspectos geométricos. Los algoritmos jerárquicos también toman en cuenta una noción de similitud o disimilitud. Lo mismo ocurre para el caso de Spectral Clustering. En cambio, en los métodos basados en modelos probabilísticos la *razón* de agrupación de los datos cambia. Para el caso de mixtura de distribuciones, la *razón* es que los datos provienen de una cierta distribución, y es por esa razón que el método consiste en estimar la distribución. Los clusterings basados en modas y estimación de conjuntos de nivel basan su criterio en el principio de que los clústers corresponden a modas en una cierta distribución de la que provienen. Por eso mismo el método consiste en estimar la función de densidad. **En la segunda parte de la tesis**, en todos los capítulos que se expusieron los métodos basan su **razón (o criterio)** para clasificar en la similitud (o disimilitud, también llamada distancia) entre los datos. Son métodos que asumen que el criterio fundamental (y el único, en general) es la similitud entre los datos. Hasta donde se ha podido encontrar ([39]), incluso en los artículos más recientes, siempre que se ha hecho un intento de formalizar y estudiar el clustering en general, *la razón que hace a los datos agrupables* es la similitud (o disimilitud) entre ellos.

7.1.2. Tipos de clasificación

Ahora que se ha hecho claro qué se estudia en clasificación estadística y matemática, es posible poner nombre a sus diversas ramas de estudio. Lo que variará es la razón que hace a los datos agrupables:

1. **Clasificación supervisada:** Se agrupan los datos en función de un conjunto de datos previos, ya clasificados, que se consideran fiables. Generalmente el conjunto de datos es grande en volumen. Dentro de esta categoría se incluye:

    a) **Clasificación semisupervisada:** La razón de su agrupabilidad consiste en que se dispone de un pequeño conjunto de datos ya clasificados, pero se combinan con criterios geométricos o espaciales.

2. **Clasificación no supervisada:** Se busca agrupar datos pero sin disponer de un conjunto previo de datos ya clasificados. Dentro de estos métodos pueden encontrarse:

    a) **Clasificación por similitud:** Este es, a mi juicio, el nombre correcto para lo que se ha denominado Clustering, o Clasificación no supervisada, por los autores en la segunda parte de este documento. Es importante notar que siempre se considera el dato de la similitud (o disimilitud) como el criterio que guía los pasos de la clasificación; es la razón por la cual los datos son agrupables: *porque datos similares deben estar juntos y datos muy distintos lejos*. Esto es fundamental, porque no corresponde a esta ciencia el estudiar cuándo un dato es o no similar, sino que ya lo toma como un dato recibido del cual va a hacer usufructo.

    b) **Clasificación geométrica:** El criterio es la figura geométrica que forman los datos en el espacio. A esto corresponde lo visto en las primeras partes del primer capítulo.

    c) **Clasificación por distribución:** El criterio es la distribución del vector aleatorio. Es el caso visto cuando se estiman densidades y se analizan los conjuntos de nivel.

Entre otras posibles. Lo importante, y lo que se propone en este apartado, es **no diferenciar entre clasificación supervisada, semisupervisada y no supervisada**, como si en cada uno de ellas hubiese una unidad. Porque esto supondría que el criterio fundamental de distinción está en el dataset

inicial. Es verdad que es importante esa distinción, pero no es la fundamental. Lo que distingue un área de la clasificación de otra dándole unidad es *la razón por la cual los datos son agrupables*. Por eso, es imposible tener un conjunto de axiomas para clasificación no supervisada que contemple todos los casos en los que no haya un dataset, porque se tendrán situaciones y métodos tan dispares como: clasificar datos porque forman un círculo, clasificarlos por estimación de la densidad o por la similitud entre ellos. El verdadero criterio debería ser la razón que hace -y se asume- a los datos agrupables.

Por esa razón, si se desea particionar los datos en función de una figura geométrica (como fue visto en el capítulo 2), los axiomas, el punto de partida de dicho estudio, las propiedades relevantes de las funciones de clustering, etc. si bien pueden tener un punto en común con la teoría desarrollada por Ackerman, la teoría general del clustering que se propone no captura adecuadamente esos casos. Lo mismo ocurre con el estudio de cómo particionar los datos en función de si provienen de mixturas de distribuciones, o a partir de la estimación de conjuntos de nivel. Aquí es importante remarcar una diferencia: una cosa es el algoritmo que particiona datos y otra es el estudio matemático detrás. Cuando se distinguen las diferentes clasificaciones, se hace una distinción en la ciencia, en el estudio específico que se está realizando. Eso no significa que métodos que se originaron en otros espacios no puedan utilizarse en otros campos de estudio haciendo adaptaciones pertinentes. Por ejemplo, es perfectamente válido utilizar un algoritmo que asume que los datos provienen de una mixtura de normales para clusterizar a los datos y analizar a dicha función según los criterios de clasificación por similitud, si se utiliza la distancia entre los datos como una medida de disimilitud y se lo adapta para no tomar en consideración aspectos geométricos.

Pero hay que tener en cuenta que una clasificación por mixtura de distribuciones, o por estimación de conjuntos de nivel, como área de estudio, tendrá diferentes principios matemáticos que pueden llegar a ser incompatibles con la clasificación por similitud. Un ejemplo concreto es el axioma de invarianza por representación, que no lo cumplen esos algoritmos debido a que toman en consideración la figura geométrica que forman los datos en el espacio. Otros ejemplos son los axiomas de abundancia, consistencia, etc. Ninguno de ellos aplica para esos casos debido a que esas propiedades se definen a partir de la noción de similitud o disimilitud entre los datos, y no toman en consideración aspectos geométricos.

7.1.3. La importancia de la palabra

Un aspecto fundamental en la fundamentación y estudio de cualquier teoría está en el nombrar los diversos aspectos de la realidad. El acto de nombrar hace evidente al intelecto humano la realidad que está presente, pero que en una primera mirada escapa a nuestra visión. En este sentido, se dice que nombrar un aspecto de la realidad lo hace visible a nuestra inteligencia, y ese es todo el objetivo de la ciencia: hacer evidente verdades que están ahí. Se hará una analogía con Topología General. Todas las definiciones que se ven: conexión, conexión por caminos, arcoconexión, conexión local, compacidad, compacidad local, Lindelöf, etc. sirven para manifestar las diversas propiedades que ya están presentes en los diversos espacios topológicos, pero que si no se le pusiese un nombre no las veríamos; o no se harían tan manifiestas a nuestra inteligencia. No solo hace visible una realidad presente, aunque sutil, sino que además evita caer en el error. Tomemos como ejemplo: $[0,1]^{[0,1]}$, ¿Qué garantiza que dado un punto de dicho espacio existe una base de entornos compactos? La respuesta no importa, para este ejemplo, lo importante es que dicha pregunta pone de manifiesto, y presupone, que se está pensando en la posibilidad de que no ocurra eso; la verdadera pregunta es: ¿Es localmente compacto? Sin el concepto de compacidad local se es más susceptible al error, porque se puede asumir como evidente que para todo abierto siempre hay un compacto dentro de él con interior no vacío. Nombrar hace evidente un aspecto de la diversidad de posibilidades de fenómenos en los espacios topológicos. Este es el fundamento de lo que se quiere mostrar: hay que nombrar, clasificar las diversas propiedades, dar nombres a diversas posibilidades dentro de la clasificación; al modo que Margareta Ackerman hizo en su tesis con: localidad, invarianza por escala, abundancia, etc.

Este nombrar no debe confundirse con palabrerío vacío de contenido, sino el estudio de propiedades, fenómenos, situaciones reales. Una vez comprendida la realidad, y qué ocurre, así como la diferencia entre algoritmos, es posible nombrar.

7.1.4. Qué es la clasificación matemática general

Toda ciencia tiene un punto de partida, un método o modo de alcanzar verdades y un término. El punto de partida siempre son los principios del razonamiento (usualmente denominados axiomas); el método o modo de razonamiento varía, pero en matemática siempre es a través de deducciones lógicas, generalmente involucrando cálculos; y por último, el término es el conocimiento alcanzado.

Ahora bien, cuando Margareta Ackerman dice que la cuestión de los axiomas del clustering es un problema abierto, lo que quiere decir es que no está claro el punto de partida; no significa que no lo haya, solo que no es claro cuál es. Cuando, a partir del artículo de Kleinberg, se plantea la imposibilidad de axiomatizar el clustering, se apunta a una dificultad real: no parece posible axiomatizar el clustering si se lo piensa como clasificación sin ningún criterio previo que indique cómo agrupar los datos. Pero, al mismo tiempo, la implementación de algoritmos concretos como k-medias o single-linkage muestra que efectivamente se está operando sobre ciertos principios de clasificación, aunque estos no siempre hayan sido claramente identificados o formulados.

Se definió la clasificación estadística y matemática general como la ciencia que tiene por objeto material los datos cuantitativos (de ahí lo matemático) y por objeto formal los datos en cuanto clasificables o agrupables. Resulta relevante observar que esta ciencia no puede ser exclusivamente matemática, porque estudia la diversidad de posibilidades de clasificar datos cuantitativos en función de la variedad de razones por las que podrían ser agrupables. La pregunta sobre los fundamentos, o los axiomas y principios de un estudio, escapa al método matemático; requiriendo necesariamente el método discursivo y dialéctico utilizado por la filosofía. Este tipo de análisis posee una dimensión filosófica que excede el marco puramente matemático, pues se ocupa de cuestiones relativas a qué significa ser clasificable, los diversos modos en que un conjunto de datos puede agruparse y los distintos principios o supuestos involucrados en cada forma de clasificación. Como explica Sanguinetti en [38]: "*La matemática es una ciencia esencialmente deductiva. Opera partiendo de principios formales, no necesariamente reales, que son enunciados básicos y primeros que formulan ciertas características de los objetos matemáticos;*". Esto podría explicar parcialmente por qué aún no se ha encontrado un sistema axiomático satisfactorio para el aprendizaje no supervisado. El estudio de los fundamentos de la clasificación matemática parece requerir no solamente herramientas matemáticas, sino también una reflexión epistemológica y filosófica acerca de las diversas nociones de agrupabilidad involucradas. En este sentido, resulta necesario distinguir y ordenar los diversos campos del conocimiento implicados, reconociendo que existe un estudio más general que el de una teoría particular del clustering: el análisis de los principios que subyacen a las distintas formas de clasificación matemática y de las razones por las cuales los datos pueden considerarse agrupables, al modo como se intentó realizar en la subsección anterior. Por ello, considero que el problema de la axiomatización del clustering probablemente requiera una aproximación interdisciplinaria, donde el análisis matemático pueda complementarse con herramientas provenientes de la filosofía y la epistemo-

logía de la ciencia, especialmente en lo relativo al estudio de los principios y supuestos fundamentales. Sin una reflexión suficientemente rigurosa sobre dichos principios resulta difícil ordenar jerárquicamente el conocimiento, establecer con claridad los puntos de partida y determinar qué propiedades son verdaderamente relevantes para el estudio estadístico de la clasificación.

7.2. Clasificación por similitud

Es posible concluir que lo estudiado por Margareta Ackerman, y por buena parte de la literatura clásica sobre clustering, corresponde fundamentalmente a una teoría de la *clasificación por similitud*, en un sentido amplio que incluye también nociones de disimilitud. Desde este punto de vista, resulta posible reconsiderar el problema de la axiomatización del clustering.

Los autores que trabajaron sobre este problema buscaron construir un conjunto de definiciones y axiomas que caracterizasen de manera general qué es una función de clustering o, más ampliamente, qué significa clasificar por similitud. Sin embargo, podría sostenerse que parte de las dificultades encontradas provienen de intentar formular estas nociones exclusivamente mediante estructuras formales o lógico-simbólicas, dejando en un segundo plano la reflexión acerca de las nociones intuitivas y semánticas involucradas en la clasificación. En palabras de Sanguinetti en [38]: "*...los principios matemáticos ahora no se formulan según un criterio de evidencia material (por inducción a partir de la realidad), sino por simple evidencia formal (en el sentido de no-contradicción). Pero ya veremos que no pueden eliminarse en matemáticas algunos axiomas en el sentido clásico. (...) Existen principios matemáticos reales, leyes de la cantidad como tal, obtenidos por inducción, inmediatamente evidentes, y que están implícitos en todo razonamiento matemático. Por ejemplo, «dos cantidades iguales a una tercera son iguales entre sí», «el todo es mayor que la parte», son enunciados que se aplican a las cantidades reales, y ni siquiera de un modo aproximado, sino exacto.*"

En este sentido, resulta interesante recordar el caso de los axiomas y definiciones de Euclides. Conceptos como punto o línea no aparecen formulados originalmente mediante lenguaje lógico-formal moderno, sino a través de descripciones intuitivas orientadas a poner de manifiesto ciertas propiedades fundamentales de los objetos geométricos. Esto sugiere que, al menos en algunos contextos, la formulación matemática puede apoyarse también en nociones semánticas e intuitivas previas, y no exclusivamente en definiciones construidas a partir de simbolismos formales.

7.2.1. Punto de partida

El punto inicial debe ser la definición de lo que será el objeto principal de análisis: la función de clustering o algoritmo de clustering. La definición presentada a continuación es análoga a la de estimador paramétrico (esta última es: variable aleatoria que sirve para estimar un parámetro).

Definición 7.2.1. Una función de clustering es en primer lugar una función que recibe como datos de entrada el par (X, s) (o (X, d)), donde X es el conjunto de los datos y s la función de similitud entre ellos (en el otro caso d es la distancia o disimilitud) y devuelve una partición Γ de los datos. En segundo lugar, es una función que toma en consideración la similitud para agrupar los datos más similares entre sí y separar a los datos menos similares. Una función de k-clustering es una función de clustering que además recibe como dato de entrada la cantidad de clústers en que se desea particionar los datos y devuelve una k-partición de los datos; esto es: k clústers.

Para que dicha definición tenga sentido, es necesario notar que se utilizan implícitamente los siguientes principios semánticos:

Principio y fundamento de la función de similitud: la función de similitud utilizada debe reflejar, al menos de manera aproximada, alguna noción relevante de semejanza entre los datos respecto del problema considerado.

Principio y fundamento de la función de clasificación por similitud: todo algoritmo de clasificación por similitud debe buscar agrupar los datos utilizando el principio de que *datos similares deben estar agrupados y datos disimilares separados.*

Estos principios ya tienen consecuencias inmediatas en la práctica estadística. Por ejemplo, supóngase que se dispone de datos en el espacio euclídeo y se desea utilizar k-medias para realizar un clustering de los datos utilizando como métrica de disimilitud la distancia usual: ¿es realmente la distancia euclídea una medida adecuada, o aproximativa, de la disimilitud entre los datos, en función del problema que se desea estudiar? Porque si simplemente se la utiliza sin una reflexión previa sobre esta cuestión, es esperable que el resultado no diga nada real de los propios datos, pero si se logra una medida de la similitud entre ellos que sea bastante ajustada a la realidad, entonces toda esta teoría tiene aplicación consistente a dicho problema.

Otro principio axiomático interesante es el de invarianza por representación propuesto por Margareta Ackerman. Este principio centra el estudio exclusivamente en clasificar según la similitud, y no según otro criterio. Se lo propone como primer axioma de la clasificación por similitud (o disimilitud):

Axioma 1: Las funciones de clustering que clasifican por similitud son invariantes por representación al modo descrito por Margareta y expuesto en el capítulo 6.

Esto con el propósito de restringir la clasificación exclusivamente bajo el criterio de la similitud o disimilitud entre los datos; excluyendo del análisis cualquier otra variable que pertenezca a los datos. Resta cuestionarse las siguientes propiedades: las que refieren a la *abundancia*, a la *consistencia* y a la *invarianza por escala*.

Sobre la invarianza por escala

La propiedad de invarianza por escala no parece seguirse necesariamente de la noción semántica de clasificación por similitud y, por lo tanto, no parece constituir una propiedad esencial de toda función de clustering. Además, si bien se está clasificando por similitud, el modo, o criterio, por el cual se clasifica puede considerar el valor cuantitativo, en sí mismo como significativo, o solamente en relación con los demás; este último caso es el contemplado por invarianza por escala. Cuando se dice que una función de clustering es invariante por escala se está diciendo que no importa cuánto vale la similitud entre los datos, sino las proporciones con respecto al todo. Que $s(x, y) = 3$ y $s(y, z) = 6$ no importa, lo que toma en cuenta es que el primero es la mitad del segundo. Ahora bien, clasificando por similitud es posible querer considerar el caso de que el valor numérico signifique algo, como en el algoritmo de single-linkage que tiene como criterio de parada un valor mínimo de similitud (o una distancia máxima, como fue visto al revisar la contribución de Kleinberg). Otra razón no contemplada es que la medida de similitud (o disimilitud) puede no ser lineal; por ejemplo, que $s(y, z) = 6$ puede no significar que sean dos veces más similares esos datos que $s(x, y) = 3$, debido a que la medición no sigue una proporción lineal. Un ejemplo de esto fue visto en 3.4.3 con la siguiente medición de similitud: $s(x_i, x_j) = e^{-d^2(x_i,x_j)}$; en la cual puede observarse que tiene una relación no lineal con la distancia. Por lo tanto, estas son las razones por las que invarianza por escala no se propone como axioma, o propiedad fundamental de las funciones que clasifican por similitud.

Sobre la k-abundancia

Para el caso de abundancia, no parece ser una propiedad esencial, porque el hecho de que todos los algoritmos contemplados por Margareta lo cumpla no significa que no pueda existir un algoritmo interesante que no lo cumpla. Considérense los siguientes ejemplos:

1. **Middle-means:** Es como el k-medias para $k = 2$ funcionando según lo visto en la definición 3.1.5, pero solamente considerando 2-particiones que tengan la mitad de los datos (o la mitad más o menos un dato, en el caso impar). Este algoritmo siempre devuelve 2-particiones que dividen a los datos en dos particiones de tamaños prácticamente iguales. Por lo cual no son 2-abundante.

2. **Middle-linkage:** Es como el single-linkage pero con la siguiente modificación: inicia buscando los dos datos más similares y los agrupa en el mismo clúster. Luego busca cuál es el dato menos similar a esos dos y lo clasifica en otro clúster. Luego busca el dato más similar a este último y lo agrega a dicho clúster. Continúa alternando entre un clúster y otro, con el criterio de buscar cuál es el más similar al primer clúster, luego al segundo, y así, hasta dividir el dataset en dos partes casi del mismo tamaño. Este algoritmo tampoco es 2-abundante.

Por lo cual, parecería reducir la teoría el hecho de establecer como axioma la k-abundancia; descartaría este tipo de algoritmos que buscan particiones de tamaños fijos según una cantidad mínima, o según una proporción establecida respecto del total.

Sobre la localidad

La propiedad de localidad no es axiomática porque los algoritmos disociativos en general no la cumplen, por lo visto en el teorema 6.11 cuando se trató del tema de sensibilidad al contexto.

Sobre la consistencia

¿Qué pensar de la propiedad de consistencia, ya sea la exterior como la interior? Si bien, a priori, parece natural la consistencia externa, así como la interna ¿Qué pensar de cada una? En la taxonomía realizada por Ackerman

(ver 6.2.2) la mayoría de los algoritmos no la cumplen. Esto no es casualidad. Es debido al hecho de la combinación de consistencia con invarianza por escala. Cuando aumenta la similitud entre los datos de un mismo clúster, al ser invariante por escala, cambia la proporción total y puede ocurrir que los datos de un mismo clúster se dividan en más de uno, mientras los datos de distintos clústers se combinan. Esto ocurre por ese hecho: aumentar la similitud intraclúster puede aumentar la similitud entre datos de distintos clústers; debido a la invarianza por escala. Incluso es coherente que si a datos similares se los hace más similares, pueda ocurrir una desproporción de todo el conjunto de similitudes, causando que se unan datos que antes estaban separados para así separar datos que antes estaban unidos.

La consistencia externa es otra historia, y sobre ella se ha de decir dos cosas. La primera es que parece razonable considerar que hay que buscar la forma de salvar ese axioma y convertirlo en el segundo axioma de una función de clustering, porque es natural que si a datos ya clasificados se los hace más distintos entre sí entonces mantengan esa misma separación; en tanto sea una función de k-clustering, porque caso contrario puede ocurrir lo mismo que fue descrito para la consistencia interna. Un segundo punto de discusión importante es el hecho de que Ratio-Cut y Normalized-Cut no cumplen la consistencia externa (ver 6.2.2). A mi juicio podría discutirse la manera en que esos algoritmos fueron definidos por Ackerman en su tesis. Ella expone que dichos algoritmos funcionan maximizando las funciones de errores vistas en las definiciones 6.2.16 y 6.2.17; pero, analizando la lógica de lo que mide cada una de esas sumatorias, nótase que debería ser al revés, se debería buscar el mínimo y no el máximo. Porque el cut consiste en medir la similitud entre los datos de un clúster y su complemento, y, luego, se cocienta por, o bien el tamaño del clúster, o bien por su volumen (la similitud intracluster). Es decir, mide qué tanta similitud hay entre clústers distintos sopesando por el tamaño del clúster; una especie de promedio de similitud extracluster. Pero, si se busca que datos en distintos clústers sean lo menos similares posibles entonces se debería buscar minimizar dichas funciones y no maximizarlas. Aún así considerando a los algoritmos que devuelven el clúster que minimiza, y no maximiza, no se ha podido dar con una prueba de la consistencia externa; aunque se conjetura su validez. Por lo cual, en el caso de las funciones de k-clustering se propone como segundo la consistencia externa:

Axioma 2 (para funciones de k-clustering): Toda función de k-clustering debe ser exteriormente consistente.

A modo de resumen, quedaría así la propuesta axiomática:

1. **Principios fundacionales**

   a) **Principio y fundamento de la función de similitud:** la función de similitud utilizada debe reflejar, al menos de manera aproximada, alguna noción relevante de semejanza entre los datos respecto del problema considerado.

   b) **Principio y fundamento de la función de clasificación por similitud:** todo algoritmo de clasificación por similitud debe buscar agrupar los datos utilizando el principio de que *datos similares deben estar agrupados y datos disimilares separados.*

2. **Funciones de clustering en general**

   a) *Axioma 1:* Toda función de clustering basada en similitud es invariante por representación.

3. **Funciones de k-clustering**

   a) *Axioma 1:* Toda función de clustering basada en similitud es invariante por representación.

   b) *Axioma 2:* Toda función de k-clustering es exteriormente consistente.

El primer axioma es la consecuencia inmediata de afirmar que se está clasificando por similitud, porque entonces lo único que debe importar es la similitud entre los datos y no qué son los datos. Si se buscase estudiar clasificación por similitud combinada con otra técnica, entonces este axioma debería retirarse. El último axioma, para las funciones de k-clustering, es una propuesta a ser estudiada con más detalle, porque parte más de una corazonada que de una investigación profunda. Parecería razonable estudiar qué ocurre con la consistencia exterior en combinación con la propiedad de sensibilidad al contexto, como en los algoritmos disociativos, así como intentar responder la pregunta para los casos de Ratio-cut y Normalized-cut. Una propiedad vista por Margareta Ackerman en su tesis es la siguiente: *Si una función de clustering F depende únicamente de las distancias intraclústers de forma monótona entonces es exteriormente consistente*; por lo cual, para estudiar este axioma es necesario estudiar dos casos: las funciones que no dependen solamente de las similitud intraclústers y de las funciones que sí dependen solamente pero no monótonamente.

7.2.2. Metodología

Una vez propuestos los principios y axiomas como puntos de partida, es posible cuestionarse la metodología a emplear. Para responder a esta pregunta, se propone la siguiente subdivisión de la clasificación por similitud:

1. **Taxonomía de funciones de clustering**: Estudia la diversidad de propiedades y funciones de clustering, así como las relaciones entre dichas propiedades.

2. **Medidas de calidad**: Estudia el problema de cómo medir qué tan buena es una función de clustering o qué tan bueno es un clustering.

3. **Medidas de similitud o disimilitud**: Cómo medir las diversas formas de similitud para distintas situaciones o tipos de datos.

El primer aspecto consiste en el estudio taxonómico de los algoritmos, esto es, investigar las diversas propiedades que nacen del comportamiento del algoritmo en función de su input y output, tal como lo han hecho Kleinberg, Ackerman, entre otros. Esto permite entender a un nivel más macroscópico y general cómo funcionan ciertos algoritmos y qué los diferencia. Por ejemplo, el saber que los métodos aglomerativos son locales y difieren de los disociativos, como CUBT (Árboles de decisiones), que son sensibles al contexto. Como se dijo en el apartado de la importancia de la palabra, el poner nombre a distintas situaciones, propiedades, particularidades, ayuda a entender más las posibilidades de clasificar y cómo funcionan esos métodos. La introducción de conceptos como localidad, dependencia exclusiva de las similitudes intraclúster o sensibilidad al contexto da ideas claras al investigador de cuáles son las diferencias esenciales de esos métodos.

El segundo aspecto corresponde a las medidas de calidad. Se propone que debería englobar, en principio, al menos dos asuntos: el propuesto por Margareta en su tercer capítulo y el estudio de la robustez y la clusterabilidad de un conjunto de datos en su sexto capítulo. Si bien en esta tesis no se alcanzó una reflexión con profundidad sobre este apartado, aun así se remarca que es un tema muy importante que debe explorarse. Esto incluiría reflexionar sobre los axiomas propuestos por Margareta y establecer métodos y resultados relativos a la robustez, sensibilidad y clusterabilidad de un conjunto de datos.

El tercer apartado es el de dar métodos y criterios, sobre cómo establecer medidas de similitud para un conjunto de datos. También estudiar propiedades de esas medidas, por ejemplo, para los casos de medidas discretas, continuas, etc. y cómo eso se relaciona con las funciones de clustering

y sus propiedades. Este problema de la medición fue nombrado por William E. Wright en su artículo [36], como se comentó en su apartado específico. Métodos de construcción de grafos de similitud también fueron vistos en el apartado de construcción de grafos de similitud en el capítulo 3.

Más sobre la taxonomía

Un aspecto sobre el estudio taxonómico que se desea resaltar es el siguiente: libertad en el estudio de las propiedades y sus relaciones. Lo importante de este estudio es observar propiedades y ver las relaciones entre ellas, más allá de buscar en ellas un corpus axiomático para el análisis del clustering. Teniendo presente los principios y axiomas como punto de partida se pueden idear diversos algoritmos y estudiar sus propiedades y relaciones; profundizando la diversidad de posibilidades de clasificación basada en similitud. Otro aspecto a profundizar es el estudiar datos en un espacio topológico y similitudes que sean continuas, o datos en un espacio de medida y similitudes que sean funciones medibles. Esta idea no es nueva, sino que fue expuesta en el tercer capítulo en el apartado de k-medias y probabilidad. Es interesante no solo estudiar las propiedades propuestas por Ackerman en función del input-output, sino también incorporar análisis probabilístico y topológico. El hecho de agregar probabilidad o continuidad no modifica el tipo de clasificación, sino que se mantiene la clasificación por similitud, pero haciendo más específico el dominio de los datos o las funciones de similitud.

7.3. Crítica a los autores de la segunda parte

7.3.1. William E. Wright

Sin lugar a dudas uno de los pioneros en proponer una teoría general del clustering. Si bien es interesante su propuesta, es muy rescatable el problema de la medición planteado por él. Este problema es un problema real e importante. Si se va a clasificar por similitud debemos tener métodos de medir dicha similitud. El asunto es que para él no es una única medición de la similitud, sino varias; esto es un problema, porque la medición de la similitud debe ser solamente una, sino no es una medición de similitud. Cuando se tienen varias mediciones de similitud, lo que se tiene son varias similitudes con respecto a ciertos aspectos de los datos, por ejemplo: una torta de chocolate es muy similar a una torta de limón con respecto al hecho de ser torta, o ser un postre, pero es muy poco similar con el hecho de ser de chocolate y la otra ser de limón; sería, según esta segunda medición, más similar a una

tableta de chocolate. Por lo que, hay que hacer una diferencia entre clasificar por similitud y clasificar por similitudes. Y Wright propone un estudio sobre clasificación por similitudes y no por similitud, aunque tienen su punto de unión. Otra idea a rescatar es justamente esta, no solamente clasificar por similitud sino por similitudes, algo que no fue profundizado por los siguientes autores ya citados en capítulos anteriores, y que no hay mucho desarrollo. Porque considerar que las diversas similitudes dan un punto en el espacio euclídeo y aplicar los métodos ya vistos no parece resolver completamente el problema, porque es necesario utilizar el hecho de que esas mediciones son similitudes. Un estudio en profundidad sobre este asunto puede que sea un problema abierto. Un aspecto discutible del artículo de Wright es, justamente, este: la idea de que las mediciones dan un punto en el espacio que representa su posición.

Otro aspecto que merece destacarse es el hecho de hablar de masa de los datos; un aspecto interesante que podría considerarse en futuros estudios, especialmente cuando se tienen datos discretos que pueden superponerse. El clustering de datos discretos es un tema que no he encontrado mucho desarrollo, y parece plausible que estas ideas de masa puedan ser útiles.

Para concluir, la principal deficiencia de la propuesta de Wright está en su complejidad. Son muchos axiomas para una teoría general del clustering. A su vez, la definición de función de clustering se limita al caso de funciones de clustering basadas en optimizar una función de costo. Esta complejidad se ve patente cuando él expone un ejemplo de función de clustering.

7.3.2. Puzicha, Hofmann y Buhmann

En el artículo de los autores, los mismos utilizan en el título que el clustering que se está realizando es basado en proximidad (similitud). Esto refuerza la idea propuesta en esta tesis de que un posible error al buscar axiomatizar una teoría general del clustering radica en la indefinición del área de estudio; al concretizarla en clasificación por similitud se vuelven más claros sus axiomas y finalidad.

La propuesta del artículo está en axiomatizar una clasificación por similitud basada en funciones objetivo. Buscar una teoría general de la clasificación por similitud basada en funciones de costo es algo sumamente deseable. Un punto donde se debería profundizar más es en el axioma de monotonía de las funciones de costo (criterios de clustering, según los autores), porque no queda claro que se debería cumplir esa propiedad para cualquier clustering

de los datos; sí parece razonable que se debería cumplir para el caso óptimo, o los óptimos (que minimizan el criterio), pero considero cuestionable que se debiere cumplir para cualquier partición de los mismos.

Para los autores, el núcleo axiomático se encuentra en los axiomas de invarianza, que, según ellos, eliminaría cualquier sesgo en la escala de los datos. Estos axiomas ya fueron citados en su respectivo apartado, y son los de *invarianza por escala* e *invarianza por traslación*. No se pretende aquí adoptar una posición definitiva sobre si deberían ser, o no, axiomas, aunque como se sostuvo anteriormente, en principio no es la pregunta fundamental, sino más bien: ¿cuáles son las consecuencias de dichas propiedades y cómo se relacionan entre sí? así como también: ¿qué consecuencias trae a la dinámica input-output y a su relación con otras propiedades como localidad, consistencia exterior, etc? También es mucho más interesante ver cómo dichas propiedades modifican y caracterizan la propia función de costo; como lo demuestran los autores en su artículo al estudiar funciones de costo que son, además, aditivas.

Este artículo, junto con lo desarrollado por Kleinberg y Ackerman, parece constituir un importante punto de partida para una teoría general de la clasificación por similitud basada en funciones de costo.

7.3.3. Kleinberg

Hasta donde ha alcanzado esta investigación, de todos los desarrollos sobre el tema, y formalizaciones, pienso que Kleinberg ha encontrado un enfoque particularmente adecuado de cómo formalizar una teoría general de la clasificación por similitud en su apartado taxonómico. Mientras los autores anteriores han dado diversos axiomas y propiedades y resultados matemáticamente más complejos, él, en su sencillo artículo, da un método investigativo simple y potente de cómo analizar las propiedades de las funciones de clustering en general. Este método es el utilizado por Ackerman en su tesis y publicaciones; consiste en estudiar las funciones de clustering a partir de cómo varía el output en función del input. Este cambio metodológico permite, sí, el objetivo tan deseado: una teoría general del clustering. El principal desarrollo por el cual es conocido es su *teorema de imposibilidad*. Este teorema tiene el detalle de que solamente es válido para funciones basadas en disimilitud (o distancia), mientras que para funciones basadas en similitud, como se comentó en la observación del teorema 5.1, la que devuelve las componentes conexas del grafo de similitudes asociado cumple con los tres axiomas propuestos por Kleinberg. Al final de este apartado se demostrará un resultado,

hasta donde se ha investigado, novedoso: esa función es la única que cumple los tres axiomas propuestos por él.

Otros aportes que conviene destacar de su artículo, porque son muy valiosos para una teoría general de la clasificación por similitud, son:

- **Algoritmos linkage-based con otros métodos de parada**. A diferencia de los vistos en la tesis de Ackerman, Kleinberg contempla la posibilidad de otros criterios de parada; que han sido vistos en el teorema 5.2.
- **Estructura de la imagen de una función de clustering**. En su artículo, él demuestra que el conjunto imagen de una función de clustering invariante por escala y consistente forma una anticadena. Este tipo de análisis está ausente en la tesis de Ackerman y es interesante tenerlo presente en futuros desarrollos.
- **Propiedades flexibilizadas:** Ackerman tiene versiones flexibilizadas de ciertas propiedades, como exteriormente consistente, interiormente consistente, k-abundante, exteriormente abundante, etc. Pero Kleinberg propuso una serie de flexibilizaciones de invarianza por escala, abundancia y consistencia que deberían tomarse en consideración para futuros análisis taxonómicos. (ver 5.0.6)

Como se comentó en el teorema 5.1, las funciones de similitud permiten que exista similitud cero entre los datos, lo que hace posible que el grafo pesado asociado (Γ_s) tenga más de una componente conexa (algo imposible para funciones de disimilitud). Luego, la función de clustering que devuelve la cantidad de componentes conexas ($f(s) = \Gamma_s$) cumple los tres axiomas propuestos por Kleinberg (Ver la proposición 1 en 5.0.3). Por lo tanto, como posible resultado original de esta investigación, se demuestra que esta función de clustering, que devuelve en la cantidad de componentes conexas, es la única que satisface simultáneamente los tres axiomas propuestos por Kleinberg cuando se trabaja con funciones de similitud. Hasta donde alcanza la revisión bibliográfica realizada, no se ha encontrado en la literatura una demostración ni una discusión explícita sobre este punto. Este hallazgo, por tanto, constituye una posible extensión y formalización de la observación mencionada en [39].

Teorema 7.1. *(**Teorema de unicidad**) Existe una única función de clustering f basada en similitud que es consistente, abundante e invariante por escala, y esta es la que devuelve las componentes conexas:*

$$f(s) = \Gamma_s$$

Antes de iniciar la prueba se aclara la siguiente licencia terminológica: cuando se haga referencia a: *componente conexa de s*; se está refiriendo a la componente conexa del grafo asociado a la función de similitud s ya expuesto en 5.0.3.

Demostración. Consideremos f una función de clustering basada en similitud consistente, abundante e invariante por escala. Como f es abundante $\forall \Gamma \in \text{Part}(X)\ \exists s \in \Omega_s\ : f(s) = \Gamma$.

La idea de la demostración consiste en modificar dicha similitud para construir otras, sin alterar el clustering producido por f, utilizando consistencia e invariancia por escala, hasta forzar que el clustering coincida necesariamente con las componentes conexas mostrando que no podría ser de otra forma.

Como f no necesariamente devuelve las componentes conexas, pueden ocurrir dos casos:

1. $\exists x, y \in X, \quad x \nsim_\Gamma y$, pero $x \sim_{\Gamma_S} y$.
2. $\exists x, y \in X, \quad x \sim_\Gamma y$, pero $x \nsim_{\Gamma_S} y$.

El primer caso corresponde a la existencia de dos datos que no están en el mismo clúster de Γ pero sí en la misma componente conexa del grafo asociado a s. El segundo caso corresponde a dos datos que sí están en el mismo clúster de Γ pero no en la misma componente conexa. Si ocurre el primer caso, como están en la misma componente conexa asociada a s, entonces existe un camino $x = x_1, x_2, \ldots, x_{n-1}, x_n = y$, tal que $s(x_i, x_{i+1}) > 0$, $\forall i \in \{1, 2, \ldots, n-1\}$. Esto implica que existen dos datos x_i y x_{i+1} tales que $x_i \nsim_\Gamma x_{i+1}$, pero $s(x_i, x_{i+1}) > 0$. Si ocurre el segundo caso entonces $s(x, y) = 0$, porque caso contrario pertenecerían a la misma componente conexa asociada a s. Por lo tanto, se concluye que los dos casos anteriores implican:

1. $\exists x, y \in X, \quad x \nsim_\Gamma y$, pero $s(x, y) > 0$.
2. $\exists x, y \in X, \quad x \sim_\Gamma y$, pero $x \nsim_{\Gamma_S} y$ y $s(x, y) = 0$.

Se define la siguiente función de similitud s' de la siguiente forma:

1. $s'(x, y) = 0$ cuando $x \nsim_\Gamma y$
2. $s'(x, y) > \text{máx}_{x \neq y}\, s(x, y)$ cuando $x \sim_\Gamma y$ (en caso de que sea cero, se coloca $s'(x, y) = 1$).

Entonces s' es un cambio s-consistente y por lo tanto $f(s') = f(s) = \Gamma$; de lo que se concluye:

$$\forall s \; \exists s' \; : \; f(s) = f(s'). \tag{7.1}$$

Y $f(s') = \Gamma_{s'}$, son las componentes conexas de s'.

El objetivo es probar que no puede existir una función de similitud s tal que $f(s)$ no devuelva sus componentes conexas. Se recuerda que Γ_s corresponde al clustering formado por las componentes conexas de una función de similitud s.

Entonces, juntando 7.1 con la propiedad de abundancia, se cumple:

$$\forall s \; \exists s^* \; : \; f(s^*) = \Gamma_s = \Gamma_{s^*},$$

y además $s^*(x, y) > 0$ siempre que $x \sim_{\Gamma_s} y$.

Cada vez que se escribe s^* se hace referencia a esta similitud. La primera igualdad es consecuencia de la propiedad de abundancia, la segunda se obtiene aplicando 7.1.

Dado Γ, se obtuvo por abundancia que este clustering tiene asociado dos similitudes, una s y otra derivada de esta, s', tal que $\Gamma = f(s) = f(s')$ y Γ son las componentes conexas de s'. Para probar el teorema, se demostrará que las siguientes dos situaciones son imposibles:

1. $\exists x \nsim_{\Gamma} y$ tales que $s(x, y) > 0$.
2. $\exists x \sim_{\Gamma} y$ tales que $x \nsim_{\Gamma_s} y$.

Y con ello quedaría demostrado el resultado, ya que $f(s)$ serían sus componentes conexas.

Para el primer caso:

Sean x y y tales que $x \nsim_{\Gamma} y$ y $s(x, y) > 0$.

1. $s'(x, y) = 0$ y $s^*(x, y) > 0$, porque x está en la misma componente conexa que y en Γ_s.
2. Sea $r = \text{mín}\{s(x, y) \; : \; s(x, y) > 0\}$, entonces existe una constante $c > 0$ tal que
$$c \cdot s^*(x, y) < r. \tag{7.2}$$
3. Por invarianza por escala $f(c \cdot s^*) = f(s^*) = \Gamma_s$. Se llamará $s_3 = c \cdot s^*$.

4. Sea s_4 definida del siguiente modo: $s_4(x,y) = s_3(x,y)$ cuando $x \nsim_\Gamma y$, y cuando $x \sim_\Gamma y$ se define $s_4(x,y) = s(x,y)$.

5. s_4 es s_3-consistente porque

   a) si $x \nsim_{\Gamma_\mathcal{S}} y$ y $x \nsim_\Gamma y$, entonces $s_4(x,y) = s_3(x,y)$.

   b) si $x \nsim_{\Gamma_\mathcal{S}} y$ y $x \sim_\Gamma y$, entonces $s_4(x,y) = s(x,y) = 0 = s_3(x,y)$.

   c) si $x \sim_{\Gamma_\mathcal{S}} y$ y $x \nsim_\Gamma y$, entonces $s_4(x,y) = s_3(x,y)$.

   d) si $x \sim_{\Gamma_\mathcal{S}} y$ y $x \sim_\Gamma y$, entonces $s_4(x,y) = s(x,y) > c \cdot s^*(x,y) = s_3(x,y)$.

   Luego, $f(s_4) = f(s_3) = \Gamma_\mathcal{S}$.

6. s_4 es s-consistente porque:

   a) si $x \nsim_{\Gamma_\mathcal{S}} y$ y $x \nsim_\Gamma y$, entonces $s_4(x,y) = s_3(x,y) = 0 = s(x,y)$.

   b) si $x \nsim_{\Gamma_\mathcal{S}} y$ y $x \sim_\Gamma y$, entonces $s_4(x,y) = s(x,y) = 0$.

   c) si $x \sim_{\Gamma_\mathcal{S}} y$ y $x \nsim_\Gamma y$, entonces $s_4(x,y) = s_3(x,y) = c \cdot s^*(x,y) < s(x,y)$.

   d) si $x \sim_{\Gamma_\mathcal{S}} y$ y $x \sim_\Gamma y$, entonces $s_4(x,y) = s(x,y)$.

   Por lo tanto, $f(s_4) = f(s) = \Gamma$.

7. Pero por hipótesis $\Gamma_\mathcal{S} \neq \Gamma$, lo que es una contradicción.

Para el segundo caso:

1. $s'(x,y) > 0$ y $s^*(x,y) = 0$, porque x no está en la misma componente conexa de y en $\Gamma_\mathcal{S}$.

2. Por lo visto en el primer caso, es imposible que haya pares de puntos x, y tales que $s(x,y) > 0$ y $s'(x,y) = 0$. Esto implica que si $s'(x,y) = 0$ entonces $s(x,y) = 0$.

3. Por lo tanto, $\Gamma_\mathcal{S}$ es un refinado de Γ como clústers.

4. Sea $a = \max s(x,y)$ y $c > 0$ tal que $c \cdot s^*(x,y) > a$ cuando $s^*(x,y) > 0$. Notar que $s(x,y) > 0 \Rightarrow s^*(x,y) > 0$. Se define $s_3 = c \cdot s^*$.

5. Por invarianza por escala $f(s_3) = f(s^*) = \Gamma_\mathcal{S}$.

6. A su vez, s_3 es un cambio s-consistente, porque si $x \nsim_\Gamma y$ entonces $s(x,y) = 0 = s^*(x,y) = c \cdot s^*(x,y) = s_3(x,y)$; y si $x \sim_\Gamma y$ entonces $s_3(x,y) = c \cdot s^*(x,y)$, y aquí se bifurca en dos posibilidades:

a) Si $s(x,y) = 0$ entonces $s^*(x,y) = 0$, luego $s_3(x,y) = s(x,y)$.

b) Si $s(x,y) > 0$ entonces $s^*(x,y) > 0$, luego $s_3(x,y) = c \cdot s^*(x,y) > a \geq s(x,y)$.

7. Por lo tanto, $f(s_3) = f(s) = \Gamma$. Lo que es una contradicción.

Se concluye, por lo tanto, que no puede existir una función f *consistente*, *invariante por escala* y *abundante* que admita como imagen de una función de similitud s algo distinto de sus componentes conexas.

□

7.3.4. Margareta Ackerman

Se ha dicho mucho sobre su trabajo en esta tesis. Solamente se agregarán dos comentarios, uno a su trabajo en medidas de calidad de clustering y otro sobre su trabajo en robustez.

Medidas de calidad

Según la propia autora, el objeto de las mediciones de calidad es determinar qué tan bueno es un clustering de un dataset en particular para diversas finalidades; algunas pueden ser para comparar entre distintos algoritmos, otros para comparar entre un mismo algoritmo, etc. Es un trabajo importante el analizar estas funciones con más detalle, y la propuesta de Ackerman, para dar una respuesta satisfactoria. En este comentario se dará una opinión sobre su propuesta y cuáles son los posibles problemas que se observaron.

El primero es la fundamentación de su propuesta axiomática. Ella busca dar axiomas de medidas de calidad, no solamente inspirándose en los tres axiomas propuestos por Kleinberg para funciones de clustering, sino intentando dar una traducción directa de ellos. La misma autora no da una razón de por qué los axiomas de medidas de calidad deberían ser análogos a los de funciones de clustering, y tampoco queda claro su razón al analizar cada uno de dichos axiomas.

El segundo problema es de orden filosófico. Cuando se quiere medir qué tan bueno es algo, es necesario con anterioridad tener presente cuál es el ideal de perfección para luego medir los grados de cuánto se aproxima o no. Téngase como ejemplo la salud humana. Una medida de calidad para la salud sería contar la cantidad de órganos y partes del cuerpo defectuosas (o sanas). Si la medida da cero significa que está totalmente saludable. Pero, para poder

saber que esa medida es legítima, es decir, que es capaz de realizar lo que se propone, es necesario saber de antemano qué es un ser humano perfectamente saludable, o qué es un órgano perfectamente saludable, etc. Por lo cual, resulta difícil construir un corpus axiomático para las medidas de calidad sin dar una noción previa, al menos incipiente, de qué es un clustering perfecto. Retornando al ejemplo de la salud, algunas medidas medirán la cantidad de órganos defectuosos, otras, cuántos parámetros de vitaminas y minerales en sangre no se cumplen, tal vez un tercer medidor mida cuántas veces al mes va al médico el paciente, etc. Solo posteriormente se podrá visualizar qué cumplen todas esas medidas en común y ahí fundamentar un corpus axiomático. Porque es posible dar un conjunto de axiomas, que las mejores medidas de calidad cumplan, pero aún así no es posible juzgar si realmente son mediciones correctas de calidad. El axioma debe servir como criterio para saber si se está o no ante una función que mide la calidad de un clustering.

Por último, los tres axiomas propuestos por Ackerman, a priori, no parecen ser esenciales para lo que es una medida de calidad, es decir, no parecen ser conclusiones al menos indirectas de lo que se entiende semánticamente por una medida de calidad. No es inmediato, ni directo, el por qué una medida de calidad debería ser independiente de la escala. O por qué debería mejorar ante cambios consistentes de la similitud, más aún teniendo en cuenta lo comentado sobre el fenómeno que ocurre al mezclar consistencia interna con invarianza por escala. El axioma de abundancia tampoco es natural, o intuitivo. Otro aspecto es que ella misma muestra que las principales funciones de costo que sirven como medidas de calidad fallan al menos uno de los tres axiomas; esto pone en cuestión su carácter fundamental.

Estas medidas de calidad se asemejan más a funciones de costo, y, por lo tanto, mantienen una relación estrecha con la teoría de funciones de clustering. Una medición de calidad interesante, y no contemplada, es medir la calidad de una función de clustering.

Robustez y oligarquías

A mi juicio, este apartado contiene algunos de los desarrollos más interesantes de la tesis de Ackerman respecto al problema de evaluar la calidad de funciones de clustering y de clusterings particulares. En especial, la noción de datasets bien clusterizables resulta sumamente relevante, ya que permite comenzar a estudiar formalmente qué tan apto es un conjunto de datos para ser clusterizado. De manera análoga, los resultados sobre robustez y oligar-

quías permiten analizar la sensibilidad y estabilidad de distintas funciones de clustering frente a perturbaciones de los datos.

Estos conceptos parecen constituir un posible punto de partida para una teoría más general de medición de calidad en clustering. Por ejemplo, si se considera que un "buen clustering" es aquel que satisface cierto grado de α-separación —particularmente para $\alpha \geq 1$—, entonces podría definirse una medida de calidad de la forma

$$m(\Gamma, (X, d), k),$$

que devuelva el supremo de los valores de α tales que el k-clustering Γ es α-separable. Bajo esta perspectiva, la calidad del clustering se evalúa en función del ideal de separabilidad adoptado.

Además, los resultados de Ackerman muestran que ciertos algoritmos poseen capacidades específicas para detectar determinados tipos de estructuras en los datos. Por ejemplo, la propia autora demuestra que el algoritmo single-linkage detecta clusterings 1-separables. Esto sugiere la posibilidad de desarrollar metodologías donde, una vez obtenido un clustering mediante un algoritmo particular, se evalúe posteriormente su grado de separabilidad utilizando criterios adicionales. Incluso podrían compararse distintos algoritmos con el objetivo de aproximarse al clustering que maximice cierto criterio de separación.

En todos estos casos se observa un aspecto importante: las medidas de calidad parecen depender necesariamente de algún ideal previo de clustering. En este ejemplo, dicho ideal está dado por la noción de α-separación. Otras propiedades, como el β-balance, podrían desempeñar un papel complementario o secundario dentro de este tipo de análisis.

Por otra parte, el estudio de la robustez introduce una tensión particularmente interesante. Los resultados presentados por Ackerman sugieren que algoritmos muy eficaces para detectar ciertos tipos de clusterings bien separados pueden, al mismo tiempo, resultar menos robustos frente a perturbaciones de los datos. Esta relación entre eficacia y robustez parece revelar un fenómeno estructural importante dentro de la teoría del clustering.

Un estudio más profundo de estas cuestiones —así como de su relación con medidas de calidad y con distintos ideales de clustering— podría contribuir significativamente al desarrollo de una teoría más general sobre la evaluación de funciones y algoritmos de clustering.

7.3.5. Estudios posteriores a Ackerman

Hay poco desarrollo posterior a la tesis de Margareta Ackerman que aborde la problemática de fundamentar una teoría general del clustering. Uno de los trabajos más recientes encontrados es [39], titulado "*Axioms for Graph Clustering Quality Functions*". En él, se insiste en que persiste la problemática de fundamentar una teoría general del clustering, y que aún no se tiene claro qué es un clustering, o una función de clustering. Los autores proponen un corpus axiomático para las funciones de calidad de clusterings de grafos. Se inspiran en el trabajo de Ackerman sobre funciones de calidad, adaptándolo y complementándolo para el caso de grafos. Pero no parece observarse todavía un desarrollo ampliamente consolidado de una teoría verdaderamente general del clustering en el sentido discutido a lo largo de esta tesis.

7.4. Conclusión y futuros trabajos

A partir de la propuesta axiomática y metodológica, hay varias posibilidades para continuar la investigación y el trabajo. Un trabajo interesante sería compilar de forma organizada los resultados de las investigaciones hechas hasta el momento, como se realizó en este documento, pero estructurando el documento en función de la temática y no del autor; como los libros de texto de cálculo, topología, etc. Una continuación directa de este trabajo es hacer una revisión de los teoremas de imposibilidad a la luz del teorema de unicidad (7.1), buscando nuevos teoremas de unicidad. También sería importante estudiar la diferencia entre funciones de clustering basadas en disimilitud y las basadas en similitud, analizar si la consistencia externa debería ser un axioma para funciones de k-clustering, investigar sobre la axiomática de las mediciones de calidad, así como una fundamentación teórica sobre qué es una medida de calidad. Sería interesante profundizar sobre las diversas posibilidades de realizar mediciones de similitud, cuando son continuas, medibles, discretas, etc., y cómo varían los teoremas vistos en el estudio taxonómico en esos casos. Como puede observarse, el terreno es muy fértil y muy interesante, hay muchas líneas de investigación y desarrollo que podrían realizarse y profundizarse.

Por lo tanto, la principal conclusión de esta tesis es que los principales desarrollos de los autores estudiados en la segunda parte, especialmente Kleinberg y Ackerman, pueden entenderse como ***clasificación por similitud***; y que una de las grandes dificultades de axiomatizar el clustering estaba en que se buscaba axiomatizar la clasificación no supervisada, pero implícitamente

consideraban que la razón que hace a los datos agrupables es la similitud entre ellos, o su disimilitud. Esto parecía conducir a dos dificultades: no tener claro qué es una función de clustering y qué es un clustering, y una teoría aparentemente general, pero cuyo alcance efectivo parecería más restringido; porque, como se vio, el alcance que tiene está limitado para ciertas situaciones, no contemplando métodos como los de clasificar en razón de la figura geométrica, o en función de distribuciones mixtas, etc. Una vez identificado el objeto material y formal de lo que se está estudiando, estas dificultades parecen disminuir; permitiendo además identificar para qué casos es útil el desarrollo realizado por dichos autores.

Bibliografía

[1] Jon Kleinberg *"An Impossibility Theorem for Clustering"*. Advances in Neural Information Processing Systems(NIPS) 15,2002.

[2] Rudolf Scitovski , Kristian Sabo , Francisco Martínez-Álvarez , Šime Ungar, *Cluster Analysis and Applications*. Springer. (2021).

[3] T. Maroševic, Data clustering for circle detection. Croat. Oper. Res. Rev. 5, 15–24 (2014)

[4] M.A. Kashiha, C. Bahr, S. Ott, C.P. Moons, T.A. Niewold, F.T.D. Berckmans, Automatic monitoring of pig locomotion using image analysis. Livestock Sci. 159, 141–148 (2014)

[5] S. Rueda, S. Fathima, C.L. Knight, M. Yaqub, A.T. Papageorghiou, B. Rahmatullah, A. Foi, M. Maggioni, A. Pepe, J. Tohka, R.V. Stebbing, J.E. McManigle, A. Ciurte, X. Bresson, M.B. Cuadra, C. Sun, G.V. Ponomarev, M.S. Gelfand, M.D. Kazanov, C.-W. Wang, H.-C. Chen, C.-W. Peng, C.-M. Hung, J.A. Noble, Evaluation and comparison of current fetal ultrasound image segmentation methods for biometric measurements: a grand challenge. IEEE Trans. Med. Imaging 10, 1–16 (2013)

[6] R. Scitovski, K. Sabo, The adaptation of the k-means algorithm to solving the multiple ellipses detection problem by using an initial approximation obtained by the DIRECT global optimization algorithm. Appl. Math. 64, 663–678 (2019)

[7] R. Scitovski, T. Maroševic, Multiple circle detection based on center-based clustering. Pattern Recogn. Lett. 52, 9–16 (2014)

[8] Pranjal Awasthi and Maria-Florina Balcan. Center based clustering: A foundational perspective. In Christian Hennig, Marina Meila, Fionn Murtagh, and Roberto Rocci, editors, Handbook of cluster analysis. CRC Press, 2015.

[9] Avrim Blum, John Hopcroft, and Ravindran Kannan, Foundations of Data Science, 2018.

[10] Jon Kleinberg. "An Impossibility Theorem for Clustering." Advances in Neural Information Processing Systems (NIPS) 15, 2002.

[11] Shai Ben-David, Margareta Ackerman. Measures of Clustering Quality: A Working Set of Axioms for Clustering. Advances in Neural Information Processing Systems (NIPS) 21, 2008.

[12] Margareta Ackerman. "Towards Theoretical Foundations of Clustering". University of Waterloo, Ontario, Canada, 2012.

[13] Reza Bosagh Zadeh, Shai Ben-David. A Uniqueness Theorem for Clustering. 2012.

[14] U Von Luxburg, S Ben-David. Towards a statistical theory of clustering. Fraunhofer IPSI, Darmstadt, Germany. 2005.

[15] J Correa-Morris. An indication of unification for different clustering approaches. 2013.

[16] Vincent Cohen-Addad, Varun Kanade, Frederik Mallmann-Trenn. Clustering Redemption–Beyond the Impossibility of Kleinberg's Axioms. Advances in Neural Information Processing Systems 31 (NeurIPS 2018).

[17] Sanchez Garcia, Ruben and Strazzeri, Fabio. Morse Theory and an impossibility theorem for graph clustering. 2018.

[18] R. Scitovski, S. Scitovski, A fast partitioning algorithm and its application to earthquake investigation. Comput. Geosci. 59, 124–131 (2013)

[19] G. Asencio-Cortés, S. Scitovski, R.Scitovski, F. Martínez-Álvarez, Temporal analysis of croatian seismogenic zones to improve earthquake magnitude prediction. Earth Sci. Inf. 10, 303–320 (2017)

[20] Morales-Esteban A, Scitovski S, Scitovski R (2014) A fast partitioning algorithm using adaptive mahalanobis clustering with application to seismic zoning. Comput Geosci 73:132–141

[21] Lu, Z., Wahlström, J., Nehorai, A. Community Detection in Complex Networks via Clique Conductance. Sci Rep 8, 5982 (2018). https://doi.org/10.1038/s41598-018-23932-z

[22] Hennig, C., Meila, M., Murtagh, F., and Rocci, R. (Eds.). (2015). Handbook of Cluster Analysis (1st ed.). Chapman and Hall/CRC. https://doi.org/10.1201/b19706

[23] Boyd, S. and Vandenberghe, L. (2004). Convex Optimization. Cambridge University Press, Cambridge.

[24] Biggs, N. (1993). Algebraic Graph Theory. Cambridge Mathematical Library, Issue 67. Cambridge University Press, Cambridge, 205 pp.

[25] G.W.Stewart and Ji-guang Sun. Matrix Perturbation Theory. Academic Press, 1990.

[26] Rinaldo, A. (2017). "Lecture 15: October 23". En "36-755: Advanced Statistical Theory I" (scribe: J. Wang). Carnegie Mellon University. https://www.stat.cmu.edu/~arinaldo/Teaching/36755/F17/Scribed_Lectures/F17_1023.pdf

[27] J. Lei and A. Rinaldo, Consistency of spectral clustering in stochastic block models,*Annals of Statistics* **43** (2015), no.1, 215–237. https://doi.org/10.1214/14-AOS1274.

[28] Saerens, M., Fouss, F., Yen, L., Dupont, P. (2004). The Principal Components Analysis of a Graph, and Its Relationships to Spectral Clustering. In: Boulicaut, JF., Esposito, F., Giannotti, F., Pedreschi, D. (eds) Machine Learning: ECML 2004. ECML 2004. Lecture Notes in Computer Science(), vol 3201. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-30115-8_35

[29] G. Biau, L. Devroye, and G. Lugosi, On the performance of clustering in Hilbert spaces, *IEEE Transactions on Information Theory* **54** (2008), no. 2, 781–790. https://doi.org/10.1109/TIT.2007.913516

[30] D. Pollard, Strong consistency of k-means clustering, *Annals of Statistics* **9** (1981), no. 1, 135–140. Available at: https://doi.org/10.1214/aos/1176345339.

[31] D. Pollard, A central limit theorem for k-means clustering, *Annals of Probability* **10** (1982), no. 4, 919–926. Available at: https://doi.org/10.1214/aop/1176993713.

[32] D. Pollard, Quantization and the method of k-means, *IEEE Transactions on Information Theory* **28** (1982), no. 2, 199–205. Available at: https://doi.org/10.1109/TIT.1982.1056481.

[33] A. Cuevas and R. Fraiman, Set estimation, in *New Perspectives in Stochastic Geometry*, W. S. Kendall and I. Molchanov (eds.), Oxford University Press, Oxford, 2009; online edn, Oxford Academic, 1 Feb. 2010. Available at: `https://doi.org/10.1093/acprof:oso/9780199232574.003.0011` (accessed 2 Sept. 2025).

[34] W. Polonik, Measuring mass concentrations and estimating density contour clusters—an excess mass approach, *Annals of Statistics* **23** (1995), no. 3, 855–881. Available at: `https://doi.org/10.1214/aos/1176324626`.

[35] Cuevas, A., González-Manteiga, W. and Rodríguez-Casal, A. (2006). Plug-in estimation of general level sets. Aust. N. Z. J. Stat., 48, 7–19

[36] William E. Wright. A formalization of cluster analysis. *Pattern Recognition*, 5(3):273–282, 1973. ISSN 0031-3203. `https://doi.org/10.1016/0031-3203(73)90048-4`.

[37] Jan Puzicha, Thomas Hofmann, and Joachim M. Buhmann. A theory of proximity based clustering: structure detection by optimization. *Pattern Recognition*, 33(4):617–634, 2000. ISSN 0031-3203. `https://doi.org/10.1016/S0031-3203(99)00076-X`.

[38] Juan José Sanguineti. Lógica *Ediciones Universidad de Navarra S.A. (EUNSA)*

[39] T. van Laarhoven and E. Marchiori, *Axioms for Graph Clustering Quality Functions*, Journal of Machine Learning Research, vol. 15, no. 6, pp. 193–215, 2014.

[40] G. W. Milligan, *A Monte Carlo study of thirty internal criterion measures for cluster analysis*, Psychometrika, vol. 46(2), pp. 187–199, 1981.

[41] F. B. Baker and L. J. Hubert, *Measuring the power of hierarchical cluster analysis*, Journal of the American Statistical Association, pp. 31–38, 1975.

[42] L. J. Hubert and J. R. Levin, *A general statistical framework for assessing categorical clustering in free recall*, Psychological Bulletin, vol. 83(6), p. 1072, 1976.

[43] J. MacQueen, *Some methods for classification and analysis of multivariate observations*, in L. Le Cam y J. Neyman (eds.), *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, vol. 2, pp. 281–297, University of California Press, Berkeley, 1967.

[44] J. A. Hartigan, *Clustering Algorithms*, John Wiley & Sons, New York, 1975.

[45] A. K. Jain y R. C. Dubes, *Algorithms for Clustering Data*, Prentice-Hall, Englewood Cliffs, NJ, 1988.

[46] R. Fraiman, B. Ghattas y M. Svarc, *Clustering Using Unsupervised Binary Trees: CUBT*, arXiv preprint, Computing Research Repository (CoRR), 2010.

[47] A. P. Dempster, N. M. Laird y D. B. Rubin, *Maximum likelihood from incomplete data via the EM algorithm (with discussion)*, Journal of the Royal Statistical Society, Series B, vol. 39, pp. 1–38, 1977.

[48] G. J. McLachlan y T. Krishnan, *The EM Algorithm and Extensions*, Wiley, New York, 1997.