Universidad de la República

Facultad de Ciencias Económicas y de Administración

Licenciatura en Estadística

Trabajo Final de Grado para la Licenciatura en Estadística

Aprendizaje Estadístico Supervisado basado en Núcleos

Autor: Sebastián Vallejo Massolo

Tutores:

Prof. Dr. Ricardo Fraiman

Prof. Dr. Leonardo Moreno

Tribunal:

. Dr. Ricardo Fraiman . Dr. Marco Scavino . Dr. Leonardo Moreno

Pando, Uruguay. Marzo, 2026

A la memoria de Edmundo Canalda,
pilar invaluable durante tantos años y gran mentor;
impulsor incondicional no solo de mi labor en la investigación matemática,
sino también de mi participación en los espacios donde se construye la
realidad común.

Agradecimientos

Este trabajo está dedicado a Verónica, Sol y Luz.

Deseo manifestar un reconocimiento muy especial a la Profesora Economista Alicia Guglielmo, cuyo consejo estratégico y apoyo constante constituyeron los pilares necesarios para el inicio y la prosecución de mis estudios en esta Licenciatura.

Al Dr. Richard Fariña, mi sincera gratitud por haberme permitido colaborar en diversos trabajos de investigación, tales como la reciente publicación sobre relatividad biológica extendida Fariña (2026). Su constante apoyo y agudeza intelectual han sido pilares fundamentales para acrecentar mi interés y fortalecer mi nivel de entrenamiento en la estadística matemática.

Al Dr. Marco Scavino, integrante del tribunal de esta monografía, por las enseñanzas recibidas a lo largo de las diversas asignaturas que tuve el privilegio de cursar bajo su tutela. Sus lecciones no solo me permitieron profundizar en los fundamentos teóricos de la disciplina, sino que también despertaron un especial interés por la obra de Alexandr Borovkov, cuya rigurosa exposición de la teoría de la probabilidad ha sido una referencia invaluable en mi formación académica.

Deseo expresar mi especial reconocimiento al Profesor Dr. Leonardo Moreno. Tras los años compartidos forjando un enriquecedor vínculo —tanto desde la experiencia de ser compañeros de estudio como desde sus enseñanzas como profesor—, quiero destacar hoy su invaluable labor, rigor metodológico y guía permanente en su rol de cotutor de esta monografía.

Finalmente, quisiera extender mi más profunda gratitud al Profesor Dr. Ricardo Fraiman. Ha sido un inmenso honor no solo haberme formado bajo su tutela como alumno en múltiples cursos, sino también contar con su invaluable supervisión y sabiduría para el desarrollo de este trabajo monográfico.

Resumen

Este trabajo explora varias técnicas de aprendizaje automático y sus elementos subyacentes de la teoría del aprendizaje estadístico, con un enfasis específico en el aprendizaje supervisado basado en kernels. La investigación se centra en un problema de regresión aplicado a datos quimiométricos, logrando resultados competitivos con publicaciones recientes en el campo. Se explora la transición de la minimización del riesgo empírico (ERM) a métodos de regularización como las máquinas de vectores de soporte (SVM) y los mínimos cuadrados regularizados kernel (KRLS). En una aplicación práctica, la investigación aplica diferentes técnicas de aprendizaje automático a un conjunto de datos proporcionado por un laboratorio en Montevideo, con el objetivo de predecir los valores de los componentes químicos a partir de datos de espectroscopia de infrarrojo cercano (NIRS).
El objetivo es desarrollar un sistema capaz de predecir automáticamente valores químicos basados en datos NIRS, con un coeficiente de determinación objetivo (R^2) de 0,92 establecido por expertos de campo que proporcionaron los datos. Los resultados muestran que al eliminar el ruido y reducir la dimensionalidad mediante coeficientes wavelet, SVM logra valores de R^2 que superan el objetivo.
Palabras clave: Quimiometría, Reducción de Dimensionalidad, Eliminación de Ruido, Métodos basados en Núcleos, Regresión, Aprendizaje Estadístico Supervisado

Abstract

This paper explores various machine learning techniques and their underlying statistical learning theory elements, with a specific emphasis on kernel-based supervised learning. The research focuses on a regression problem applied to chemometric data, achieving results competitive with recent publications in the field. It explores the transition from Empirical Risk Minimization (ERM) to regularization methods like Support Vector Machines (SVM) and Kernel Regularized Least Squares (KRLS). In a practical application, the research applies different machine learning techniques to a dataset provided by a laboratory in Montevideo, aiming to predict chemical component values from Near-Infrared Spectroscopy (NIRS) data. The goal is to develop a system capable of automatically predicting chemical values based on NIRS data, with a target coefficient of determination (R^2) of 0.92 set by field experts. The results

show that by eliminating noise and reducing dimensionality using wavelet coefficients, SVM achieves R^2 values exceeding the target for both training and test data.

Keywords Chemometrics, Dimensionality Reduction, Noise Elimination, Kernel-based Methods, Regression, Supervised Statistical Learning

Correo electrónico del autor sebastian.vallejo@gmail.com

Índice general

Índice de tablas

Índice de figuras

Introducción

Entender la inteligencia es uno de los problemas más importantes de la ciencia actual y es considerado el problema del siglo XXI (Poggio & Smale 2005) como lo fuera en la segunda mitad del siglo pasado el descifrar el código genético.

El problema del aprendizaje es la puerta de entrada para la comprensión de la inteligencia en cerebros y máquinas, para descubrir cómo funciona el cerebro humano y para fabricar máquinas inteligentes que aprendan de la experiencia, mejorando sus habilidades.

Las técnicas de aprendizaje estadístico hacen posible el desarrollo de software que puede ser adaptado rápidamente para manejar el creciente volumen de información y flujo de datos que nos circunda.

Este trabajo presenta distintas técnicas de Aprendizaje Automático, así como algunos de los elementos de la Teoría del Aprendizaje Estadístico que las sustentan. Haremos foco especialmente en técnicas de aprendizaje supervisado basadas en núcleos y presentaremos ejemplos del funcionamiento de una serie de algoritmos. Usando datos reales, aplicaremos distintos modelos de regresión para un conjunto de datos quimiométricos, obteniendo resultados que compiten con las últimas publicaciones del tema.

Herbert Simon, Premio Nobel en Economía de 1978, definió el aprendizaje (Simon 1983) de la siguiente forma:

"El aprendizaje denota cambios adaptativos en el sistema en el sentido que le permite a éste realizar la misma tarea o tareas sorteadas de la misma población más eficientemente y en forma mas efectiva la próxima vez."

En el contexto de este trabajo se hace referencia a aprendizaje como el proceso de inferir reglas generales a través de observar muestras.

El Aprendizaje Estadístico (Steinwart & Christmann 2008) es una formulación matemática específica del concepto general de aprendizaje, que busca relacionar cierto tipo de valores de entrada o medidas a ciertos valores de sali-

da o respuesta.

Resulta de interés en varias aplicaciones de la vida real conocer si existe una estructura de dependencia entre las entradas y las salidas y, si es así, qué relación funcional la describe.

Muchos organismos muestran alguna grado de habilidad para aprender. Por ejemplo, un niño puede aprender qué es un perro, si se le muestran ejemplos de los objetos que son perros y objetos que no lo son. No es necesario explicar ninguna regla sobre qué es lo que hace que un objeto sea un perro, simplemente pueden aprender el concepto de "perro" observando ejemplos.

La Teoría de Aprendizaje Estadístico no estudia el proceso de aprendizaje en los organismos vivos, sino que estudia el proceso de aprendizaje en abstracto.

Su objetivo es inferir una regla general que permita explicar la muestra observada, para luego generalizar sobre nuevos elementos no observados.

Más precisamente, el objetivo de la Teoría del Aprendizaje Estadístico (Boucheron *et al.* 2004b) es proporcionar un marco de trabajo para estudiar el problema de la inferencia. Para esto se realizan ciertos supuestos de naturaleza estadística sobre el fenómeno subyacente que genera los datos.

El proceso de inferencia inductiva puede ser resumido en los siguientes pasos:

- Observar el fenómeno.
- Construir un modelo del fenómeno.
- Hacer predicciones utilizando el modelo.

El objetivo del Aprendizaje Automático[1] es en concreto automatizar el proceso de inferencia inductiva, y el objetivo de la Teoría del Aprendizaje Estadístico es formalizarlo.

De esta forma, la Teoria del Aprendizaje Estadístico proporciona la base teórica a muchos de los algoritmos de Aprendizaje Automático. Pero esta no es la su única motivación (von Luxburg & Schölkopf 2008), sino una de corte filosófico, que consiste en buscar responder la pregunta de qué es lo que nos permite obtener conclusiones válidas a partir de los datos empíricos.

La Teoría del Aprendizaje Estadístico busca entonces responder preguntas fundamentales como:

- Cuáles tareas de aprendizaje pueden ser realizadas por computadoras en general.

[1] En la literatura se utiliza habitualmente el término inglés *Machine Learning*.

- Qué tipo de supuestos tenemos que establecer para que el aprendizaje estadístico sea exitoso.
- Cuáles son las propiedades clave que un algoritmo de aprendizaje necesita satisfacer para ser exitoso.
- Qué niveles de desempeño podemos esperar de los resultados de los algoritmos de aprendizaje.

Estructura del Trabajo

En la primera parte se desarrollan alguno de los conceptos fundamentales de aprendizaje y específicamente las características del aprendizaje supervisado, planteando formalmente el problema de la regresión.

Luego, se revisan definiciones y resultados en lo que hace a la minimización del error empírico, para luego plantear soluciones alternativas como la minimización del error estructural y la regularización.

Más adelante, se presentan los métodos Kernel partiendo del método básico de regresión lineal.

Se plantean resultados aplicables a la implementación de la selección del modelo.

Proseguimos realizando un estudio de las Máquinas de Vectores de Soporte[2] aplicadas tanto a problemas de clasificación como de regresión y el algoritmo de Mínimos Cuadrados Regularizados Kernel[3] aplicado a regresión.

En los capítulos dedicados ejemplos prácticos y la aplicación concreta se presentan distintas aplicaciones prácticas del Aprendizaje Automático y una aplicación con datos reales quimiométricos.

Finalmente se resumen las conclusiones obtenidas y se realizan comentarios sobre posibles investigaciones futuras.

En los anexos se presentan las demostraciones detalladas de los Teoremas de Representación (Scholkopf *et al.* 2001), de una cota para el Error Muestral (Wu *et al.* 2006) aplicable a la minimización del error empírico y un Teorema que nos permite obtener una solución alternativa para el problema de clasificación (Cucker & Zhou 2007). Finalmente se presenta un resumen de las lecturas complentarias realizadas sobre análisis funcional [4], análisis convexo y la historia

[2] En inglés *Support Vector Machines*: SVM.
[3] En inglés *Kernel Regularized Least Squares*: KRLS.
[4] En base al mathcamp del curso (Rifkin *et al.* 2003) y otras obras de referencia.

de la investigación del problema de aprendizaje[5].

[5]Preparado principalmente en base las introducciones de (Vapnik 1995) y (Vapnik 1998).

Capítulo 1

Aprendizaje supervisado

El proceso de aprendizaje está ligado directamente a la existencia de cierto objetivo a alcanzar. Muchas veces este objetivo de aprendizaje no puede ser especificado explícitamente por contar con información a priori insuficiente, es decir el objetivo del aprendizaje no esta completamente definido. En el caso opuesto, cuando el objetivo de aprendizaje esta dado en una forma explícita, no hay necesidad de aprendizaje dado que el objetivo puede ser alcanzado, por ejemplo, diseñando un sistema.

En este capítulo desarrollaremos una introducción a las distintas metodologías de aprendizaje, para luego hacer foco en el aprendizaje supervisado y los problemas de regresión y clasificación, abordando el problema del aprendizaje para introducirnos luego en algunos problemas clave a resolver.

1.1. Metodologías de aprendizaje

La característica distintiva del aprendizaje es que la carencia de información completa a priori, es decir, una definición incompleta sobre el objetivo del aprendizaje, es compensada por el necesario procesamiento de la información disponible.

Una posible clasificación de los métodos de aprendizaje surge de cuán completa sea la información disponible a priori (Tsypkin 1973). Se distiguen dos tipos de aprendizaje: supervisado y no supervisado.

- En el aprendizaje supervisado, se conoce la respuesta del sistema de aprendizaje y se usa la diferencia entre la respuesta deseada[1] y la obteni-

[1] Según el modelo utilizado.

da, es decir el error del sistema de aprendizaje, para corregir dicho comportamiento.

- En cambio, en el aprendizaje no supervisado, no se conoce la respuesta del sistema de aprendizaje, y de esta forma no se puede obtener una fórmula explícita ni utilizar el error del sistema de aprendizaje para mejorar su comportamiento.

Los problemas de aprendizaje (Hastie *et al.* 2009) pertenecen principalmente a una de estas dos categorías.

Se realiza aprendizaje supervisado cuando en la muestra disponible para cada observación de la medida predictora x_i (para $i = 1, ..., n$) hay asociada una medida de respuesta y_i.

Se desea ajustar entonces un modelo que relacione la respuesta con los predictores, para poder predecir la respuesta para observaciones futuras (predicción) o comprender mejor la relación entre la respuesta y los predictores (inferencia).

Varios métodos clásicos tales como la regresión lineal y la logística están en el dominio del aprendizaje supervisado (Puntanen 2008).

En cambio, el aprendizaje no supervisado describe una situación distinta: para cada observación se cuenta con un vector de medidas x_i pero no una respuesta asociada. Una herramienta de aprendizaje estadístico que podemos utilizar en este ámbito es el análisis de conglomerados[2] cuyo objetivo consiste en definir, en base a la muestra, cuándo las observaciones pertenecen a ciertos grupos relativamente diferenciados.

Algunos autores (Chapelle *et al.* 2006) definen un tercer grupo: los problemas de aprendizaje semi-supervisado se encuentran entre los dos anteriores y se caracterizan por la presencia de conjuntos de datos etiquetados y datos no etiquetados. Es el caso de problemas de análisis de datos que involucran muestras muy grandes en las cuales etiquetar[3] requeriría la intervención humana o procedimientos muy costosos, lo que ocasiona que quede una gran cantidad de datos sin etiquetar.

[2]En inglés *clustering*.
[3]Definir la variable de respuesta.

1.2. Características del aprendizaje supervisado

Cuando se utilizan computadoras para resolver problemas prácticos es usual que el método para obtener la salida requerida a partir de un conjunto de datos de entrada pueda se descrito en forma explícita.

La tarea del diseñador del sistema que implemente la solución requerida será traducir el método en una secuencia de instrucciones las cuales la computadora deberá seguir para lograr el efecto deseado.

Cuando se intenta utilizar las computadoras para resolver problemas más complejos, aparecen situaciones en las que no existe un método conocido para calcular la salida deseada a partir de un conjunto de entrada. Por ejemplo, modelar la relación entre el espectro cercano al infrarojo de una substancia y un valor químico medido, o la clasificación de una solicitud de crédito en la cual se desea saber quién pagará el préstamo y quién no a partir de ciertas características del solicitante.

Un método alternativo (Cristianini & Shawe-Taylor 2000, Capítulo 1) para resolver este tipo de problemas es que la computadora intente aprender la relación funcional entre las entradas y las salidas a partir de ejemplos. Estas técnicas son denominadas técnicas de aprendizaje y en particular cuando los ejemplos son pares de entradas y salidas, es llamado aprendizaje supervisado.

Entonces, en el aprendizaje supervisado, o aprendizaje a partir de muestras, en lugar de programar directamente el método de cálculo de la salida, una máquina es entrenada para realizar una tarea determinada sobre varios pares de entrada/salida (Evgeniou & Pontil 2000).

El proceso de entrenamiento consiste en elegir la función que mejor describa la relación entre las entradas y las salidas.

En un problema de aprendizaje supervisado, el sistema debe predecir las etiquetas de los patrones (Anthony & Bartlett 2009), donde las etiquetas pueden ser una clase o un número real[4]. Durante el entrenamiento el sistema recibe información parcial sobre la verdadera relación entre los patrones y sus etiquetas en la forma de un conjunto de patrones etiquetados.

Esta forma de aprendizaje tiene dos ventajas destacables: permite un ahorro importante de esfuerzo de diseño y puede ser utilizado para problemas que no pueden ser especificados en forma precisa, dado que su entorno cambia.

En este contexto, la tarea principal de la teoría de aprendizaje (Evgeniou & Pontil 2000) será determinar cuan bien la función elegida generaliza, o cuan

[4]También se incluye el caso para $\mathbb{R}^k$.

bien estima la salida para entradas no conocidas previamente.

1.3. El proceso generador de datos

Se cuenta con un conjunto de datos $\mathbf{z} = \{x_i, y_i\}_{i=1}^n$, que se asume provienen del conjunto $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ según cierta distribución de probabilidad desconocida.

La relación entre los elementos de $\mathcal{X}$ e $\mathcal{Y}$ es probabilística, dado que un elemento de $\mathcal{X}$ no determina únicamente un elemento de $\mathcal{Y}$ sino una distribución de probabilidad en $\mathcal{Y}$.

No son necesarios supuestos específicos en los espacios $\mathcal{X}$ e $\mathcal{Y}$, pero sí se realizan ciertos supuestos (von Luxburg & Schölkopf 2008) sobre el mecanismo que genera los puntos de entrenamiento. Asumimos que existe una distribución de probabilidad conjunta ρ sobre $\mathcal{X} \times \mathcal{Y}$ y que los elementos de la muestra de entrenamiento $\{x_i, y_i\}_{i=1}^n$ son sorteados en forma independiente con la distribución ρ.

Se destacan algunos supuestos importantes (von Luxburg & Schölkopf 2008) en este contexto:

- *No hay supuestos sobre ρ.* En el escenario estándar de la Teoria del Aprendizaje Estadístico (SLT [5]) no se realiza ningún supuesto sobre la distribución de probabilidad ρ: puede ser cualquier distribución sobre $\mathcal{X} \times \mathcal{Y}$. En este sentido, SLT trabaja en un escenario agnóstico distinto al de la estadística paramétrica, donde usualmente se asume que la distribución de probabilidad pertenece a cierta familia de distribuciones y cuyo objetivo es estimar sus parámetros. En la década de 1920 (Vapnik 1998) comenzó el análisis de métodos de inferencia estadística. Dos eventos principales señalaron su comienzo.

  Fisher (Fisher 1992) introdujo los principales modelos de inferencia estadística en el marco de trabajo unificado de la estadística paramétrica. El describió diferentes problemas para estimar funciones en base a datos dados como problemas de estimación de parámetros de modelos específicos (paramétricos) y sugirió un método para estimar los parámetros desconocidos para esos modelos: el método de la máxima verosimilitud.

  Por otra parte Glivenko, Cantelli y Kolmogorov comenzaron un análisis general de inferencia estadística. Glivenko y Cantelli (Salnikov 2021) probaron que la función de distribución empírica siempre converge a la

[5] En inglés SLT: *Statistical Learning Theory.*

función de distribución. Kolmogorov, por su parte, encontró la tasa asintótica exacta de convergencia (Arak 1982). Esta tasa es rápida (exponencial) e independiente de la función de distribución (desconocida).
Estos trabajos determinaron dos enfoques principales en el abordaje de la inferencia estadística. La inferencia paramétrica, que busca crear métodos estadísticos simples de inferencia que pueden ser usados para problemas de la vida real y una inferencia general, que busca encontrar un método para cualquier problema de inferencia estadística.
La filosofía que lleva a la creación de la estadística de inferencia paramétrica está basada en el principio que el investigador conoce bastante bien el problema a analizar. Conoce la ley física que genera las propiedades estocásticas de los datos y la función a encontrar salvo un número finito de parámetros. Estimar esos parámetros usando los datos es la escencia del problema de la inferencia estadística.
La filosofía que lleva a la inferencia estadística general es diferente: no se cuenta con suficiente información a priori sobre la ley subyacente al problema o sobre la función que se desea aproximar. Es necesario encontrar un método para obtener una función aproximada en esta nueva situación.

- *Etiquetas no determinísticas debido al ruido o al solapamiento de las clases.* Las etiquetas y_i no son necesariamente funciones determinísticas de los objetos x_i, sino que pueden ser al azar. Hay dos razones para esto. La primera es que el proceso de generación de datos puede tener ruido en la obtención de las etiquetas. Es decir, puede ocurrir que una etiqueta y_i que tenemos en la muestra de entrenamiento pueda ser incorrecta. Este es un supuesto realista e importante.

  Por ejemplo para generar los datos de entrenamiento de un detector de correo eletrónico no deseado (spam[6]) se le pide a humanos que etiqueten correos manualmente en dos clases "spam" y "no spam", y los humanos se pueden equivocar.

  La segunda razón por la cual se puede llegar a etiquetas no determinísticas es el caso de las clases solapadas. Por ejemplo, tomemos la tarea de predecir el género de una persona basándonos en su estatura. Es claro que una persona de 1.80 metros de altura, puede ser hombre o mujer, por lo tanto no podemos asignar una etiqueta única y al dato de entrada $x=$

[6]El *spam* es una comunicación no solicitada y masiva que se envía de forma cuestionable y no deseada.

1.80.

- *Muestreo independiente.* Es un supuesto importante de la Teoría del Aprendizaje Estadístico que los datos sean muestreados en forma independiente. Este es un supuesto fuerte, que es justificado en muchas aplicaciones, pero no en todas. Por ejemplo consideremos el reconocimiento de patrones de dígitos manuscritos. Dadas algunas imágenes de dígitos manuscritos, la tarea es entrenar una máquina que automáticamente reconozca nuevos dígitos escritos. Para esta tarea el conjunto de entrenamiento consiste usualmente de una gran colección de dígitos escritos por distintas personas. Aquí es seguro asumir que los dígitos forman una muestra independiente proveniente de la población de todos los dígitos manuscritos. Como un ejemplo en el que el supuesto de independencia es violado, consideremos el caso del descubrimiento de drogas. Esta es un área de la farmacéutica donde se busca identificar compuestos que podrían ser útiles para diseñar una nueva droga. El aprendizaje automático es usado para este objetivo: los datos de entrenamiento consisten en un conjunto de compuestos químicos x_i con una etiqueta asociada y_i que indica si el compuesto es útil o no para el diseño de drogas. Resultaría costoso encontrar cuál compuesto químico posee ciertas propiedades que lo constituyan en una droga apropiada porque esto implicaría realizar muchos experimentos de laboratorio. De esta forma, solo unos pocos compuestos x_i tienen una etiqueta conocida y_i y estos compuestos fueron elegidos cuidadosamente en primer lugar. Asi es que no podemos asumir que los x_i son una muestra representativa obtenida en forma independiente de alguna distribución de los compuestos químicos, dado que los compuestos etiquetados han sido elegidos a mano según un proceso que no es al azar.

- *La distribución ρ es fija.* En el escenario estándar del Aprendizaje Estadístico no tenemos ningún parámetro de tiempo. En particular, no asumimos ningún orden en particular de las muestras de entrenamiento ni que la distribución de probabilidad cambia en el tiempo. Este supuesto no sería correcto si se tratara con series de tiempo.

- *La distribución ρ es deconocida durante el proceso de aprendizaje.* Es importante tener en cuanta que en el momento del entrenamiento la distribución no es conocida. Si ρ fuera conocida, el aprendizaje sería trivial dado que se podría obtener el mejor clasificador con una fórmula dada.

En su lugar, tendremos solo acceso a la ρ en forma indirecta, observando a la muestra de entrenamiento. Intuitivamente esto quiere decir que si obtenemos suficientes elementos de la muestra, podemos estimar todas las propiedades importantes de ρ en forma bastante precisa, pero de todas formas presentando errores. La Teoría del Aprendizaje Estadístico proporciona un marco de trabajo para realizar enunciados sobre ese error.

1.4. El Problema del Aprendizaje

El problema de aprendizaje consiste entonces en, dado el conjunto de datos de entrenamiento, encontrar un estimador, es decir una función $f : \mathcal{X} \to \mathcal{Y}$ que pueda ser usada para, dado un nuevo valor de $x \in \mathcal{X}$ predecir un valor y.

Por ejemplo $\mathcal{X}$ podría ser el conjunto de todas las imágenes posibles, $\mathcal{Y}$ el conjunto $\{-1, 1\}$ y $f(x)$ una función indicatriz que indica si una imagen contiene cierto objeto ($y = 1$) o no ($y = -1$).

Otro ejemplo podría ser el caso donde $\mathcal{X}$ es el conjunto de los espectros posibles de una substancia, e $\mathcal{Y}$ son los valores posibles de presencia de cierto compuesto químico, y $f(x)$ la función regresión que mapea los espectros con el valor químico.

El método para resolver el problema de aprendizaje consistirá en definir una funcion de riesgo que mida el error promedio asociado con el estimador para luego poder buscar el estimador con el menor riesgo.

Definimos la función de pérdida o discrepancia $V(y, f(x))$ que mide el error que se produce al predecir y con $f(x)$ (Vapnik 1999).

El error promedio o riesgo esperado es[7]:

$$\mathcal{E}_f = \mathcal{R}(f) = \int_{\mathcal{X}\times\mathcal{Y}} V(y, f(x)) d\rho.$$

Asumimos que el riesgo esperado esta definido en una amplia clase de funciones $\mathcal{F}$ y anotaremos f_0 a la función que minimiza el riesgo esperado en $\mathcal{F}$. La función f_0 que es nuestro estimador ideal, es conocida como la función objetivo. Formalizaremos el problema específicamente para la regresión, haciendo referencia a que puede ser demostrada su existencia y unicidad[8](Cucker & Zhou

[7]En la literatura aparecen distintas notaciones para referirnos a las funciones de pérdida y los espacios.

[8]Bajo ciertas restricciones que detallaremos más adelante.

2007, págs. 10, 11 y 46).

1.4.1. Formalización del problema de regresión

Los casos de aprendizaje supervisado cuando $\mathcal{Y}$ es finito o en particular $\mathcal{Y} = \{+1, -1\}$ son denominados clasificación. Cuando $\mathcal{Y} = \mathbb{R}^n$ son llamados problemas de regresión.

El objetivo de aprendizaje consiste en aprender de la muestra de entrenamiento una función $f : \mathcal{X} \to \mathcal{Y}$ que nos permita predecir el valor de y para un nuevo x.

Se puede formalizar el problema de la regresión utilizando una función de pérdida cuadrática (Cucker & Zhou 2007).

Dado que se estudia el aprendizaje a partir de muestreo al azar, el objeto primario de nuestro desarrollo es entonces una medida de probabilidad ρ desconocida, que gobierna el muestreo.

Sea $\mathcal{X}$ un espacio métrico compacto e $\mathcal{Y} = \mathbb{R}^k$. Por conveniencia manejaremos en general $k = 1$.

Sea ρ una medida de probabilidad de Borel sobre $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$.

Aplicando la definición de riesgo definimos entonces, para ese caso particular, el error de generalización utilizando el error de mínimos cuadrados de f definido por:

$$\mathcal{E}(f) = \mathcal{E}_\rho(f) = \int_{\mathcal{Z}} (f(x) - y)^2 d\rho.$$

Es decir que definimos:

$$V(x, f(x)) = (f(x) - y)^2.$$

Para cada entrada $x \in \mathcal{X}$ y salida $y \in \mathcal{Y}$, $(f(x) - y)^2$ es el error que se ocasiona al usar f como un modelo del proceso generador de y a partir de x. Este es un error local. Integrando sobre $\mathcal{X} \times \mathcal{Y}$ con respecto a ρ se está promediando este error entre todos los pares posibles (x, y).

El problema esta planteado: ¿Cuál es la función f que minimiza el error $\mathcal{E}(f)$?.

Para cada $x \in \mathcal{X}$, definimos $\rho(y|x)$ como la probabilidad condicional medida sobre $\mathcal{Y}$ con respecto a x.

Definimos también a ρ_X la probabilidad marginal de ρ en $\mathcal{X}$, es decir, la medida en $\mathcal{X}$ definida por $\rho_X(S) = \rho(\pi^{-1}(S))$, siendo $\pi : \mathcal{X} \times \mathcal{Y} \to \mathcal{X}$ la proyección.

Para cada función integrable $\varphi : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ una versión del Teorema de Fubini nos permite relacionar ρ, $\rho(y|x)$ y ρ_X de la siguiente forma:

$$\int_{\mathcal{X}\times\mathcal{Y}} \varphi(x,y)d\rho = \int_{\mathcal{X}} \left(\int_{\mathcal{Y}} \varphi(x,y)d\rho(y|x) \right) d\rho_X.$$

Definimos $f_\rho : \mathcal{X} \to \mathcal{Y}$ como la función regresión de ρ como :

$$f_\rho(x) = \int_{\mathcal{Y}} y d\rho(y|x).$$

Para cada $x \in \mathcal{X}$, $f_\rho(x)$ es el promedio de la coordenada y en $\{x\} \times \mathcal{Y}$.

Asumiremos que la función resgresión es acotada.

Fijemos ahora $x \in \mathcal{X}$ y consideremos la función de $y \to \mathbb{R}$ que mapea y en $(y - f_\rho(x))$.

Dado que la esperanza de esta función es 0, su varianza es

$$\sigma^2(x) = \int_{\mathcal{Y}} (y - f_\rho(x))^2 d\rho(y|x).$$

Ahora se promedia en $\mathcal{X}$ para obtener

$$\sigma_\rho^2 = \int_{\mathcal{X}} \sigma^2(x) d\rho_X = \mathcal{E}(f_\rho).$$

Proposición 1. *Primera descomposición del error*

$$\mathcal{E}(f) = \int_{\mathcal{X}} (f(x) - f_\rho(x))^2 d\rho_X + \sigma_\rho^2.$$

Demostración. De la definicion de $f_\rho(x)$ para cada $x \in \mathcal{X}$, $\int_{\mathcal{Y}} (f_\rho(x) - y) = 0$. Luego

$$\begin{aligned}
\mathcal{E}(f) &= \int_{\mathcal{Z}} (f(x) - f_\rho(x) + f_\rho(x) - y)^2 \\
&= \int_{\mathcal{X}} (f(x) - f_\rho(x))^2 + \int_{\mathcal{X}} \int_{\mathcal{Y}} (f_\rho(x) - y)^2 \\
&\quad +2 \int_{\mathcal{X}} \int_{\mathcal{Y}} (f(x) - f_\rho(x))(f_\rho(x) - y) \\
&= \int_{\mathcal{X}} (f(x) - f_\rho(x))^2 + \sigma_\rho^2 + 2 \int_{\mathcal{X}} (f(x) - f_\rho(x)) \int_{\mathcal{Y}} (f_\rho(x) - y) \\
&= \int_{\mathcal{X}} (f(x) - f_\rho(x))^2 + \sigma_\rho^2.
\end{aligned}$$

□

El sumando $\int_{\mathcal{X}}(f(x) - f_\rho(x))^2$ es un promedio en $\mathcal{X}$ del error ocasionado por utilizar f como un modelo para f_ρ.

Además, al ser σ_ρ^2 independiente de f, la proposición anterior implica que f_ρ tiene el menor error posible entre todas las funciónes $f : \mathcal{X} \to \mathcal{Y}$. De esta forma σ_ρ^2 es una cota inferior del error $\mathcal{E}$ y sólo depende de la medida ρ.

Entonces nuestro objetivo se convierte en "aprender", es decir encontrar una buena aproximación de f_ρ a partir de muestras al azar sobre $\mathcal{Z}$.

Consideremos ahora la muestra. Sea

$$\mathbf{z} \in \mathcal{Z}^m, \mathbf{z} = \{(x_1, y_1), ..., (x_m, y_m)\}.$$

una muestra en $\mathcal{Z}^m$, es decir m elementos extraídos en forma independiente segun ρ. Definimos el error empírico de f con respecto a $\mathbf{z}$ como

$$\mathcal{E}_{\mathbf{z}}(f) = \frac{1}{m}\sum_{i=1}^{m} V(y_i, f(x_i)).$$

En particular, para el caso en que $V(y, f(x)) = (f(x) - y)^2$

$$\mathcal{E}_{\mathbf{z}}(f) = \frac{1}{m}\sum_{i=1}^{m}(f(x_i) - y_i)^2.$$

1.5. Espacios de Hipótesis y Funciones Objetivo

En esta sección presentaremos formalmente dónde viven las funciones objetivo y objetivo empírica y demostraremos su existencia.

Como se indica en (Cucker & Zhou 2007), los procesos de aprendizaje no tienen lugar en el vacío. Alguna estructura debe estar presente en el comienzo del proceso. Asumiremos que esta estructura consiste en una clase de funciónes.

El objetivo del proceso de aprendizaje será encontrar la mejor aproximación de f_ρ dentro de esa clase.

Sea $\mathscr{C}(\mathcal{X})$ el espacio de Banach[9] de funciones contínuas en $\mathcal{X}$ equipado con

[9]En un anexo de este trabajo se presenta un resumen de Análisis Funcional.

la norma infinita:

$$\|f\|_\infty = sup_{x\in\mathcal{X}}||f(x)||.$$

Consideraremos un subconjunto $\mathcal{H}$ de $\mathscr{C}(\mathcal{X})$ que llamaremos espacio de hipótesis donde los algoritmos trabajarán para encontrar la mejor aproximación de f_ρ.

Si $f_\rho \in \mathcal{H}$ las cosas se simplificarían, pero en general no podemos asumir siquiera que $f_\rho \in \mathscr{C}(\mathcal{X})$ por lo que tendremos que considerar una función objetivo $f_\mathcal{H} \in \mathcal{H}$.

Definimos $f_\mathcal{H}$ como cualquier función que minimice el error $\mathcal{E}$ entre las $f \in \mathcal{H}$, es decir, cualquier minimizador de (para el caso del error cuadrático):

$$min_{f\in\mathcal{H}} \int_\mathcal{Z} (f(x) - y)^2 \, d\rho.$$

Dado que $\mathcal{E}(f) = \int_\mathcal{X} (f - f_\rho)^2 + \sigma_\rho^2$ entonces $f_\mathcal{H}$ es también un optimizador de

$$min_{f\in\mathcal{H}} \int_\mathcal{X} (f - f_\rho)^2 \, d\rho_X.$$

Sea una muestra $\mathbf{z} \in \mathcal{Z}^m$. Definimos la función objetivo empírica $f_{\mathcal{H},\mathbf{z}} = f_\mathbf{z}$ como una función que minimiza el error empirico $\mathcal{E}_\mathbf{z}(f)$ entre todas las $f \in \mathcal{H}$, es decir un optimizador de:

$$min_{f\in\mathcal{H}} \frac{1}{m} \sum_{i=1}^{m} (f(x_i) - y_i)^2.$$

Este problema de minimización depende de ρ solo a través de su dependencia de $\mathbf{z}$, pero una vez conocida $\mathbf{z}$ no es necesaria más información sobre ρ para buscar $f_\mathbf{z}$.

1.5.1. Existencia de las funciones objetivo

Es posible demostrar la existencia de $f_\mathcal{H}$ y $f_\mathbf{z}$ bajo ciertos supuestos sobre $\mathcal{H}$. (Cucker & Zhou 2007, pág. 10 y 11).

Sea $f : \mathcal{X} \to \mathcal{Y}$ y $\mathbf{z} \in \mathcal{Z}^n$. Definimos el defecto de f con respecto a $\mathbf{z}$ como:

$$L_\mathbf{z}(f) = \mathcal{E}(f) - \mathcal{E}_\mathbf{z}(f).$$

El error teórico $\mathcal{E}(f)$ no puede ser medido, pero el $\mathcal{E}_\mathbf{z}(f)$ sí. Por lo tanto una cota en el defecto antes definido nos permitirá acotar en cuánto se separará el

error teórico del error observado. Más adelante en este trabajo se establecerán cotas a estos errores.

Dadas f_1,$f_2 \in \mathscr{C}(\mathcal{X})$ estimaremos la cantidad:

$$|L_{\mathbf{z}}(f_1) - L_{\mathbf{z}}(f_2)|.$$

Proposición:

Para $j = 1, 2, |f_j(x) - y| \leq M$ en un conjunto de medida completa $U \subseteq \mathcal{Z}$ luego para todas muestra $\mathbf{z} \in U^n$ (Gallagher *et al.* 2022) (Kolountzakis 2023).

$$|L_{\mathbf{z}}(f_1) - L_{\mathbf{z}}(f_2)| \leq 4M\|f_1 - f_2\|_\infty.$$

Demostración:

Dado que $(f_1(x) - y)^2 - (f_2(x) - y)^2 = (f_1(x) + f_2(x) - 2y)(f_1(x) - f_2(x))$, entonces:

$$\begin{aligned} |\mathcal{E}(f_1) - \mathcal{E}(f_1)| &= \left| \int_{\mathcal{Z}} (f_1(x) + f_2(x) - 2y)(f_1(x) - f_2(x)) d\rho \right| \\ &\leq \int_{\mathcal{Z}} |(f_1(x) - y) + (f_2(x) - y)| \|f_1 - f_2\|_\infty d\rho \\ &\leq 2M\|f_1 - f_2\|_\infty. \end{aligned}$$

También para toda muestra $\mathbf{z} \in U^n$ tenemos:

$$\begin{aligned} |\mathcal{E}_{\mathbf{z}}(f_1) - \mathcal{E}_{\mathbf{z}}(f_1)| &= \frac{1}{n} \left| \sum_{i=1}^{n} (f_1(x) + f_2(x) - 2y)(f_1(x) - f_2(x)) \right| \\ &\leq \frac{1}{n} \sum_{i=1}^{n} |(f_1(x) - y) + (f_2(x) - y)| \|f_1 - f_2\|_\infty \\ &\leq 2M\|f_1 - f_2\|_\infty. \end{aligned}$$

Entonces:

$$\begin{aligned} |L_{\mathbf{z}}(f_1) - L_{\mathbf{z}}(f_2)| &= |\mathcal{E}(f_1) - \mathcal{E}_{\mathbf{z}}(f_1) + \mathcal{E}(f_2) - \mathcal{E}_{\mathbf{z}}(f_2))| \\ &\leq |\mathcal{E}(f_1) - \mathcal{E}(f_2)| + |\mathcal{E}_{\mathbf{z}}(f_1) - \mathcal{E}_{\mathbf{z}}(f_2)| \leq 4M\|f_1 - f_2\|_\infty. \end{aligned}$$

A partir de este resultado, se puede demostrar que $\mathcal{E}, \mathcal{E}_{\mathbf{z}} : \mathcal{H} \to \mathbb{R}$ son contínuas.

Aplicando la continuidad anterior y suponiendo que $\mathcal{H}$ es compacto, final-

mente se puede demostrar la existencia de $f_{\mathcal{H}}$ y $f_{\mathbf{z}}$. (Cucker & Zhou 2007, página 11)

Estas funciones $f_{\mathcal{H}}$ y $f_{\mathbf{z}}$ no son necesariamente únicas. Sin embargo, se plantea un resultado de unicidad para $f_{\mathcal{H}}$ demostrado en (Cucker & Zhou 2007, página 46) cuando $\mathcal{H}$ es convexo que a continuación transcribimos.

Sea $\mathcal{H}$ un subconjunto convexo de $\mathcal{L}^2(X)$ tal que $f_{\mathcal{H}}$ existe. Entonces $f_{\mathcal{H}}$ es única como un elemento de L^2 y, para toda $f \in \mathcal{H}$,

$$\int_{\mathcal{X}} (f_{\mathcal{H}} - f)^2 \leq \mathcal{E}_{\mathcal{H}(f)}.$$

En particular, si $\rho_{\mathcal{X}}$ es no-degenerada, entonces $f_{\mathcal{H}}$ es única en $\mathcal{H}$.

1.5.2. Ejemplos de espacios de hipótesis

A continuación mencionamos ejemplos de espacios de hipótesis (Cucker & Zhou 2007).

Ejemplo 1: Espacio de los Polinomios homogéneos.

Sea $\mathcal{H}_d = \mathcal{H}_d(\mathbb{R}^{n+1})$ el espacio lineal de polinomios homogéneos de grado d en $x_0, x_1, ..., x_n$. Sea $X = S(\mathbb{R}^{n+1})$ la esfera unitaria n-dimensional. Un elemento en $\mathcal{H}_d$ define una función $\mathcal{X} \to \mathbb{R}$ que puede ser escrita como:

$$f = \sum_{\alpha=d} w_\alpha x^\alpha.$$

Aquí $\alpha = (\alpha_0, ..., \alpha_n) \in \mathbb{N}^{n+1}$ es un "multi-índice", $|\alpha| = \alpha_0 + ... + \alpha_n$ y $x^\alpha = x_0^{\alpha_0}...x_n^{\alpha_n}$. Por lo tanto $\mathcal{H}_d$ es un espacio vectorial de dimensión finita.

Podemos considerar

$$\mathcal{H} = \{f \in \mathcal{H}_d \text{ con } \|f\|_\infty \leq 1\},$$

como un espacio de hipótesis. Debido a la escala $f(x) = x^\alpha f(x)$, tomar el límite $\|f\|_\infty \leq 1$ no causa pérdida.

Los $x_1...x_n$ son las variables del polinomio homogéneo (García *et al.* 2024) y los respectivos w_i son los coeficientes.

Ejemplo 2: Espacios de funciones finito-dimensionales.

Formulamos una generalización del primer ejemplo. Sean $\varphi_1, ..., \varphi_N \in \mathscr{C}(\mathcal{X})$ y $\mathbb{E}$ el subespacio lineal de $\mathscr{C}(\mathcal{X})$ generado por $\{\varphi_1, ..., \varphi_N\}$. En este caso tomaremos $\mathcal{H} = \{f \in \mathbb{E} | \|f\|_\infty \leq R\}$ para cierto $R > 0$.

Los siguientes ejemplos manejan espacios lineales infinito-dimensionales, tema que desarrollamos en el anexo correspondiente que se refiere a Análisis Funcional.

Definición: Sea $J : E \to F$ una aplicación lineal entre los espacios de Banach E y F. Decimos que J está acotada cuando existe $b \in \mathbb{R}$ tal que para todo $x \in E$ con $||x|| = 1$, se tiene $||J(x)|| \leq b$. La norma del operador de J es:

$$||J|| = \sup_{||x||=1} ||J(x)||.$$

Si J no está acotada, entonces escribimos $||J|| = \infty$. Decimos que J es compacta cuando la clausura $\overline{J(B)}$ de $J(B)$ es compacta para cualquier conjunto acotado $B \subset E$.

Ejemplo 3: Espacios de Sobolev (Adams 1975): Sea X un dominio en $\mathbb{R}^n$ con frontera suave. Para cada $s \in \mathbb{N}$ podemos definir un producto interno en $\mathcal{C}^\infty(X)$ por:

$$\langle f, g \rangle_s = \int_X \sum_{|\alpha| \leq s} D^\alpha f D^\alpha g.$$

Aquí estamos integrando con respecto a la medida de Lebesgue μ en X heredada del espacio Euclidiano. Denotaremos por $||\cdot||_s$ la norma inducida por $\langle \cdot, \cdot \rangle_s$. Nótese que cuando $s = 0$, el producto interno anterior coincide con el de $L^2(X)$. Es decir, $||\cdot||_0 = ||\cdot||_{L^2}$.
Definimos el espacio de Sobolev $\mathcal{H}^s(\mathcal{X})$ como la completación de $\mathcal{C}^\infty(X)$ con respecto a la norma $||\cdot||_s$. El Teorema de inmersión de Sobolev afirma que para todo $r \in \mathbb{N}$ y todo $s > n/2 + r$, la inclusión:

$$J_s : \mathcal{H}^s(\mathcal{X}) \hookrightarrow C^r(\mathcal{X}),$$

está bien definida y acotada, en particular, para todo $s > \frac{n}{2}$.

Del Teorema de Rellich se obtenemos que si $\mathcal{X}$ es compacto, esta última inclusión es compacta también. Así, si B_R denota la bola cerrada de radio R en $\mathcal{H}^s(\mathcal{X})$, podemos tomar $\mathcal{H}_{R,s} = H = \overline{J_s(B_R)}$.

1.5.3. Reproducing Kernel Hilbert Spaces

A continuación reproducimos un conjunto de definiciones y observaciones aplicables a los Espacio de Hilbert de Kernels (Cucker & Zhou 2007, Sección 2.4).

Definiciones. Sea $\mathcal{X}$ un espacio métrico.

- Decimos que $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ es simétrica cuando $K(x,t) = K(t,x)$ para todo $x, t \in \mathcal{X}$ y se dice semidefinida positiva cuando para todos los conjuntos finitos $\mathbf{x} = \{x_1, x_2, \cdots, x_n\} \subset \mathcal{X}$ luego la matrix $n \times n$ $K[\mathbf{x}]$ cuya entrada (i,j) es $K(x_i, x_j)$ es semidefinida positiva.
- Decimos que K es un Kernel de Mercer si es contínua, simétrica y semidefinida positiva.
- La matriz $K[\mathbf{x}]$ es llamada el Gramiano[10] (o determinante de Gram) de K en el conjunto $\mathbf{x}$.

Se observa que ser semidefinida positiva implica que $K(x,x) \geq 0$ para cada $x \in \mathcal{X}$. Definimos:

$$\mathbf{C}_K := \sup_{x \in \mathcal{X}} \sqrt{K(x,x)}.$$

Luego

$$\mathbf{C}_K := \sup_{x,t \in \mathcal{X}} \sqrt{|K(x,t)|}.$$

Dado que la matriz $K[\{x,t\}]$ es semidefinida positiva, para todo $x, t \in \mathcal{X}$,

$$(K(x,t))^2 \leq K(x,x)K(t,t).$$

Para $x \in \mathcal{X}$ sea K_x la función

$$\begin{aligned} K_x : \quad & \mathcal{X} \to \mathbb{R} \\ & t \mapsto K(x,t). \end{aligned}$$

Teorema .Existe un único espacio de Hilbert $(\mathcal{H}_K, \langle \cdot, \cdot \rangle_{\mathcal{H}_K})$ de funciones en $\mathcal{X}$ que satisfacen las siguientes condiciones:

- Para todo $x \in \mathcal{X}$, $K_x \in \mathcal{H}_K$.
- El span del conjunto $\{K_x | x \in \mathcal{X}\}$ es denso en $\mathcal{H}_K$.
- Para toda $f \in \mathcal{H}_K$ y $x \in \mathcal{X}$, $f(x) = \langle K_x, f \rangle_{H_k}$.

Adicionalmente $\mathcal{H}_K$ consiste de funciones continuas y la inclusión $I_K : \mathcal{H}_K \to \mathscr{C}(\mathcal{X})$ esta acotada con $\|I_k\| \leq C_K$.

[10] En (Cucker & Zhou 2007) se utiliza el término en inglés *Gramian*.

Se puede consultar la demostración de este Teorema en la página 23 de (Cucker & Zhou 2007).

El espacio de Hilbert $\mathcal{H}_K$ en el Teorema anterior es referido como un Espacio de Hilbert de Kernels de Reproducción[11]. Destacamos la tercer propiedad del mismo, a la que nos referiremos como propiedad de reproducción.

1.6. Error muestral, de aproximación y de generalización

Pasamos ahora a desarrollar algunas definiciones y resultados relevantes tomados de (Cucker & Zhou 2007).

Dado un espacio de hipótesis $\mathcal{H}$, el error en $\mathcal{H}$ de una función $f \in \mathcal{H}$ es el error nomalizado:

$$\mathcal{E}_{\mathcal{H}}(f) = \mathcal{E}(f) - \mathcal{E}(f_{\mathcal{H}}).$$

Siguiendo el desarrollo planteado en (Cucker & Zhou 2007) transcribimos el siguiente resultado: Proposicición. Sea $f : X \to Y$ una función. Entonces,

$$\mathcal{E}(f) = \int_X (f(x) - f_\rho(x))^2 \, d\rho_X + \sigma_\rho^2.$$

Es de destacar que este tratamiento de descomposición del error ya es tratado en (Cox 1988) y (Niyogi & Girosi 1996) citados en (Cucker & Zhou 2007).

Ya hemos establecido entonces que σ_ρ^2 es una cota inferior del error $\mathcal{E}$ que sólo depende de la medida ρ.

El error de generalización $\mathcal{E}(f_z)$ de f_z depende de ρ, $\mathcal{H}$, la muestra z, y del esquema que define de f_z.

La distancia cuadrada $\int_X (f_z - f_\rho)^2 d\rho_X$ es el exceso de error de generalizacion de f_z.

Consideremos ahora la suma $\mathcal{E}_{\mathcal{H}}(f_z) + \mathcal{E}(f_{\mathcal{H}})$. El segundo sumando depende de la eleccion de $\mathcal{H}$ pero es independiente de la muestra. Se le denomina error de aproximación.

El error de la aproximación se puede considerar como la suma

$$\mathcal{A}(\mathcal{H}) + \sigma_\rho^2,$$

[11] En inglés *Reproducing Kernel Hilbert Space*: RKHS.

donde

$$\mathcal{A}(\mathcal{H}) = \int_X (f_{\mathcal{H}} - f_\rho)^2 d\rho_X.$$

De esta forma σ_ρ^2 es una cota inferior del error de aproximación.

El primer sumando, $\mathcal{E}_{\mathcal{H}}(f_z)$, es llamado el error muestral o error de estimación.

Nuestro objetivo pasa entonces de estimar $\mathcal{E}(f_z)$ a descomponerse en otros dos problemas: estimar el error muestral y el de aproximación.

Según (Cucker & Zhou 2007) estimar el error muestral $\mathcal{E}_H(f_z)$ en el espacio $\mathcal{H}$, y su dependencia de f_ρ es independiente de la muestra z. Por el contrario, los límites para el error de aproximación $\mathcal{E}_H$ no dependerán de propiedades de f_ρ si bien el límite se mantendrá con una confianza de al menos $1 - \delta$.

Si bien la dependencia del comportamiento de f_ρ parece inevitable en las estimaciones del error de aproximación (y por lo tanto en el error de generalización $\mathcal{E}(f_z)$ de f_z).

El objetivo es planteado en (Cucker & Zhou 2007) consiste en que dada una muestra z, seleccionar un espacio de hipótesis $\mathcal{H}$ y calcular laf_z sin suposiciones sobre f_ρ, estableciendo límites sobre $\int_X (f_z - f_\rho)^2 \, d\rho_X$ que sean razonablemente buenos dependiendo del comportamiento de f_ρ.

Capítulo 2

Minimización del Error Muestral para Regresión

Se pueden establecer cotas para el error muestral de un espacio de hipótesis compacto y convexo. Estas propiedades de convexidad y compacidad son requisitos para varios resultados que mencionaremos a lo largo de este trabajo.

Para un tamaño fijo de la muestra, el error muestral se incrementa con el tamaño de $\mathcal{H}$. Estableceremos que este comportamiento con respecto a una medida especial del tamaño de $\mathcal{H}$: su capacidad medida por los covering numbers (Cucker & Zhou 2007) Capítulo 3 y (Cucker & Smale 2001).

Comenenzamos entonces abordando un algoritmo de aprendizaje basado en minimización de riesgo empírico[1] para un problema de regresión asociado a una función de pérdida genérica ψ citando a (Wu *et al.* 2006).

Sea $\mathcal{H}$ un conjunto medible uniformemente acotado de funciones en $\mathcal{X}$ y $\psi : \mathbb{R} \to \mathbb{R}_+$ una función de pérdida de regresión. La función empírica objetivo f_z asociada a la muestra $z = \{(x_i, y_i)\}_{i=1}^m \in \mathcal{Z}^m$ se define como:

$$f_z := \arg\min_{f \in \mathcal{H}} \frac{1}{m} \sum_{i=1}^{m} \psi(y_i - f(x_i)).$$

Definimos $\mathcal{H}$ como el espacio de hipótesis para el problema de regresión.

Si definimos el error empírico o riesgo asociado a la función de pérdida ψ como

$$\mathcal{E}_z(f) = \frac{1}{m} \sum_{i=1}^{m} \psi(y_i - f(x_i)),$$

entonces $f_z = \arg\min_{f \in \mathcal{H}} \mathcal{E}_z(f)$.

[1] ERM: en inglés *Empirical Risk Minimization.*

La teoría de ERM estudia cómo f_z se aproxima (con respecto al error) a la función objetivo definida como:

$$f_{\mathcal{H}} := \arg\min_{f\in\mathcal{H}} \mathcal{E}(f) = \arg\min_{f\in\mathcal{H}} \int \psi(y - f(x))dP.$$

Para describir esta aproximación descomponemos el error muestral:

$$\mathcal{E}(f_z) - \mathcal{E}(f_{\mathcal{H}}) = \mathcal{E}(f_z) - \mathcal{E}_z(f_z) + \mathcal{E}_z(f_z) - \mathcal{E}_z(f_{\mathcal{H}}) + \mathcal{E}_z(f_{\mathcal{H}}) - \mathcal{E}(f_{\mathcal{H}}).$$

Dado que f_z minimiza el error empírico en $\mathcal{H}$, se cumple entonces que $\mathcal{E}_z(f_z) - \mathcal{E}_z(f_{\mathcal{H}}) \leq 0$. Tenemos entonces la siguiente descomposición del ERM:

$$\mathcal{E}(f_z) - \mathcal{E}(f_{\mathcal{H}}) \leq \{\mathcal{E}(f_z) - \mathcal{E}_z(f_z)\} + \{\mathcal{E}_z(f_{\mathcal{H}}) - \mathcal{E}(f_{\mathcal{H}})\}.$$

Definición: Sea S un espacio métrico y $\eta > 0$. Definimos el *covering number* $\mathcal{N}(S, n)$ como el mínimo $l \in \mathbb{N}$ tal que existen l discos en S de radio η que cubren S. Cuando S es compacto este numero es finito.

Definición: Sea $M > 0$ y ρ una medida de probabilidad en Z. Decimos que el conjunto $\mathcal{H}$ de funciones desde $X\mathbb{R}$ es *M-acotado* cuando:

$$sup|f(x) - y| \leq M,$$

se cumple casi en todo Z, siendo y la salida o resultados de la muestra.

Exponemos entonces el resultado principal, que se demuestra en el Anexo B de este trabajo.

Sea $\mathcal{H}$ un subconjunto compacto y convexo de $\mathscr{C}(X)$. Si $\mathcal{H}$ es M-acotado, luego, para todo $\varepsilon > 0$,siendo m el tamaño de la muestra,

$$\text{Prob}_{z\in Z^m}\{\mathcal{E}_{\mathcal{H}}(f_z) \leq \varepsilon\} \geq 1 - \mathcal{N}\left(\mathcal{H}, \frac{\varepsilon}{12M}\right) \exp\left\{-\frac{m\varepsilon}{300M^2}\right\}.$$

Recordemos la definición de la función defecto: $L_z(f) = \mathcal{E}(f) - \mathcal{E}_z(f)$.

2.1. Inecuaciones Exponenciales en Probabilidad

Se transcriben algunos resultados importantes, que manifiestan distintos enfoques de acotar el error muestral tomadas (Lin & Bai 2011), (Cucker & Zhou 2007) y (Boucheron *et al.* 2004a).

2.1.1. Desigualdad de Bennett

Teorema 1. Desigualdad de Markov

Sea ξ una variable alatoria no negativa y $t > 0$, luego

$$\xi \geq \xi \mathbb{1}_{\xi \geq t} \geq t \mathbb{1}_{\xi \geq t}.$$

Dado que $\text{Prob}\{\xi \geq t\} \leq \frac{\text{E}(\xi)}{t}$.

Teorema 2. Desigualdad de Bennett

Sean $\{\xi_i\}_{i=1}^m$ variables independientes en un espacio de probabilidad Z con medias $\{\mu_i\}$ y varianzas $\{\sigma_i^2\}$. Definimos $\Sigma^2 := \sum_{i=1}^m \sigma_i^2$.

Si para cada i, $|\xi_i - \mu_i| \leq M$ se cumple casi en cualquier lado, luego para cada $\varepsilon > 0$ se cumple que:

$$\text{Prob}\left\{\sum_{i=1}^m [\xi_i - \mu_i] > \varepsilon\right\} \leq \exp\left\{-\frac{\varepsilon}{M}\left\{\left(1 + \frac{\Sigma^2}{M\varepsilon}\right)\ \log\left(1 + \frac{M\varepsilon}{\Sigma^2}\right) - 1\right\}\right\}.$$

Demostración (Cucker & Zhou 2007).

Sin perder generalidad, asumimos $\mu_i = 0$. Luego la varianza de ξ_i es $\sigma_i^2 = E\left(\xi_i^2\right)$.

Sea c una constante arbitraria positiva que será determinada más adelante en la demostración. Luego:

$$I := \text{Prob}\left\{\sum_{i=1}^m [\xi_i - \mu_i] > \varepsilon\right\} = \text{Prob}\left\{\exp\left\{\sum_{i=1}^m c\xi_i\right\} > e^{c\varepsilon}\right\}.$$

Por la desigualdad de Markov y la independencia de ξ_i , tenemos:

$$I \leq e^{-c\varepsilon}\text{E}\left(\exp\left\{\sum_{i=1}^n c\xi_i\right\}\right) = e^{-c\varepsilon}\prod_{i=1}^m \text{E}\left(e^{c\xi_i}\right).$$

Dado que $|\xi_i| \leq M$ casi en todas partes y $E(\xi_i) = 0$, el desarrollo de Taylor para e^{ξ_i} se expresa de la forma siguiente:

$$E(e^{\xi_i}) = 1 + \sum_{\ell=2}^{+\infty} \frac{E(\xi_i^\ell)}{\ell!} \leq 1 + \sum_{\ell=2}^{+\infty} \frac{M^\ell \sigma_i^2}{\ell!}.$$

Aplicando que $1 + x \leq e^x$, entonces:

$$E(e^{\xi_i}) \leq \exp\left(\sum_{\ell=2}^{+\infty} \frac{c^\ell M^\ell - cM^\ell \sigma_i^2}{\ell!}\right) = \exp\left(\frac{e^{cM} - 1 - cM}{M^2}\sigma_i^2\right),$$

y por lo tanto

$$1 \leq \exp\left(-c\xi_i + \frac{e^{cM} - 1 - cM}{M^2}\sum_i \sigma_i^2\right).$$

Ahora elegimos la constante c como el minimizador de la cota de la derecha:

$$c = \frac{1}{M}\log\left(1 + \frac{M\xi}{\sum_i \sigma_i^2}\right).$$

Es decir, $e^{cM} - 1 = \frac{M\varepsilon}{\Sigma^2}$. Con esta elección,

$$1 \leq \exp\left\{-\frac{\varepsilon}{M}\left[1 + \left(\frac{\Sigma^2}{M\varepsilon}\right)\log\left(1 + \frac{M\varepsilon}{\Sigma^2}\right) - 1\right]\right\}.$$

Esto prueba la desigualdad planteada.

2.1.2. Cotas expresadas como inecuaciones exponenciales

Citamos ahora otras inecuaciones que nos permiten acotar el error tomadas de (Cucker & Zhou 2007), incluyendo la de Bennett, antes demostrada.

Proposición: Sea $\{\xi_i\}_{i=1}^m$ un conjunto de variables aleatorias independientes en un espacio de probabilidad Z con medias $\{\mu_i\}_{i=1}^m$ y varianzas $\{\sigma_i^2\}_{i=1}^m$, y que satisfacen $|\xi_i(z) - \mathbb{E}[\xi_i]| \leq M$ para todo i y casi todo $z \in Z$. Sea $\Sigma^2 := \sum_{i=1}^m \sigma_i^2$. Entonces, para todo $\epsilon > 0$, se tiene:

Desigualdad generalizada de Bennett

$$\mathbb{P}\left(\left|\sum_{i=1}^m (\xi_i - \mu_i)\right| > \epsilon\right) \leq 2\exp\left(-\frac{\epsilon}{2M}\log\left(1 + \frac{M\epsilon}{\Sigma^2}\right)\right).$$

Desigualdad de Bernstein

$$\mathbb{P}\left(\left|\sum_{i=1}^m (\xi_i - \mu_i)\right| > \epsilon\right) \leq 2\exp\left(-\frac{\epsilon^2}{2(\Sigma^2 + \frac{1}{3}M\epsilon)}\right).$$

Desigualdad de Hoeffding

$$\mathbb{P}\left(\left|\sum_{i=1}^{m}(\xi_i - \mu_i)\right| > \epsilon\right) \leq 2\exp\left(-\frac{\epsilon^2}{2mM^2}\right).$$

2.2. El error muestral

Definimos entonces el error muestral como:

$$\mathcal{E}_{\mathcal{H}}(f_{\mathbf{z}}) = \mathcal{E}(f_{\mathbf{z}}) - \mathcal{E}(f_{\mathcal{H}}).$$

Podemos expresarlo como:

$$\mathcal{E}_{\mathcal{H}}(f_{\mathbf{z}}) = \mathcal{E}(f_{\mathbf{z}}) - \mathcal{E}_{\mathbf{z}}(f_{\mathbf{z}}) + \mathcal{E}_{\mathbf{z}}(f_{\mathbf{z}}) - \mathcal{E}_{\mathbf{z}}(f_{\mathcal{H}}) + \mathcal{E}_{\mathbf{z}}(f_{\mathcal{H}}) - \mathcal{E}(f_{\mathcal{H}}).$$

Dado que $f_{\mathbf{z}}$ minimiza $\mathcal{E}_{\mathbf{z}}$ en $\mathcal{H}$ luego $\mathcal{E}_{\mathbf{z}}(f_{\mathbf{z}}) - \mathcal{E}_{\mathbf{z}}(f_{\mathcal{H}}) \leq 0$ Luego:

$$\mathcal{E}_{\mathcal{H}}(f_{\mathbf{z}}) \leq \{\mathcal{E}(f_{\mathbf{z}}) - \mathcal{E}_{\mathbf{z}}(f_{\mathbf{z}}) + \mathcal{E}_{\mathbf{z}}(f_{\mathcal{H}}) - \mathcal{E}(f_{\mathcal{H}})\}.$$

Entonces, para encontrar una cota para $\mathcal{E}_{\mathcal{H}}(f_{\mathbf{z}})$ alcanza con acotar $\mathcal{E}(f_{\mathbf{z}}) - \mathcal{E}_{\mathbf{z}}(f_{\mathbf{z}})$ y $\mathcal{E}_{\mathbf{z}}(f_{\mathcal{H}}) - \mathcal{E}(f_{\mathcal{H}})$. Este es un resultado clave, que se utiliza en el Teorema que a continuación se expone.

2.2.1. Acotando el error para una función ψ genérica

Definición previa: Decimos que ψ es una función *s-Lipschitz* $(0 < s \leq 1)$ en $[-M, M]$ si existe una constante $C \geq 0$ tal que

$$|\psi(t) - \psi(t')| \leq C|t - t'|^s.$$

Teorema

En (Wu *et al.* 2006) se presenta una cota para una función ψ genérica, que cumpla las condiciones que siguen.

Sea $\mathcal{H}$ un subconjunto de $\mathcal{C}(\mathcal{X})$ [2] tal que para toda $f \in \mathcal{H}$ se cumple que $|f(x) - y| \leq M$ casi seguramente. Si $\psi : \mathbb{R} \to \mathbb{R}$ es una función de pérdida regresiva que satisface la condición de s-Lipschitz antes mencionada.

$$\text{Prob}\{\mathcal{E}_{\mathcal{H}}(f_{\mathbf{z}}) \leq \epsilon\} \geq 1 - \mathcal{N}(\mathcal{H}, \left(\frac{\epsilon}{8C}\right)^{1/s})2\exp\left\{-\frac{m\epsilon^2}{128C^2M^{2s}}\right\}.$$

[2] Conjuntos de las funciones contínuas (Kowalczyk 2014).

La demostración se presenta en el Apéndice B de este trabajo.

2.3. Medidas de Capacidad

Capacidad es una cantidad que mide la *complejidad* del espacio (Evgeniou *et al.* 2002). Se definen distintas medidas de capacidad en la teoría. Las más populares son la VC-dimension (Vapnik & Chervonenkis 1971) o versiones sensibles a escala (Kearns & Schapire 1994).

Se puede formular (von Luxburg & Schölkopf 2008) cotas para la generalización en términos del coeficiente de separación $\mathcal{N}$.

La desventaja es que es difícil de evaluar. Pero existen distintos conceptos de capacidad, obviamente, con sus ventajas y desventajas.

El más conocido es la dimension VC, cuyo objetivo principal es caracterizar el comportamiento del crecimiento del coeficiente de separación usando solamente un valor.

Decimos que una muestra Z_n de tamano n es **separada**[3] por una clase de funciones $\mathcal{F}$ si la clase de funciones puede realizar cualquier etiquetado sobre la muestra proporcionada, es decir $|\mathcal{F}_{Z_n}| = 2^n$.

La dimensión VC (Vapnik 1982) de $\mathcal{F}$ que se anota como como $VC(\mathcal{F})$ se define como el número n mas grande tal que exista una muestra de tamaño n la cual es separada por $\mathcal{F}$. Formalmente:

$$VC(\mathcal{F}) = max\{n \in \mathbb{N} \text{tal que} |\mathcal{F}_{Z_n}| = 2^n \text{ para alguna } Z_n\}$$

Si no existe el máximo se dice que la dimension VC es infinita.

Lemma 2.1. *(Vapnik 1982) Sea $\mathcal{F}$ una clase de funciones con dimesión VC finita d.*
Luego

$$\mathcal{N}(\mathcal{F}, n) \leq \sum_{i=0}^{d} \binom{n}{i},$$

para todo $n \in \mathbb{N}$. En particular, para todo $n \geq d$ tenemos que

$$\mathcal{N}(\mathcal{F}, n) \leq \left(\frac{n}{d}\right)^d.$$

La importancia de este enunciado radica en que si $n \geq d$ entonces el coefi-

[3]En los textos originales en inglés se utiliza el termino shattered, que se puede traducir como destrozado.

ciente de separación se comporta como una funcion polinómica de el tamaño n de la muestra.

Concluimos entonces que en la minimización del error muestral para regresión, podemos concluir que para un tamaño fijo de la muestra, el error muestral aumenta con el tamaño del espacio de hipótesis. Para acotar este error, se presentan inecuaciones exponenciales como las desigualdades de Hoeffding y Bennett. La capacidad, medida por los covering numbers, es clave para establecer cotas para el error muestral de un espacio de hipótesis compacto y convexo. En este contexto, se presenta un teorema que proporciona una cota para una función de pérdida genérica que satisface la condición de Lipschitz.

Capítulo 3

Métodos Kernel

Los métodos kernel son un poderoso conjunto de técnicas para descubrir relaciones no lineales en los datos. Combinan la eficiencia y las propiedades bien entendidas de los algoritmos lineales con la flexibilidad para modelar patrones complejos. La clave es mapear los datos en un espacio de mayor dimensión, llamado *espacio de características*[1], donde las relaciones lineales pueden capturar la estructura no lineal deseada. Esto se logra implícitamente mediante el uso de una *función kernel* que calcula productos internos en este espacio de alta dimensión directamente a partir de los datos originales.

Se tratarán a continuación cuatro pasos clave de este enfoque:

- Incorporación de datos en un espacio de características.
- Detección de relaciones lineales en este espacio de características.
- Implementación de algoritmos que se basan únicamente en productos internos, no en vectores de características explícitos.
- Cálculo eficiente de productos internos directamente a partir de datos originales utilizando funciones kernel.

Esta estrategia ofrece un enfoque modular, que combina el conocimiento específico del dominio codificado en la función kernel con algoritmos de aprendizaje robustos de propósito general.
La eficiencia de los métodos kernel, que requieren únicamente recursos computacionales polinómicos incluso cuando el espacio integrado crece exponencialmente, los hace particularmente atractivos.

[1] En inglés *Feature Space*.

Ilustraremos estos principios utilizando el ejemplo del aprendizaje supervisado con regresión de mínimos cuadrados. Esto demostrará cómo los métodos kernel proporcionan un marco poderoso y flexible para descubrir patrones complejos en los datos.

Según (Shawe-Taylor & Cristianini 2004a, Chapter 2) la implementación de un método Kernel consta de dos etapas:

- Un módulo que realiza el mapeo en el espacio de características.
- Un algoritmo de aprendizaje diseñado para encontrar patrones lineales en ese espacio.

Hay dos razones (Shawe-Taylor & Cristianini 2004a) por las cuales este enfoque debiera funcionar. Primero, detectar relaciones lineales ha sido el foco de gran parte de la investigación en estadístico y Aprendizaje Estadístico durante décadas, por lo que los algoritmos resultantes son tanto bien entendidos como eficientes. En segundo lugar se cuenta con un atajo computacional que hace posible representar patrones lineales eficientemente en espacios de alta dimensión para asegurar el poder de representación adecuado. El atajo es la denominada función kernel.

De esta forma es posible detectar patrones estables en forma robusta y eficiente a partir de una muestra finita.

La estrategia es embeber los datos en un espacio donde los patrones pueden ser descubiertos como relaciones lineales.

Esto se realiza de forma modular. Tenemos entonces dos componentes:

- El primer componente de mapeo es definido en forma implícita por la llamada función kernel. Este componente dependera de los tipos de datos específicos y dominio de conocimiento referente a los patrones esperados para la fuente de datos en particular.
- El segundo componente de análisis de patrones de propósito general, robusto. Es mas, típicamente viene acompañado con el análisis estadístico de su estabilidad. El algoritmo es tambien eficiente, que requiere recursos de cómputo polinómicos en el tamaño y el número de los datos, inclusive cuando la dimensión del espacio embebido crezca en forma exponencial.

Esta estrategia recuerda el enfoque de la ingenieria de software para los sistemas de aprendizaje, dado que se divide la tarea en subcomponentes donde

se puede reutilizar los módulos clave.

Presentaremos a través del ejemplo de aprendizaje supervisado, utilizando la regresion de mínimos cuadrados, los principales ingredientes de los métodos kernel.

Se detallan a continuación cuatro aspectos clave del enfoque planteado:

- Los datos predictores son embebidos en un espacio vectorial denominado espacio de caracterísicas (feature space).
- Se observan relaciones lineales con las imágenes de los datos predictores en el espacio de características.
- Se implementan algoritmos de tal forma en que los puntos proyectados no son necesarios, solamente sus productos internos a pares.
- Los productos internos pueden ser calculados en forma eficiente directamente a partir de los datos predictores originales.

De esta forma, si bien optimizaremos funciones lineales, con este enfoque se desarrollarán un extenso y flexible conjunto de herramientas eficientes y bien fundamentadas por métodos para descubrir relaciones no lineales en los datos a través del mapeo de embebido no lineal.

3.1. Regresión Lineal

"... la mayoría de los otros métodos de regresión son en realidad simplemente elaboraciones o modificaciones de la regresión lineal. Es casi imposible *entender*, como opuesto de *usar*, redes neuronales o support vector machines sin una buena comprensión de la metodología de la regresión lineal." Extraído del prefacio de (Weisberg 2013).

La función de regresión y mínimos cuadrados ordinarios(OLS) El modelo de regresión lineal (Hocking 1996) se basa en el supuesto de que la variable de respuesta y está relacionada linealmente con los datos de entrada x.

Planteamos el problema de encontrar una función lineal:

$$g(x) = \langle w, x \rangle = w'x = \sum_{i=1}^{d} w_i x_i,$$

que interpole mejor una muestra de entrenamiento $S = \{(x_1, y_1), ..., (x_n, y_n)\}$ de puntos x_i provenientes de $X \subseteq \mathbb{R}^n$ con sus correspondientes etiquetas y_i provenientes de $Y \subseteq \mathbb{R}$.

En el caso exacto, cuando los datos fueron generados de $(x, g(x))$, para $g(x) =< w, x >$ y hay exactamente $d = n$ puntos linealmente independientes, es posible encontrar los parámetros w resolviendo el sistema de ecuaciones lineales:

$$Xw = Y,$$

donde utilizamos la matriz X donde las filas son los vectores fila $x'_1, ..., x'_l$ y Y es el vector $(y_1, ..., y_n)'$ con $n = d$.

Si la muestra tiene d elementos igual la dimensión de x_i $(n = d)$, obtener w se reduce a resolver el sistema de ecuaciones $Xw = Y$. El caso en que $n > d$ es el problema estudiado por Gauss, conocido como aproximación de mínimos cuadrados, que implica resolver el sistema $X'Xw = X'Y$, conocido como ecuaciones normales. Si existe la inversa de $X'X$ la solución puede ser expresada como:

$$w = (X'X)^{-1}X'Y$$

Si hay menos puntos que dimensiones, hay varios posibles w que describen los datos exactamente, siendo necesario un criterio para elegirlo. En este caso elegiremos el vector w que tenga la norma mas pequeña.

Si hay mas puntos que dimensiones y hay ruido en el proceso de generación no encontraremos una función exacta siendo necesario entonces un criterio de aproximación. En este caso elegiremos la función que genere el menor error.

En general, si trabajamos con conjuntos de datos pequeños con ruido, una combinación de las estrategias anteriores será necesario.

Buscamos entonces una función con un error de entrenamiento pequeño. La suma de los cuadrados de los errores la más usada como medida de discrepancia total entre los datos de entrenamiento y una función en particular

$$L(g, S) = L(w, S) = \sum_{i=1}^{n} (y_i - g(x_i))^2 = (y - Xw)'(y - Xw).$$

Este es un problema conocido, que se resuelve obteniendo las ecuaciones normales:

$$X'Xw = X'y.$$

Si la inversa de $X'X$ existe, la solucion puede ser expresada como

$$w = (X'X)^{-1}X'y.$$

Entonces, para minimizar la pérdida cuadratica del interpolante lineal, se necesita mantener tantos parámetros como dimensiones. Para esto, es necesario entonces resolver un sistema $d \times d$ de ecuaciones lineales, una operación de costo d^3.

La salida a predecir para un nuevo punto de datos, puede ser calculada usando la función de predicción:

$$g(x) =< w, x > .$$

La representacion dual de esta solución la podemos plantear como

$$w = (X'X)^1X'y = X'X(X'X)^2X'y = X'\alpha,$$

conviertiéndola en una combinación lineal de los puntos de entrenamiento.

Siguiendo con el razonamiento propuesto en (Cristianini & Shawe-Taylor 2000) podemos realizar la observación que continuación se desarrolla.

Observación (Pseudoinversa) Si $X'X$ es singular, se puede utilizar la pseudoinversa. Esta permite calcular el w que satisface la ecuaciones normales minimizando la norma. Podemos también intercambiar el tamaño de la norma contra la pérdida. Este es el enfoque utilizado por la regresión ridge.

Inestabilidad de la solución OLS Cuando la matriz X no es de rango completo, $X'X$ será singular, y por lo tanto los estimadores OLS de w no serán únicos. La singularidad puede producirse cuando la matriz X tiene columnas colineales o cuando hay mas variables que observaciones (Izenman 2008).

3.1.1. Métodos de regresión lineal sesgados

Una forma de solucionar el problema anterior, es abandonar el requerimiento de que el estimador de w sea insesgado. Existen diversos estimadores sesgados que son superiores al OLS del punto de vista del MSE cuando se presentan estos problemas (Izenman 2008).

3.2. Explicando KRLS: distintos enfoques para su interpretación

A continuación presentamos los distintos puntos de vista a través de los cuales podemos interpretar el uso de los kernels, según lo indicado en (Izenman 2008).

3.2.1. Similitud

Se puede interpretar a la función kernel como una medida de similitud entre dos patrones de entrada.

Por ejemplo, para el caso del kernel gaussiano la función está dada por:

$$k(x_j, x_i) = e^{-\frac{||x_j - x_i||^2}{\sigma^2}},$$

siendo $||x_j - x_i||$ la distancia euclidiana entre los vectores x_i y x_j.

Dos características importantes de este kernel son:

- Alcanza valor máximo 1 solamente cuando $x_i = x_j$.
- Se acerca a 0 cuando se agranda la distancia euclidiana entre x_i y x_j.

Según el Teorema de representación, la función objetivo puede ser aproximada por alguna función en el espacio de funciones representado por

$$f(x) = \sum_{i=1}^{n} c_i k(x, x_i),$$

dónde $k(x, x_i)$ mide la similaridad entre nuestro punto de interés x y los vectores de entrada x_i siendo c_i el peso de cada patrón de entrada.

La intuición clave de este enfoque es que no modela a y_i como una función lineal de x_i. En cambio, apalanca[2] la informacion sobre la similaridad entre las observaciones.

Consideremos un punto de conjunto de test x^* en el cual queremos evaluar la función para ciertos patrones de entrada x_i y pesos c_i.

Para este punto de test la función predictora está dada por:

$$f(x^*) = c_1 k(x^*, x_1) + c_2 k(x^*, x_2) + \cdots + c_n k(x^*, x_n).$$

[2]En las referencias en inglés se utiliza el verbo *to leverage*.

Por lo tanto el resultado de la función será una combinación lineal de las similaridades del punto objetivo a las observaciones.

A diferencia de los Modelos Lineales Generalizados[3]donde la salida es una suma ponderada de variables independientes, aquí nos basamos en la premisa de que la información está codificada en la similaridad entre las observaciones, donde observaciones más similares, se espera que tengan salidas más similares.

3.2.2. Superposición de Gaussianas

Podemos observar que para el caso del kernel gaussiano, la función $k(\cdot, x_i)$ traza una curva Gaussiana centrada en x_i. De esta forma, la función predictora f puede ser considerada como la superposición de curvas Gaussianas, centradas en los puntos de la muestra x_i y escaladas por los pesos c_i.

Según este punto de vista, dado un set de datos de entrenamiento, la función objetivo es aproximada ubicando Gaussianas sobre cada uno de los puntos de la muestra y escalándolas hasta que la superficie acumulada aproxime la función objetivo.

Hemos abordado un conjunto de técnicas para descubrir relaciones complejas en los datos. La implementación de un método Kernel consta de dos etapas: un módulo que realiza el mapeo en el espacio de características y un algoritmo de aprendizaje diseñado para encontrar patrones lineales en ese espacio. Los métodos Kernel constan de cuatro aspectos clave: los datos predictores son embebidos en un espacio de características, se observan relaciones lineales con las imágenes de los datos predictores en el espacio de características, se implementan algoritmos de tal forma en que los puntos proyectados no son necesarios, solamente sus productos internos a pares, y el cálculo eficiente de productos internos directamente a partir de datos originales utilizando funciones kernel. La intuición clave de este enfoque es que se apalanca la información sobre la similaridad entre las observaciones.

[3]Del inglés *Generalized Linear Models*: GLM.

Capítulo 4

De ERM a Regularización

Como vimos anteriormente usamos el principio de Minimización del Error Empírico[1] para solucionar problemas específicos de aprendizaje.

Una expresión general[2] que nos permite acotar el error es presentada en (Evgeniou *et al.* 2002) es la siguiente con probabilidad η:

$$R[f] < R_{\text{emp}}[f] + \Phi\left(\sqrt{\frac{h}{\ell}}, \eta\right).$$

Siendo h la capacidad y Φ una función creciente en $\frac{h}{\ell}$ y η.

ERM funciona en forma eficaz cuando tenemos una muestra grande (Vapnik 1999), donde el error real está cercano al valor del error empírico. De esta forma un valor reducido del riesgo empírico implica un valor reducido del riesgo esperado, dado que para un valor grande de $\frac{h}{\ell}$ el segundo sumando del lado derecho de la inecuación anterior se hace más pequeño.

4.1. Minimización del Riesgo Estructural

Pero, cuando $\frac{h}{\ell}$ se hace pequeño, un menor error empírico no garantiza que se reduzca el error esperado. En este caso, para minimizar el riesgo o error, requeriremos de un nuevo principio, basado en la minimización simultánea de los dos términos de la inecuación antes mencionada. El primer sumando depende del riesgo empírico y el segundo depende de la complejidad del conjunto de funciones en que buscamos (espacio de hipótesis).

[1] En inglés *Empirical Risk Minimization*: ERM.

[2] Una expresión más ajustada de esta cota puede ser encontrada por ejemplo en (Vapnik 1999).

De esta forma, un método alternativo a ERM para lograr mejor generalización consiste en buscar la mejor forma en que se compensan el riesgo empírico y la complejidad del conjunto de funciones.

Este método es la Minimización del Error Estructural [3] (Evgeniou *et al.* 2002) (Pontil 2003) que consiste en definir una secuencia de espacios de hipótesis anidados $\mathcal{H}_1 \subset \mathcal{H}_2 \subset \cdots \subset \mathcal{H}_M$, donde el espacio $\mathcal{H}_i$ tiene una capacidad finita h_i y menor que la de los espacios anteriores, es decir $h_1 \leq h_2 \leq \cdots \leq h_M$.

Por ejemplo, el espacio $\mathcal{H}_k$ podría ser el conjunto de los polinomios de grado k, o el conjunto de splines de k nodos.

Usando entonces una secuencia anidada de espacios de hipótesis, SRM consiste en elegir el minimizador del riesgo empírico en el espacio $\mathcal{H}_{k^*}$ para el cual la cota del error estructural medido por el segundo sumando de la inecuación anterior es minimizada. Más información de SRM puede ser obtenida en (Devroye *et al.* 1996) y (Vapnik 1999).

4.2. Alternativas al SRM

La implementación del método SRM antes descrito no es práctica (Evgeniou *et al.* 2002) (Pontil 2003) porque requeriría buscar la solución en una gran cantidad, en principio infinita, de problemas de optimización con restricciones.

Definiremos ahora el error regularizado como:

$$\mathcal{E}_\gamma(f) = \int_{\mathcal{Z}} V(y, f(x)) d\rho + \gamma ||f||_K^2.$$

Para una muestra $\mathbf{z}$ definimos el error empírico regularizado como:

$$\mathcal{E}_{\mathbf{z},\lambda}(f) = \frac{1}{m} \sum_{i=1}^{m} V(y_i, f(x_i)) + \gamma ||f||_K^2.$$

La constante γ es conocida como el parámetro de regularización.

Citando a (Evgeniou & Pontil 2000) el funcional H[f] contiene tanto el riesgo empírico asi como la norma (complejidad o suavidad) de f en el RKHS, en forma similar a los funcionales considerados en la teoría de la regularización (Tikhonov & Arsenin 1977). El parámetro de regularización penaliza las funciones de gran capacidad: cuando más grande sea, más pequeña será la norma RKHS de la solución.

[3] SRM: en inglés *Structural Risk Minimization*

4.3. Apliciones concretas de la regularización

Según (Evgeniou *et al.* 2002) buscamos la minimización de

$$\frac{1}{l}\sum_{i=1}^{l} V(y_i, f(x_i)) + \lambda||f||_K^2.$$

Utilizando distintas formas de la función se obtienen distintos métodos según (Evgeniou *et al.* 2002).

- Redes de regularización:

$$V(y_i, f(x_i)) = (y_i - f(x_i))^2.$$

- Clasificación SVM:

$$V(y_i, f(x_i)) = |1 - y_i f(x_i)|_+,$$

donde $|u|_+ = u$ si $u > 0$ o cero en otro caso.

- Regresión SVM:

$$\mathcal{V}(y_i, f(\mathbf{x}_i)) = |y_i - f(\mathbf{x}_i)|_\varepsilon.$$

La función de pérdida ε-insensible está definida como:

$$\mathcal{L}_\varepsilon(y_i, f(\mathbf{x}_i)) = |y_i - f(\mathbf{x}_i)|_\varepsilon,$$

donde

$$|x|_\varepsilon = \begin{cases} 0, & \text{si } |x| < \varepsilon \\ |x| - \varepsilon, & \text{si no.} \end{cases}$$

4.3.1. Redes de regularización

El esquema de aproximación que surge de la minimización del funcional cuadrático:

$$\frac{1}{l}\sum_{i=1}^{l}(y_i - f(x_i))^2 + \lambda||f||_K^2,$$

para un λ fijo es una forma especial de regularización. Es posible demostrar (ver por ejemplo (Girosi *et al.* 1998)) que los coeficientes c_i del minimizador antes definido se resuelve mediante el siguiente sistema lineal de ecuaciones:

$$(G + \lambda I)c = y,$$

donde I es la matriz identidad, y hemos definido

$$(y)_i = y_i,$$
$$(c)_i = c_i,$$
$$(G)_{ij} = K(x_i, x_j).$$

Dado que los coeficientes c_i satisfacen un sistema lineal, la Ecuación puede reescribirse como:

$$f(x) = \sum_{i=1}^{l} y_i b_i(x),$$

con $b_i(x) = \sum_{j=1}^{l}(G + \lambda I)_{ij}^{-1} K(x_i, x)$. La Ecuación anterior nos proporciona la representación dual.

Nótese la diferencia con la representación dual: en la primera Ecuación los coeficientes c_i se aprenden de los datos, mientras que en la segunda las funciones base b_i se aprenden, siendo el coeficiente de la expansión igual a la salida de los ejemplos. Referimos a (Girosi *et al.* 1998) para más información sobre la representación dual.

4.3.2. Máquinas de Vectores de Soporte

Discutiremos ahora las Máquinas de Vectores de Soporte[4] (Cortes & Vapnik 2009), (Vapnik 1999). Distinguiremos entre problemas de salida real (regresión) y salida binaria (clasificación). El método de regresión SVM corresponde a la siguiente minimización:

$$\min_f \frac{1}{l} \sum_{i=1}^{l} [y_i - f(x_i)]^2 + \lambda ||f||_K^2.$$

Mientras que el método de clasificación SVM corresponde a:

[4]En inglés *Support Vector Machines*: SVM.

$$\min_f \frac{1}{l} \sum_{i=1}^{l} [1 - y_i f(x_i)]_+ + \lambda ||f||_K^2.$$

Una propiedad notable de las SVM es que las respectivas funciones de pérdida conducen a soluciones dispersas. Esto significa que, a diferencia del caso de las Redes de Regularización, típicamente solo una pequeña fracción de los coeficientes c_i en la Ecuación inicialmente definida son distintos de cero. Los puntos de datos X_i asociados con los c_i no nulos se denominan vectores de soporte. Si todos los puntos de datos que no son vectores de soporte fueran descartados del conjunto de entrenamiento, se encontraría la misma solución. En este contexto, una perspectiva interesante sobre las SVM es considerar sus propiedades de compresión de información. Los vectores de soporte representan los puntos de datos más informativos y comprimen la información contenida en el conjunto de entrenamiento: para el propósito de, por ejemplo, clasificación, solo se necesitan almacenar los vectores de soporte, mientras que todos los demás ejemplos de entrenamiento pueden descartarse. Esto, junto con algunas propiedades geométricas de las SVM como la interpretación de la norma RKHS de su solución como el inverso del margen (Vapnik 1999), es una propiedad clave de las SVM y podría explicar por qué esta técnica funciona bien en muchas aplicaciones prácticas.

Capítulo 5

Selección del modelo

Este Capítulo se centra en la selección del modelo en el aprendizaje automático, definiendo un algoritmo como un proceso de inferencia basado en un modelo y enfatizando la importancia de la consistencia. Se presenta un Teorema que proporciona una cota para el riesgo esperado del estimador de mínimos cuadrados regularizados. Se discute la estimación del parámetro óptimo λ abordando sus propiedades asintóticas y la noción de consistencia en la selección del modelo. Además, se mencionan métodos de implementación para la selección de λ, como la validación cruzada, y la determinación del parámetro C.

5.1. Consistencia de un algoritmo

Podemos definir a un algoritmo (De Vito *et al.* 2005), en el contexto del aprendizaje automático, como el proceso de inferencia a partir de un conjunto finitos de datos, basados en un modelo representado por el espacio de hipótesis.

Si el proceso de inferencia es correcto y el modelo realista, a medida que el tengamos mas datos disponibles, esperamos que la solución se aproxime a la mas precisa posible. Esta propiedad es usualmente llamada consistencia. Puede definirse consistencia de la siguiente forma.

Sea $(\mathcal{X}, \mathcal{A}, P)$ un espacio de probabilidad, donde:

- $\mathcal{X}$ es el espacio de entrada.
- $\mathcal{A}$ es una σ-álgebra sobre $\mathcal{X}$.
- P es una medida de probabilidad.

Consideremos una secuencia de algoritmos $\{g_n\}_{n=1}^{\infty}$, donde cada $g_n : \mathcal{X} \to \mathcal{Y}$ es una función de predicción.

Definición de Consistencia(Boucheron *et al.* 2004b): Un algoritmo es consistente si para cualquier medida de probabilidad P, se cumple que:

$$\lim_{n\to\infty} R(g_n) = R^* \quad \text{casi seguramente,}$$

donde: $R(g_n)$ es el riesgo empírico del algoritmo g_n, definido como:

$$R(g_n) = \mathbb{E}_{z\sim P} L(g_n(z), y_z),$$

siendo L una función de pérdida y y_z la etiqueta verdadera asociada a la muestra z. R^* es el riesgo Bayesiano, definido como:

$$R^* = \inf_g \mathbb{E}_{z\sim P} L(g(z), y_z),$$

donde el ínfimo se toma sobre todas las posibles funciones de predicción g.

Un algoritmo es consistente si a medida que se le proporcionan más datos su rendimiento se acerca al mejor rendimiento posible (el riesgo Bayesiano).

Un problema central de la Teoria del aprendizaje es el evaluación cuantitativa de la propiedad de inferencia de un algoritmo de aprendizaje.

Una serie de trabajos seminales citados en (De Vito *et al.* 2005) tales como (Alon *et al.* 1997),(Cucker & Smale 2001), (Devroye *et al.* 1996) o (Vapnik 1998) muestran que la característica fundamental de un algoritmo debe ser la capacidad de controlar la complejidad de la solución.

Simplificando, si el modelo es demasiado complejo, la solución a encontrar sobreajusta los datos.

En aras de superar el sobre ajuste se pueden utilizar diversas medidas de complejidad, tales como la dimension VC o los covering numbers.

Es de destacar que el buen desempeño de una amplia clase de algoritmos ha sido explicada en términos de estabilidad con respecto a las variaciones del conjunto de entrenamiento.

Se hace necesario entonces introducir una familia paramétrica de algoritmos de aprendizaje en cual los parámetros controlan las propiedades de generalización. Uno de los tipos de algoritmo son los regularizados.

En este contexto el problema principal es la elección óptima del parámetro como función del tamaño de la muestra.

Abordaremos entonces el problema de minimizar el error empírico regularizado:

$$\frac{1}{l}\sum_{i=1}^{l} V(y_i, f(x_i)) + \lambda||f||_K^2.$$

5.2. Estimación de λ

Procederemos a desarrollar los métodos planteados en (De Vito *et al.* 2005).

5.2.1. Teorema I

Dado el parametro λ, el riesgo esperado del estimador f_D^λ provisto por el algoritmo de mínimos cuadrados regularizados se concentra en torno al valor $R[f^\lambda]$. La desviación puede ser acotada por una función S que depende únicamente del nivel de confianza, el tamaño de la muestra y dos constantes κ y δ, que encodean algunas propiedades topológicas de X, Y y el kernel.

El enunciado formal del Teorema (De Vito *et al.* 2005) se transcribe a continuación.

Dados $0 < \eta < 1$, $l \in \mathbb{N}$ y $\lambda > 0$, con probabilidad al menos $1 - \eta$,

$$|R[f_D^\lambda] - R[f^\lambda]| \leq S(\lambda, l, \eta),$$

donde

$$S(\lambda, l, \eta) = \frac{\delta\kappa^2}{\lambda\sqrt{l}}\left(1 + \frac{\kappa}{\sqrt{\lambda}}\right)\left(1 + \sqrt{2log\frac{2}{\eta}}\right).$$

El término $R[f^\lambda]$ es independiente de los datos puede ser interpretado como el precio de reemplazar la función regresión f_0 por la solución regularizada f^λ, es decir el error de aproximación.

El término S es una cota de $|R[f_D^\lambda] - R[f^\lambda]|$ es decir, sobre el error empírico o muestral que resulta de aproximar la solución regularizada a través de un conjunto de entranamiento finito D.

Dado que el Teorema acota R por encima y por debajo, y que S tiende a 0 para l tendiendo a $+\infty$, el riego esperado de f_D^λ se concentra en torno al riesgo esperado de f^λ. Luego, el haber separado $R[f_d^\lambda]$ entre el error de aproximación y el error muestral resulta natural e intrínseco al problema planteado.
En resumen, este Teorema es de gran utilidad por lo siguiente:

- Control del Error de Generalización: Proporciona una cota superior para el error de generalización del estimador $f_{D,\lambda}$. Esto es crucial para entender

cómo de bien el modelo aprendido se generaliza a datos no vistos.

- Selección de Parámetros: La cota $S(\lambda, \ell, \alpha)$ puede ser utilizada para seleccionar el parámetro de regularización λ de manera óptima. Al minimizar esta cota, se puede encontrar un valor de λ que balancee adecuadamente el sesgo y la varianza, reduciendo así el riesgo de sobreajuste o subajuste.
- Estabilidad del Algoritmo: El Teorema se basa en las propiedades de estabilidad del algoritmo de mínimos cuadrados regularizados. Esto significa que el algoritmo es robusto frente a pequeñas variaciones en el conjunto de entrenamiento, lo cual es una propiedad deseable en aplicaciones prácticas.
- Asintótica y Consistencia: La cota probabilística ayuda a demostrar que el algoritmo es consistente, es decir, que el riesgo esperado del estimador converge al menor riesgo posible a medida que el número de ejemplos aumenta.
- Diseño de Algoritmos: Los diseñadores de algoritmos pueden utilizar este Teorema para desarrollar métodos de aprendizaje que sean teóricamente sólidos y que ofrezcan garantías de rendimiento en términos de error de generalización.

5.2.2. Estimación del parámetro óptimo

A partir del Teorema I podemos derivar la siguiente cota:

$$R[f^{\lambda}] \leq R[f^{*}] + S(\lambda, \ell, \eta),$$

la cual se cumple con una probabilidad de al menos $1 - \eta$.

De la forma explícita de $S(\lambda, \ell, \eta)$, tenemos que $S(\lambda, \ell, \eta)$ decrece con λ y tiende a $+\infty$ cuando λ tiende a 0. Por otro lado, es fácil verificar que $R[\hat{f}_{\lambda}]$ es una función creciente de λ y tiende a 0 cuando λ tiende a 0, ver (Cucker & Smale 2002).

Nuestra definición asegura la existencia y unicidad del estimador λ_0 del parámetro óptimo, sin embargo, aún tenemos que probar que λ_0 es finito. Ahora probaremos que la cota proporciona un estimador λ_0 que es finito para ℓ suficientemente grande.

La probabilidad de que los datos $D \notin Z$ dado que el valor de $I[f_b]$ es mayor que $E(\lambda, \ell, \eta)$ es menor o igual a η.

Para el peor escenario, la expresión anterior conduce a la siguiente regla de selección de modelo:

$$\lambda_0(\ell, \eta) = \underset{\lambda \geq 0}{\operatorname{argmin}} E(\lambda, \ell, \eta).$$

Para hacer la definición anterior más rigurosa, asumimos que E se extiende a una función continua de λ en el intervalo $[0, +\infty)$ y reemplazamos por:

$$\lambda_0(\ell, \eta) = \underset{\lambda \in [0, +\infty)}{\text{máx}} \underset{\lambda}{\operatorname{argmin}} E(\lambda, \ell, \eta).$$

5.2.3. Propiedades asintóticas y consistencia del estimador de λ

Continuando con el desarrollo de (De Vito *et al.* 2005) planteamos algunas propiedades asintóticas al agrandar el tamaño de la muestra del algoritmo de KRLS que selecciona los parámetros con los métodos antes descritos. Específicamente, describiremos las propiedades del parámetro $\lambda_0 = \lambda_o(l, n)$ de acuerdo a la noción de consistencia que se establece a continuación.

Definición. Se dice que la familia de parámetro único de estimadores $f^{\lambda(l)}$ provista de una regla de selección de modelo $\lambda()_0$ es consistente si, para todo $\epsilon > 0$, se cumple

$$\lim_{l \to \infty} \sup_{P} \mathbb{P} \left\{ D \geq l \cdot I[f^{\lambda(l)}] > \inf_{f \in H} I[f] + \epsilon) \right\} = 0,$$

donde el supremo se toma sobre el conjunto de todas las medidas de probabilidad en $X \times Y$.

En la definición anterior, el número $\text{ínf}_{f \in \mathcal{H}} I[f]$ representa una especie de error de sesgo, asociado con la elección de $\mathcal{H}$ y, por lo tanto, no puede ser controlado por el parámetro λ.

En particular, si existe $f_{\mathcal{H}} \in \mathcal{H}$ tal que $I[f_{\mathcal{H}}] = \text{ínf}_{f \in \mathcal{H}} I[f]$, el estimador $f_{\mathcal{H}}$ es la mejor descripción determinista posible que podemos dar de la relación entre x e y, dado $\mathcal{H}$. Por claridad, notamos que, para algoritmos de minimización de riesgo empírico, el error de sesgo se suele llamar error de aproximación y está controlado por la elección del espacio de hipótesis, ver (Cucker & Smale 2001) y (Niyogi & Girosi 1999).

5.2.4. Implementación de la selección de λ

En el esquema mencionado se utilizan para elegir λ métodos como la validación cruzada (Wahba 1990), la validación cruzada generalizada, el error de predicción finito o el criterio MDL. Se puede consultar (Vapnik 1998) para una revisión y comparación de los mismos.

5.3. Determinación de c

Siguiendo las pautas de (Rifkin & Lippert 2007) procedemos a minimizar la expresión:

$$\frac{1}{l}\sum_{i=1}^{l} V(y_i, f(x_i)) + \lambda||f||_K^2,$$

que simplificamos introduciendo la variable l en λ:

$$1/2\sum_{i=1}^{l}(y_i - f(x_i))^2 + \lambda/2\|f\|^2.$$

El Teorema de representación nos garantiza que la solución puede ser escrita como: $\mathrm{f}(\cdot) = \sum_{i=1}^{n} c_i k(X_i, \cdot)$, para algún $c \in \mathbb{R}^n$.
Podemos entonces reescribir problema de minimización de la siguiente forma:

$$\min_{c\in\mathbb{R}^n} \frac{1}{2}\|Y - Kc\|_2^2 + \frac{\lambda}{2}\|f\|_2^2.$$

Consideremos una función:

$$f(\cdot) = \sum_{i=1}^{n} c_i k(X_i, \cdot).$$

Para dicha función, la norma al cuadrado de f en el espacio de Hilbert inducido

por el kernel k es igual a:

$$\begin{aligned}\|f\|_k^2 &= \langle f, f\rangle_k \\ &= \left\langle \sum_{i=1}^{n} c_i k(X_i, \cdot), \sum_{j=1}^{n} c_j k(X_j, \cdot) \right\rangle_k \\ &= \sum_{i=1}^{n}\sum_{j=1}^{n} c_i c_j \langle k(X_i, \cdot), k(X_j, \cdot)\rangle_k \\ &= \sum_{i=1}^{n}\sum_{j=1}^{n} c_i c_j k(X_i, X_j) \\ &= c^T K c.\end{aligned}$$

Llegamos entonces a que la expresión a minimizar es la siguiente:

$$1/2\|Y - Kc\|_2^2 + \frac{\lambda}{2} c^T K c.$$

Esta función es convexa en la variable c y por lo tanto podemos encontrar su mínimo igualando su gradiente a 0:

$$\begin{aligned}-K\left(Y - Kc\right) + \lambda K c &= 0 \\ \left(K + \lambda I\right) c &= Y \\ c &= \left(K + \lambda I\right)^{-1} Y.\end{aligned}$$

Podemos encontrar entonces c resolviendo un sistema lineal de ecuaciones:

$$\left(K + \lambda I\right) c = Y.$$

Para cada $\lambda > 0$ la solución existe y es única pues la matriz $(K + \lambda I)$ es simétrica y definida positiva, dado que K lo era.

En este punto, nos vemos tentados a implementar alegremente un algoritmo que invirtiera la matriz $(K+\lambda I)$ para calcular un c para cada λ que probáramos para luego realizar cálculos de $LOOE$ [1]

[1]Desarrollamos el método de *Leave-One-Out Cross-Validation* o LOOCV más adelante en este trabajo.

A continuación desarrollamos resultados adicionales que facilitaran en forma crítica los cálculos necesarios.

5.4. Obtención de c para distintos λ

Realizando la descomposicion en vectores y valores propios de la matriz K:

$$K = Q\Lambda Q^t \text{ con } QQ^t = I \text{ y } \Lambda \text{ es diagonal con } \Lambda_{ij} \geq 0.$$

Luego:

$$\begin{aligned}(K + \lambda I) &= \left(Q\Lambda Q^t + \lambda I\right) = \left(\Lambda Q^t + Q\lambda I Q^t\right) \\ &= Q\left(\Lambda + \lambda I\right)Q^t.\end{aligned}$$

Por lo cual:

$$G^{-1} = (K + \lambda I)^{-1} = Q\left(\Lambda + \lambda I\right)^{-1} Q^t.$$

Una vez calculados Q y Λ, podemos encontrar el c correspondiente a ese λ:

$$c\left(\lambda\right) = Q\left(\Lambda + \lambda I\right)^{-1} Q^t Y.$$

Este último cálculo es más sencillo, dado que la matriz $(\Lambda + \lambda I)$ es diagonal.

5.5. Validación

Sabemos cómo calcular el c para cada λ, debemos utilizar un algoritmo para su selección óptima.

Validar implica verificar el comportamiento de la función en otros puntos distintos del set de entrenamiento.

5.5.1. Tipos de validación

Si tenemos una gran cantidad de datos, podemos apartar algun porcentaje de los datos, y usar ese conjunto como datos de prueba o test para seleccionar los hiperparámetros.

Una opción es la validación cruzada k-veces (k-fold cross-validation), que consiste en dividir los datos en k conjuntos iguales. Para cada i desde 1 a k entrenamos los otros $k-1$ conjuntos y testeamos en el $i$ésimo conjunto.

Cuando contamos con pocos datos, el límite de esta validación cruzada consiste en separar muestra de entrenamiento en n conjuntos de 1 elemento, es la llamada leave-one-out cross-validation.

5.5.2. Validación cruzada dejando uno fuera[2]

Para cada punto x_i se obtiene una función usando los $n-1$ datos restantes, y se mide el error en x_i.

Problema: Debemos construir n predictores diferentes, en conjuntos de datos de tamaño $n-1$.

Es posible acelerar este proceso, dado que se puede obtener una expresión analítica del error de validación cruzada (Rifkin & Lippert 2007).

$$LOOE = \frac{c}{diag_v(G^{-1})},$$

Con $G(\lambda) = K + \lambda I$.

Demostración. Definimos S^i como el conjunto de datos al que se le quitó el i-ésimo punto, es decir:

$$S^i = \{(x_1, y_1), \ldots, (x_{i-1}, y_{i-1}), (x_{i+1}, y_{i+1}), \ldots, (x_n, y_n)\}.$$

El valor i-ésimo leave-one-out es $f_{S^i}(x_i)$, es decir el valor de la función que resulta de entrenar sobre S^i evaluando en x_i. Definimos $LOOV$ como el vector formado por los valores calculados para todo el conjunto de entrenamiento.

El error i-ésimo leave-one-out es $y_i - f_{S^i}(x_i)$. Definimos $LOOE$ como el vector formado por los errores calculados para todo el conjunto de entrenamiento.

$\|LOOE\|_2^2$ es considerado un buen proxy empírico para el error en puntos futuros, entonces seleccionaremos los parámetros que minimicen esta cantidad.

5.5.3. Validación cruzada dejando uno fuera para RLS

Supongamos que ya conocieramos la función f_{S^i}.

Definimos el vector Y^ide la siguiente forma:

$$y_j^i = \begin{cases} y_j & j \neq i \\ f_{S^i}(x_i) & j = i \end{cases}$$

[2]En inglés: *Leave-One-Out Cross-Validation.*

Procedamos a resolver el problema de minimización utilizando el vector Y^ien lugar de Y:

$$\begin{aligned}\sum_{j=1}^{n}\left(f\left(x_i\right)-y_j^i\right)^2+\lambda\left\|f\right\|_K^2 &\geq \sum_{j\neq i}\left(f\left(x_i\right)-y_j^i\right)^2+\lambda\left\|f\right\|_K^2\\ &\geq \sum_{j\neq i}\left(f_{S^i}\left(x_i\right)-y_j^i\right)^2+\lambda\left\|f_{S^i}\right\|_K^2\\ &= \sum_{i=1}^{n}\left(f_{S^i}\left(x_i\right)-y_j^i\right)^2+\lambda\left\|f_{S^i}\right\|_K^2.\end{aligned}$$

Obtenemos entonces:

$$f_{S^i}(x_i)=(KG^{-1}Y^i)_i.$$

La expresión anterior es teórica, dado que para obtener Y^i necesitamos conocer $f_{S^i}(x_i)$. Siguiendo esta técnica, asumiento que tenemos el problema resuelto, y por lo tanto calculado $f_s(x)=KG^{-1}Y$, es posible obtener la siguiente expresión:

$$f_{S^i}(x_i)-f_S(x_i)=\sum_j(KG^{-1})_{ij}(y_j^i-y_j)=(KG^{-1})_{ii}(f_{S^i}(x_i)-y_i).$$

Entonces:

$$\begin{aligned}f_{S^i}(x_i) &= \frac{f_S(x_i)-(KG^{-1})_{ii}y_i}{1-(KG^{-1})_{ii}}\\ &= \frac{(KG^{-1}Y)_i-(KG^{-1})_{ii}y_i}{1-(KG^{-1})_{ii}}.\end{aligned}$$

Luego:

$$LOOV=\frac{KG^{-1}Y-\mathrm{diag}_m(KG^{-1})Y}{I-KG^{-1}}.$$

$$\begin{aligned} LOOE &= Y - LOOV = Y + \frac{\mathrm{diag}_m(KG^{-1})Y - KG^{-1}Y}{\mathrm{diag}_v(I - KG^{-1})} \\ &= \frac{\mathrm{diag}_m(I - KG^{-1})Y}{\mathrm{diag}_v(I - KG^{-1})} + \frac{\mathrm{diag}_m(KG^{-1})Y - KG^{-1}Y}{\mathrm{diag}_v(I - KG^{-1})} \\ &= \frac{Y - KG^{-1}Y}{\mathrm{diag}_v(I - KG^{-1})}. \end{aligned}$$

Podemos simplificar la expresión de $LOOE$ utilizando:

$$\begin{aligned} KG^{-1} &= Q\Lambda Q^t Q(\Lambda + \lambda I)^{-1} Q^t \\ &= Q\Lambda(\Lambda + \lambda I)^{-1} Q^t \\ &= Q(\Lambda + \lambda I - \lambda I)(\Lambda + \lambda I)^{-1} Q^t \\ &= I - \lambda G^{-1}. \end{aligned}$$

Susituyendo el resultado anterior en la expresión de $LOOE$:

$$\begin{aligned} LOOE &= \frac{Y - KG^{-1}Y}{\mathrm{diag}_v(I - KG^{-1})} \\ &= \frac{Y - (I - \lambda G^{-1})Y}{\mathrm{diag}_v(I - (I - \lambda G^{-1}))} \\ &= \frac{\lambda G^{-1} Y}{\mathrm{diag}_v(\lambda G^{-1})} \\ &= \frac{G^{-1} Y}{\mathrm{diag}_v(G^{-1})} \\ &= \frac{c}{\mathrm{diag}_v(G^{-1})}. \end{aligned}$$

Selección del parámetro del Kernel

Se deberá encontrar el óptimo del parámetro λ conjuntamente con la determinación de los parámetros de la definición de la función Kernel. Por ejemplo, para el Kernel Gaussiano es necesario determinar σ:

$$K(u, v) = e^{\|u-v\|^2/-2\sigma^2}.$$

Habiendo obtenido una forma eficiente de calcular la solución para distintos λ podremos recorrer una grilla de valores de σ y para cada caso obtener el λ óptimo.

Capítulo 6

Máquinas de Vectores de Soporte para Clasificación y Regresión

Este trabajo se hace foco en las Máquinas de Vectores de Soporte (SVM) para clasificación y regresión, un tema central en el Aprendizaje Automático. Se introduce el concepto de hiperplanos separadores, fundamentales para entender cómo las SVM clasifican datos dividiéndolos en categorías.

Este capítulo explorará:

- La idea de encontrar el hiperplano óptimo, aquel que maximiza el margen entre las clases. Se explicará también por qué este hiperplano es único.
- Una introducción general a las máquinas de vectores de soporte, destacando su importancia en el diseño de sistemas de aprendizaje.
- Cómo las SVM se utilizan para resolver problemas de clasificación, incluyendo clasificadores binarios y regularizados. Se abordará el caso separable y el caso no separable, donde se introduce el concepto de clasificador SVM de margen suave.
- El concepto del “truco del kernel", una técnica que permite realizar cálculos en un espacio de características de alta dimensión sin necesidad de calcular explícitamente la transformación de los datos.
- La aplicación de SVM en problemas de regresión, tanto para estimar funciones lineales como no lineales.

- Una conexión entre la pérdida hinge y el concepto de un hiperplano óptimo, explicando geométricamente por qué las SVM pueden ser efectivos clasificadores.

Nos basamos en la idea de que las SVM son una herramienta poderosa para el aprendizaje automático, especialmente cuando se trata de datos complejos que no son linealmente separables en su espacio original. Este enfoque permite abordar problemas de clasificación y regresión de manera efectiva, proporcionando una alternativa robusta a otros métodos. Además, este capítulo presenta las bases para la comprensión de los algoritmos SVM utilizados en la parte práctica de este trabajo.

Este capítulo utiliza como referencia trabajos clásicos de Vapnik y otros investigadores para proporcionar una base teórica sólida para poder comprender y optimizar el uso de las SVM.

6.1. Hiperplanos separadores

En (Vapnik 1998) se desarrolla el concepto de hiperplanos separadores.

Decimos que dos subconjuntos finitos de vectores $\mathbf{x}$ del conjunto de entrenamiento

$$(y_1, \mathbf{x}_1), ..., (y_l, \mathbf{x}_l), \quad \mathbf{x} \in \mathbb{R}^n, \quad y \in \{-1, 1\},$$

un subconjunto I para el cual $y = 1$, y otro subconjunto II para el cual $y = -1$, son separables por el hiperplano $\langle \mathbf{x}, \phi \rangle = c$ si existen tanto un vector unitario ϕ, con $\|\phi\| = 1$, como una constante c tales que las desigualdades:

$$\langle \mathbf{x}_i, \phi \rangle > c, \quad \text{si } \mathbf{x}_i \in I, \langle \mathbf{x}_j, \phi \rangle < c, \quad \text{si } \mathbf{x}_j \in II,$$

se cumplen donde $\langle a, b \rangle$ denota el producto interno entre los vectores a y b.

Sea ϕ un vector unitario cualquiera. Definimos dos valores:

$$c_1(\phi) = \min_{x_i \in I} \langle x_i, \phi \rangle,$$

$$c_2(\phi) = \max_{x_j \in II} \langle x_j, \phi \rangle.$$

Consideremos el vector unitario ϕ_0 que maximiza la función

$$\rho(\phi) = \frac{c_1(\phi) - c_2(\phi)}{2}, \quad ||\phi|| = 1,$$

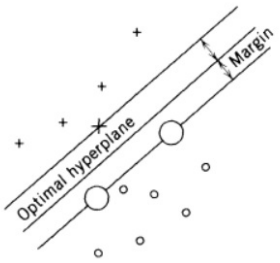

Figura 6.1: Ejemplo de hiperplano separador óptimo para dimensión 2 y el margen. Tomado de (Vapnik 1998).

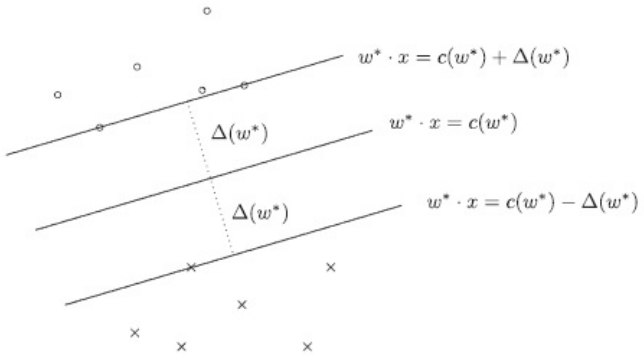

Figura 6.2: Otro ejemplo de hiperplano separador óptimo para dimensión 2 y el margen. Tomado de (Cucker & Zhou 2007).

bajo la condición de que las desigualdades antes mencionadas se cumplan. El vector ϕ_0 y la constante

$$c_0 = \frac{c_1(\phi_0) + c_2(\phi_0)}{2},$$

determinan el hiperplano que separa los vectores $x_1, ..., x_a$ del subconjunto I de los vectores $x_1, ..., x_b$ del subconjunto II, $(a + b = l)$, y tiene el margen máximo.

Llamamos a este hiperplano el hiperplano de margen máximo o hiperplano óptimo y está definido por la ecuación:

$$\langle x, \phi_0 \rangle = c_0.$$

6.2. Unicidad del hiperplano óptimo

Hacemos referencia al Teorema (Vapnik 1998) donde se demuestra la unicidad del hiperplano óptimo.

Se demuestra que el punto máximo ϕ_0 de la función continua $\rho(\phi)$, definida en el área $||\phi|| \leq 1$, existe y se alcanza en el límite $||\phi|| = 1$. La existencia del máximo se infiere la continuidad de $\rho(\phi)$ en la región acotada $||\phi|| < 1$.

Supongamos que el máximo se alcanza en algún punto interior ϕ^*. Entonces el vector

$$\phi^* = \frac{\phi_0}{||\phi_0||},$$

definiría un margen mayor

$$\rho(\phi^*) = \frac{\rho(\phi_0)}{||\phi_0||}.$$

El máximo de la función $\rho(\phi)$ no puede alcanzarse en dos puntos pues, dado que la función $\rho(\phi)$ es convexa, se alcanzaría en la línea que conecta estos dos puntos, es decir, en un punto interior, lo cual es imposible según los argumentos anteriores. Esto prueba el Teorema.

El objetivo es entonces encontrar métodos efectivos para construir el hiperplano óptimo. Para ello consideramos un enunciado equivalente del problema: encontrar un par que consiste en un vector ψ_0 y una constante b_0 conocida como umbral, tales que satisfagan las siguientes restricciones:

$$\begin{aligned} \langle x_i, \psi_0 \rangle + b_0 &\geq 1, \quad \text{si } y_i = 1, \\ \langle x_j, \psi_0 \rangle + b_0 &\leq -1, \quad \text{si } y_j = -1, \end{aligned}$$

y el vector ψ_0 tiene la norma más pequeña

$$||\psi_0||^2 = \langle \psi_0, \psi_0 \rangle.$$

6.3. Introducción a las máquinas de vectores de soporte

Según (Vapnik 1998) podemos decir que los métodos de hiperplanos separadores cumplen un rol crucial en el diseño de aprendizaje. Consideramos ahora

un tipo especial de hiperplanos separadores: los llamados hiperplanos óptimos que poseen algunas propiedades estadísticas notables. Usando el método del hiperplano separador óptimo, Vapnik introduce una nueva clase de máquinas de aprendizaje para estimar funciones indicadoras, las llamadas máquinas de vectores de soporte (SVM[1]) que son pasibles de ser generalizadas para estimar funciones de valor real, procesamiento de señales y resolución de ecuaciones de operadores lineales.

6.4. Problemas de Clasificación

Hasta ahora hemos hecho foco en el problema de aprender una función en los números reales, nos planteamos ahora problemas que consisten en aprender una función que resulte en valores binarios, o un conjunto finito, llamados problemas de clasificación.

En la práctica aparecen con frecuencia problemas tales como determinar en base a una muestra de datos clínicos si un paciente sufre de cierta enfermedad o no.

Para un clasificador binario en un espacio métrico compacto es una función $f : X \to Y$ con $Y = \{1, -1\}$, manteniendo $Z = X \times Y$.

Para medir la calidad de nuestras aproximaciones definiremos un error acorde.

Sea ρ la distribución de probabilidad en $Z = X \times Y$. El error de mal clasificaciońn $\mathcal{R}(f)$ para el clasificador $f : X \to Y$ es definido como la probabilidad de una predicción errónea, es decir, la medida del evento $\{f(x) \neq y\}$,

$$\mathcal{R}(f) := \text{Prob}_{z \in Z}\{f(x) \neq y\} = \int_X \text{Prob}_{y \in Y}\{y \neq f(x)|x\} d\rho_X.$$

Este clasificador es conocido como la regla de Bayes (Duda *et al.* 2000).

Procedemos a describir una aproximación para producir clasificadores de las muestras (y una función kernel o núcleo $\mathcal{H}_K$ cuando aplique la regularización) conocido como Support Vector Machines.

Buscaremos acotar el error de malclasificación de los clasificadores obtenidos.

Las SVM producen clasificadores de una muestra $z \in Z^m$, un numero real $\lambda > 0$ y una función kernel $\mathcal{H}_K$. Denotemos como $F_{z,\lambda}$ a ese clasificador.

[1] En inglés *Support Vector Machines*.

6.5. Clasificadores Binarios

Como vimos que para el caso de la regresión en que la función f_ρ era la función que minimizaba el error, ahora buscaremos qué clasificador binario minimiza $\mathcal{R}$ (Cucker & Zhou 2007) . Para una función $f : X \to \mathbb{R}$ definimos

$$\text{signo}\,(f(x)) = \begin{cases} -1 & si f(x) \geq 0 \\ 1 & si f(x) < 0. \end{cases}$$

Sea $K_\rho := \{x \in X : f_\rho(x) = 0\}$ y $\kappa_\rho = \rho_X(K_\rho)$.

Proposición: Para cada clasificador f,

$$\mathcal{R}(f) = \frac{1}{2}\kappa_\rho + \int_{X \setminus K_\rho} \text{Prob}_Y\,(y \neq f(x)|x)\, d\rho_X.$$

Entonces $\mathcal{R}$ es minimizado por cualquier clasificador que coincida en $X \setminus K_\rho$ con

$$f_c := \text{signo}(f_\rho).$$

El clasificador f_e se denomina regla de Bayes (Cucker & Zhou 2007).

La función que cumple la cantidad κ_ρ es similar a la de σ_ρ^2 en el contexto de la regresión. κ_ρ depende únicamente de ρ, siendo independiente de f proporcionando una cota inferior para el error de clasificación y es, nuevamente, una medida de qué tan bien condicionado está ρ.

Dado que la distribución ρ es desconocida, el mejor clasificador f_c no puede encontrarse directamente. El objetivo de los algoritmos de clasificación es encontrar clasificadores que aproximen la regla de Bayes f_c a partir de muestras $z \in Z^m$.

Una posible estrategia para lograr este objetivo es la regularización. Se podría elegir un RKHS H_k y un $\gamma > 0$, encontrar el minimizador $f_{z,\gamma}$ del error empírico regularizado, es decir,

$$f_{z,\gamma} = \arg \min_{f \in H_K} \frac{1}{m} \sum_{i=1}^{m} (f(x_i) - y_i)^2 + \gamma ||f||_K^2,$$

y luego tomar la función signo$(f_{z,\gamma})$ como una aproximación de f_c. Esta estrategia minimiza un funcional sobre un conjunto de funciones continuas de valor real y luego aplica la función signo al minimizador calculado para obtener un clasificador.

6.6. Clasificadores Regularizados

Seguimos ahora con las definiciones y resultados proporcionados por (Cucker & Zhou 2007).

Una función de pérdida $\phi(y, f(x))$ es una función que mide el error local (con respecto a ϕ) para un par dado (x, y) en el conjunto de datos Z y una función $f : X \to \mathbb{R}$.

El error de generalización asociado con la función de pérdida ϕ se define como:

$$\mathcal{E}^{\phi}(f) := \int_Z \phi(y, f(x))\, dp.$$

El error empírico asociado con la función de pérdida ϕ y una muestra $z \in Z^m$ se define como:

$$\hat{\mathcal{E}}^{\phi}(f) := \frac{1}{m} \sum_{i=1}^{m} \phi(y_i, f(x_i)).$$

Si $f \in \mathcal{H}_K$ para algún núcleo de Mercer K, entonces podemos definir versiones regularizadas de estos errores. Para $\gamma > 0$, definimos el error regularizado como:

$$\mathcal{E}^{\phi}_{\gamma}(f) := \int_Z \phi(y, f(x))\, dp + \gamma ||f||^2_K,$$

y el error empírico regularizado como:

$$\hat{\mathcal{E}}^{\phi}_{\gamma}(f) := \frac{1}{m} \sum_{i=1}^{m} \phi(y_i, f(x_i)) + \gamma ||f||^2_K.$$

Ejemplos de funciones de pérdida son la pérdida de clasificación errónea:

$$\phi_0(t) = \begin{cases} 0 & \text{si } t \geq 0 \\ 1 & \text{si } t < 0. \end{cases}$$

y la pérdida de mínimos cuadrados $\phi_{ls}(t) = (1 - t)^2$.

Se puede demostrar (Cucker & Zhou 2007) en el contexto de los los clasificadores regularizados asociados con funciones de pérdida generales que la pérdida de mínimos cuadrados ϕ_{ls} proporciona un algoritmo satisfactorio desde el punto de vista de las tasas de convergencia en su análisis de error. Aquí restringiremos nuestra exposición a una pérdida especial, llamada pérdida *hinge*[2].

[2] *Hinge loss* del inglés

La pérdida *hinge* se define como:

$$\phi_h(t) = (1-t)_+ = \text{máx}\{1-t, 0\}.$$

El clasificador regularizado asociado esta funci[on de pérdida, la máquina de vectores de soporte, ha sido utilizado extensamente y parece tener un pequeño error de clasificación en la práctica. Una propiedad interesante de la pérdida hinge ϕ_h es la eliminación del error local cuando $yf(x) > 1$. Esto significa que la solución $f_{z,\gamma}^{\phi_h}$ es dispersa en la representación $f_{z,\gamma}^{\phi_h} = \sum_{i=1}^{m} c_{z,i} K_{x_i}$. Es decir, la mayoría de los coeficientes $c_{z,i}$ en esta representación son 0. Por lo tanto, el cálculo de $f_{z,\gamma}^{\phi_h}$ puede ser, en la práctica, muy rápido. Volveremos a este tema más adelante.

Aunque la definición de la pérdida hinge puede no sugerir a primera vista ninguna razón particular para obtener buenos clasificadores, resulta que existe cierta geometría que explica por qué puede hacerse.

6.7. Hiperplanos óptimos: el caso separable

Sea $X \subseteq \mathbb{R}^n$ y $z = (z_1, ..., z_m)$ un conjunto de datos de muestra con $z_i = (x_i, y_i), i = 1, ..., m$. z consiste de dos clases con los siguientes conjuntos de indices: $I = \{i | y_i = 1\}$ y $II = \{i | y_i = -1\}$.

Sea H el hiperplano definido por $\langle w, x\rangle = b$ con $w \in \mathbb{R}^n$, $||w|| = 1$, y $b \in \mathbb{R}$.

Decimos que I y II son separables por H cuando, para $i = 1, ..., m$:

$$\begin{cases} \langle w, x\rangle > b & \text{si } i \in I \\ \langle w, x\rangle < b & \text{si } i \in II. \end{cases}$$

Es decir que los puntos x_i correspondientes a I y II caen en distintos lados de H.

Cualquier hiperplano en $\mathbb{R}^n$ induce un clasificador. Si su ecuación es $\langle w, x\rangle - b = 0$, entonces la función $x \mapsto \text{sgn}(\langle w, x\rangle - b)$ es un clasificador. Este razonamiento sugiere que el mejor clasificador entre aquellos inducidos de esta manera podría ser aquel para el cual la dirección w genere un hiperplano separador con el mayor margen posible $\Delta(w)$. Dado un conjunto de datos Z, dicha dirección se obtiene resolviendo el problema de optimización:

$$\max_{||w||=1} \Delta(w),$$

o, en otras palabras,

$$\max_{||w||=} \frac{1}{2}\left[\min_{y_i=1}\langle w, x_i\rangle - \max_{y_i=-1}\langle w, x_i\rangle\right].$$

Si w^* es un maximizador con $\Delta(w^*) > 0$, entonces el hiperplano definido por $\langle w^*, x\rangle = c(w^*)$ óptimo y $\Delta(w^*)$ se denomina margen de la muestra.

6.8. Máquinas de vectores de soporte: el enfoque de (Cucker & Zhou 2007)

Cuando el conjunto de datos Z es separable, podemos obtener un clasificador resolviendo la ecuación que más abajo transcribimos y luego, si w^* es la solución calculada, el clasificador asigna a cada punto x la señal de $\langle w^*, x\rangle - c(w^*)$. Es decir,

$$x \mapsto \text{sgn}\langle w^*, x\rangle - c(w^*)).$$

Alternativamente, podemos resolver una forma equivalente a esta ecuación:

$$\max_{||w||=1} \frac{1}{2}\left[\min_{y_i=1}\langle w, x_i\rangle - \max_{y_i=-1}\langle w, x_i\rangle\right].$$

Desarrollamos entonces la solución alternativa propuesta en (Cucker & Zhou 2007).

Teorema: Supongamos que el problema planteado tiene una solución w^* tal que $\Delta(w^*)\rangle 0$. Entonces, $w^* = \frac{w}{||w||}$, donde w es una solución del siguiente problema de optimización:

$$\begin{aligned} &\min_{w\in\mathbb{R}^n, b\in\mathbb{R}} ||w||^2 \\ &\text{sujeto a} y_i(w^T x_i - b) \geq 1, \quad i = 1, ..., m. \end{aligned}$$

Además, $\Delta(w^*) = \frac{1}{||w||}$ representa el margen.

Por lo tanto, en el caso separable, podemos proceder resolviendo ya sea el problema de optimización para la primer ecuación o el obtenido por este último Teorema. El clasificador resultante se denomina clasificador de margen duro y su margen está dado por $\Delta(w^*)$ con w^* la solución en el primer caso o para la

segunda opción.

Es entonces que hay al menos n vectores de soporte. En la mayoría de las aplicaciones de máquinas de vectores de soporte, el número de vectores de soporte es mucho menor que el tamaño de la muestra m. Esto hace que el segundo algorimo propuesto se pueda resolver en forma más rápida.

Las máquinas de vectores de soporte consisten en una familia de algoritmos de clasificación eficientes: el clasificador de margen duro SVM , que funciona para datos separables, el clasificador de margen suave SVM , para datos no separables, que desarrollaremos a continuación, y el algoritmo SVM general asociado con la pérdida hinge Φ_h y un kernel de Mercer general K. El primer y el segundo clasificador pueden expresarse en términos del kernel lineal $K(x, y) = \langle x, y\rangle + 1$, mientras que el SVM general involucra kernels de Mercer generales: el kernel polinomial $\langle x, y\rangle + 1)^d$ con $d \in \mathbb{N}$ o gaussianos $\exp\{-\frac{||x-y||^2}{2\sigma^2}\}$ con $\sigma > 0$.

Destacamos, una vez más, que estos algoritmos SVM comparten una característica especial causada por la pérdida hinge: la solución $f(x) = \sum_{i=1}^{m} c_i y_i K(x, x_i)$ a menudo tiene un vector de coeficientes $c = (c_1, ..., c_m)$ disperso, lo que hace que el algoritmo que calcula c sea más rápido.

6.9. El caso no separable - clasificador SVM de margen suave

Continuando con el razonamiento de (Cucker & Zhou 2007) observamos que en una situación no separable no existen vectores $w \in \mathbb{R}^n$ y escalares $b \in \mathbb{R}$ tales que los puntos en Z puedan ser separados en dos clases con $y_i = 1$ e $y_i = -1$ por el hiperplano $\langle w, x\rangle = b$. En este caso buscamos el clasificador de margen suave. Este se define introduciendo variables de holgura $\xi = (\xi_1, ..., \xi_m)$. Se debe resolver entonces el siguiente problema de optimización:

$$\begin{aligned} \min_{w\in\mathbb{R}^n, b\in\mathbb{R}, \xi\in\mathbb{R}^m} & ||w||^2 + \frac{1}{\gamma}\sum_{i=1}^{m}\xi_i \\ \text{sujeto a: } & y_i(\langle w, x_i\rangle - b) \geq 1 - \xi_i \\ & \xi_i \geq 0, \quad i = 1, ..., m \end{aligned}$$

Donde $\gamma > 0$ es un parámetro de regularización. Si $(\tilde{w}, \tilde{b}, \tilde{\xi})$ es una solución del problema anterior, entonces su clasificador de margen suave asociado se define como $x \mapsto \text{signo}\left(\langle \tilde{w}, x\rangle - \tilde{b}\right)$.

El problema de margen duro en el caso separable puede verse como un caso especial del problema de margen suave cuando $\frac{1}{\gamma} \to \infty$, en cuyo caso todas las soluciones tienen $\tilde{\xi} = 0$.

6.9.1. Lagrangiano del problema principal

Siendo α_i y r_i tenemos entonces que la función de Lagrange es la siguiente:

$$L(w, b, \xi, \alpha, r) = \frac{1}{2} w^t w + C \sum_i \xi_i - \sum_i \alpha_i [y_i (w^t \phi(x_i) + b) - 1 + \xi_i] - \sum_i r_i \xi_i.$$

Observamos que el valor del Lagrangiano está acotado superiormente por la función de costo original. La función de costo original a minimizar es:

$$1/2||w||^2 + \sum_i C\xi_i.$$

Por lo tanto tenemos que demostrar (Steinwart & Christmann 2008) que:

$$\sum_i \alpha_i [y_i (w^t \phi(x_i) + b) - 1 + \xi_i] + \sum_i r_i \xi_i \geq 0.$$

Entonces, utilizando los resultados que veremos más abajo en el desarrollo del problema dual:

$$\begin{aligned} & \sum_i \alpha_i [y_i (w^t \phi(x_i) + b) - 1 + \xi_i] + \sum_i r_i \xi_i \\ = & \sum_{i,j} \alpha_i \alpha_j y_i y_j \phi^t(x_j) \phi(x_i) + \sum_i \alpha_i + \sum \alpha_i \xi_i + \sum_i r_i \xi_i \\ = & \sum_{i,j} \alpha_i \alpha_j y_i y_j \phi^t(x_j) \phi(x_i) + \sum_i \alpha_i + \sum C\xi_i \end{aligned}$$

.

Luego, se cumple la desigualdad dado que:

$\sum_{i,j} \alpha_i \alpha_j y_i y_j \phi^t(x_j) \phi(x_i) \geq 0$ pues la matriz K es definida positiva por definición

$\sum_i \alpha_i \geq 0$ pues por las condiciones KKT[3] $\alpha_i \geq 0$

$\sum C\xi_i \geq 0$ pues por las condiciones KKT $\xi_i \geq 0$.

6.9.2. Problema dual

Resolver entonces el problema de optimización planteado es equivalente a resolver los siguientes problemas:

- $p^* = min_{w,b,\xi} max_{\alpha \geq 0, r \geq 0} L(w, b, \xi, r)$.
- $d^* = max_{\alpha \geq 0, r \geq 0} min_{w,b,\xi} L(w, b, \xi, r)$.

Para resolver entonces la minimización de d^* igualamos a 0 las derivadas parciales de la función definida anteriormente.

$$\frac{dL}{dw} = \frac{dL}{dw} = \frac{dL}{dw} = 0,$$

que resulta en:

$$w = \sum_i \alpha_i y_i \phi(x_i) \sum_i \alpha_i y_i = 0 \alpha_i = C - r_i.$$

Insertando los resultados anteriormente obtenidos concluimos:

$$\frac{1}{2} w^t w = \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j \phi^t(x_i) \phi(x_j).$$

$$C \sum_i \xi_i - \sum_i r_i \xi_i = \sum_i \xi_i (C - r_i) = \sum_i \xi_i \alpha_i.$$

$$\begin{aligned} -\sum_i \alpha_i [y_i (w^t \phi(x_i) + b) - 1 + \xi_i] &= -\sum_i \alpha_i y_i w^t \phi(x_i) - b \sum_i \alpha_i y_i + \sum_i \alpha_i - \sum \alpha_i \xi_i \\ &= -\sum_i \alpha_i y_i w^t \phi(x_i) + \sum_i \alpha_i - \sum \alpha_i \xi_i \\ &= -\sum_{i,j} \alpha_i \alpha_j y_i y_j \phi^t(x_j) \phi(x_i) + \sum_i \alpha_i - \sum \alpha_i \xi_i. \end{aligned}$$

Agrupando los sumandos obtenemos:

$$L(w, b, \xi, \alpha, r) = -\frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j \phi^t(x_i) \phi(x_j) + \sum_i \alpha_i.$$

Reescritura del problema dual

De los resultados del punto anterior, podemos plantear entonces que el problema dual es el siguiente:

$$max_\alpha L(\alpha) = -\frac{1}{2}\sum_{i,j}\alpha_i\alpha_j y_i y_j \phi^t(x_i)\phi(x_j) + \sum_i \alpha_i,$$

sujeto a:

$$\begin{aligned}\sum_i \alpha_i y_i &= 0\\ 0 &\leq \alpha_i \leq C.\end{aligned}$$

Dado que $K(x_i, x_j) = \phi^t(x_i)\phi(x_j)$ usando la expresión dual, el problema se convierte en la expresión matricial del mismo.

Conveniencia de resolver el problema dual

El problema de obtener soluciones sobreajustadas fue abordado por primera vez para la SVM de margen suave en (Cortes & Vapnik 2009). Para explicar su enfoque recordemos que en el problema de optimización las restricciones $y_i(w, x_i) + b \geq 1$ forzaban a los hiperplanos a no cometer errores en el conjunto de datos de entrenamiento D. El enfoque de la SVM de margen suave es relajar estas restricciones, requiriendo únicamente que (w, b) satisfaga $y_i(w, x_i) + b \geq 1 - \xi_i$ para algunas variables de holgura llamadas $\xi_i \geq 0$. Sin embargo, si estas variables de holgura son demasiado grandes, las restricciones relajadas se satisfacerían trivialmente, y por lo tanto se deben agregar salvaguardas contra tal comportamiento. Una forma de hacerlo es agregar las variables de holgura a la función objetivo.

Combinando estas modificaciones con la idea del mapeo de características, se llega al siguiente problema de optimización cuadrática:

$$\begin{aligned}\text{minimizar} \quad & \frac{1}{2}\langle w, w\rangle + C\sum_{i=1}^{n}\xi_i,\\ \text{sujeto a} \quad & y_i(w, \phi(x_i)) + b \geq 1 - \xi_i, \quad i = 1, \ldots, n,\\ & \xi_i \geq 0, \quad i = 1, \ldots, n,\end{aligned}$$

donde $C > 0$ es un parámetro que se utiliza para balancear el primer término de la función objetivo con el segundo. Nótese que, debido a la forma especial del término suplementario $C\sum_{i=1}^{n}\xi_i$, la función objetivo es aún convexa, o para

ser más precisos, cuadrática, mientras que las restricciones son todas lineales. Continuamos con el análisis publicado en (Steinwart & Christmann 2008).

El problema de optimización anterior tiene la desventaja de que debe resolverse en un espacio de Hilbert H_0, a menudo de alta o incluso infinita dimensión. En la práctica se utiliza el enfoque de Lagrange para calcular el programa dual correspondiente. Por ejemplo, para la función de pérdida hinge, este programa dual viene dado por:

$$\begin{aligned} \text{maximizar} \quad & \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} y_i y_j \alpha_i \alpha_j \langle \phi(x_i), \phi(x_j) \rangle, \\ \text{sujeto a} \quad & \sum_{i=1}^{n} y_i \alpha_i = 0, \end{aligned}$$

donde $\alpha_i \in [0, C]$ para todo i.

Si $(\alpha_1^*, ..., \alpha_n^*)$ denota una solución óptima del problema dual, entonces la solución del problema primal (w^*, b^*) puede ser obtenida de la siguiente manera:

$$\begin{aligned} w^* &= \sum_{i=1}^{n} y_i \alpha_i^* \phi(x_i), \\ b^* &= y_j - \sum_{i=1}^{n} y_i \alpha_i^* \langle \phi(x_i), \phi(x_j) \rangle, \end{aligned}$$

donde j es cualquier índice tal que $0 < \alpha_j^* < C$. La función de decisión de la SVM está dada por:

$$f(x) = \langle w^*, \phi(x) \rangle + b^* = \sum_{i=1}^{n} y_i \alpha_i^* \langle \phi(x_i), \phi(x) \rangle + b^*.$$

Observamos que tanto en el problema dual como en la función de decisión, solo aparecen productos internos de la forma $\langle \phi(x_i), \phi(x_j) \rangle$. Esto sugiere que en lugar de calcular explícitamente el mapeo de características ϕ, podemos trabajar directamente con una función núcleo K:

$$K(x, x') = \langle \phi(x), \phi(x') \rangle.$$

El uso de núcleos cuenta entonces con varias ventajas:

- Evita el cálculo explícito de ϕ: a menudo el espacio de características H es de alta dimensión o incluso infinito. Calcular $\phi(x)$ para cada punto de

datos podría ser computacionalmente costoso.

- Flexibilidad: los núcleos permiten modelar relaciones no lineales entre los datos.
- Eficiencia: existen algoritmos eficientes para resolver el problema dual utilizando núcleos.

Al utilizar núcleos, podemos aplicar la SVM a una amplia variedad de problemas de clasificación, incluso aquellos que no son linealmente separables en el espacio de entrada original.

6.10. El "truco"kernel

Obviamente, utilizar núcleos directamente en lugar de calcular primero los mapeos de características funciona para todos los métodos y algoritmos estadísticos en los que se necesitan productos internos del mapeo de características pero no el mapeo de características en sí. Al utilizar núcleos, podemos construir un algoritmo no lineal a partir de uno lineal sin cambiar el diseño central del algoritmo. Esta observación, conocida como el "truco"del núcleo establecida explícitamente por primera vez por (Schölkopf *et al.* 1998).

Desde entonces, varios algoritmos han sido "kernelizados", como el análisis de componentes principales o el análisis discriminante de Fischer. Se puede pronfundizar sobre el tema en (Schölkopf & Smola 2001) o (Shawe-Taylor & Cristianini 2004b).

Una segunda ventaja del "truco"del núcleo es que el espacio de entrada X ya no necesita ser un subconjunto de $\mathbb{R}^d$ ya que todas las computaciones se realizan en el espacio de características. Existen varios núcleos que están definidos en datos no vectoriales, como texto o secuencias de ADN. Por lo tanto, el "truco"del núcleo de hecho extiende la aplicabilidad de métodos que pueden ser "kernelizados". Referimos a (Schölkopf & Smola 2001), (Joachims 2002),(Schölkopf *et al.* 2003), y (Shawe-Taylor & Cristianini 2004b) para varios ejemplos de enfoques basados en núcleos para datos no vectoriales.

6.11. SVM para regresión

Comenzamos esta sección siguiendo los resultados propuestos en (Vapnik 1998) con ejemplos simples de tareas de estimación de regresión donde las

regresiones se definen por funciones unidimensionales y bidimensionales. Luego, consideramos la estimación de funciones de regresión lineal multidimensionales utilizando el método SVM. Construimos un modelo de regresión lineal que sea favorable para un método de selección de características y comparamos los resultados obtenidos para un método de selección de características hacia adelante con los resultados obtenidos por la máquina de vectores de soporte.

6.11.1. Estimación de funciones de regresión lineales

Describirmos a continuación experimentos con máquinas de vectores de soporte (SVMs) en la estimación de funciones de regresión lineal (Drucker 1997).

Comparamos la máquina de vectores de soporte con dos métodos diferentes para estimar la función de regresión lineal, a saber, el método de mínimos cuadrados ordinarios OLS[4] y el método de selección de características hacia adelante por pasos SFS[5].

Recordemos que el método OLS es un método que estima los coeficientes de una función de regresión lineal minimizando el funcional:

$$R(a) = \sum_{i=1}^{n} (y_i - a * x_i)^2.$$

El método SFS es un método que primero elige una coordenada del vector que proporciona la mejor aproximación de los datos. Luego, fija esta coordenada y agrega una segunda coordenada de tal manera que estas dos definan la mejor aproximación de los datos, y así sucesivamente.

Consideramos el problema de estimación de regresión lineal a partir de los datos:

$$(y_1, x_1), ..., (y_n, x_n).$$

6.11.2. Estimación de funciones de regresión no lineales

En (Vapnik 1998) se resuelven distintos experimentos en forma exitosa. Elige funciones de regresión que fueron utilizadas en muchos estudios de referencia, sugeridas en (Friedman 1991):

[4]Del inglés *Ordinary Least Squares*.
[5]Del inglés *Sequential Forward Selection*.

- Modelo de Friedman 1: se consideró la siguiente función de diez variables:

$$y = 10 * sin(\pi * x^{(1)} * x^{(2)}) + 20 * (x^{(3)} - 0.5)^2 + 10 * x^{(4)} + 5 * x^{(5)} + \epsilon.$$

  Esta función, sin embargo, depende solo de cinco variables. En este modelo, las 10 variables están distribuidas uniformemente en [0, 1] y el ruido es normal con parámetros N(0, 1).

- Modelo de Friedman 2:

$$y = \sqrt{(x^{(1)})^2 + [(x^{(2)} * x^{(3)}) - 1/(x^{(2)} * x^{(3)})]^2},$$

  tiene cuatro variables independientes distribuidas uniformemente en la siguiente región:

$$0 \leq x^{(1)} \leq 100, 40\pi \leq x^{(2)} \leq 560\pi, 0 \leq x^{(3)} \leq 1, 1 \leq x^{(4)} \leq 11.$$

  El ruido se ajusta para una relación señal-ruido (SNR[6]).

- Modelo de Friedman 3: también tiene cuatro variables independientes:

$$y = \tan^{-1}\left[\frac{x^{(2)} * x^{(3)} - 1/(x^{(2)} * x^{(4)})}{x^{(1)}}\right] + \epsilon \quad ,$$

  que están distribuidas uniformemente en la misma región. El ruido se ajustó para una SNR.

Comparamos los resultados de la máquina de regresión de soporte con las técnicas de regresión llamadas bagging (Breiman 1996) y AdaBoost (Freund & Schapire 1997) que funcionan en base a las soluciones dadas por árboles de regresión.

[6]Del inglés *signal-to-noise ratio*.

	Bagging	Boosting	SV
Friedman #1	2.2	1.65	0.67
Friedman #2	11.463	11.684	5.402
Friedman #3	0.0312	0.0218	0.026

Figura 6.3: Comparación de los resultados de los distintos métodos para resolver los problemas de Friedmann tomada de (Vapnik 1998).

La tabla tomada de (Vapnik 1998) muestra los resultados de los experimentos para estimar las funciones de Friedman usando bagging, boosting y métodos SVM. Los experimentos fueron llevados a cabo usando 240 ejemplos de entrenamiento.

Capítulo 7

El algoritmo de mínimos cuadrados regularizados

Hasta ahora, si bien hemos avanzado en plantear las ventajas teóricas de la aplicación de técnicas como KRLS, estamos lejos de ser capaces de implementar en la práctica este tipo de soluciones.

El artículo (Rifkin & Lippert 2007) nos permite dar el salto necesario entre la teoría y la implementación concreta de una solución utilizable.

7.1. Introducción, repaso kernels

Supongamos que tenemos una muestra formada por los puntos:

$$S = \{(x_1, y_1), \ldots, (x_n, y_n)\}.$$

Asumimos una función de kernel semidefinida positiva k, que generaliza la noción de producto punto en un Espacio de Hilbert con Núcleo Reproductor (RKHS). Los núcleos comúnmente utilizados según (Rifkin & Lippert 2007) incluyen:

Lineal:

$$k(X_i, X_j) = X_i X_j.$$

Polinómico:

$$k(X_i, X_j) = (X_i X_j + 1)^d.$$

Gaussiano:

$$k(X_i, X_j) = \exp\left(-\frac{||X_i - X_j||^2}{2\sigma^2}\right).$$

El orden polinomial d o el ancho de banda gaussiano σ deben ser especificados por el usuario. Definimos la matriz de kernel K de tal manera que $K_{ij} = k(X_i, X_j)$. La función de kernel k toma múltiples puntos de datos y produce una matriz de resultados: $k(X, X) = K$, y, dado un punto arbitrario X, $k(X, X+)$ es un vector columna cuya entrada i-ésima es $k(X_i, X+)$.

7.2. RLS: Mínimos cuadrados regularizados

La RLS surge como un problema de minimización de Tikhonov (Evgeniou *et al.* 2000) con una pérdida cuadrática:

$$\min_{f \in \mathcal{H}} \sum_{i=1}^{n} (y_i - f(x_i))^2 + \lambda ||f||_k^2.$$

El Teorema del representación garantiza que la solución a este problema de minimización puede ser escrita como:

$$f = \sum_{k=1}^{n} c_k k(x_k, \cdot).$$

para algún $c \in \mathbb{R}^n$. Utilizando propiedades básicas de RKHS, podemos reescribir la ecuación antes mencionada como:

$$\min_{c \in \mathbb{R}^n} \frac{1}{2} ||y - Kc||^2 + \frac{\lambda}{2} c^T K c.$$

Igualando la derivada con respecto a c a cero, vemos que c debe satisfacer:

$$(K + \lambda I)c = y.$$

Observamos que c existe y es único: K es semidefinida positiva, por lo que $K + \lambda I$ es definida positiva (para $\lambda > 0$). Definimos $G(\lambda) = K + \lambda I$. Frecuentemente, λ estará claro por el contexto, y simplemente escribiremos G.

Las predicciones en los puntos de entrenamiento estarán dadas por:

$$f(X) = Kc = K(K + \lambda I)^{-1} y = KG^{-1} y,$$

y la predicción en un nuevo punto de prueba X_* es:

$$f(X_*) = \sum_{k=1}^{n} c_k k(x_k, X_*) = k(X, X)^T c = y^T G^{-1} k(X, X_*).$$

Cabe destacar según (Cucker & Zhou 2007) no estamos sugiriendo formar G^{-1} explícitamente y manipularla. En las próximas secciones trataremos sobre los cálculos necesarios.

7.3. Minimización de la función de pérdida cuadrática

Para la función de pérdida cuadrática (Cucker & Zhou 2007) el objetivo es encontrar la función que minimiza:

$$\frac{1}{n}\sum_{i=1}^{n}\left(f\left(x_i\right)-y_i\right)^2+\lambda\left\|f\right\|_K^2.$$

Podemos incluir n en λ, obteniendo:

$$\sum_{i=1}^{n}\left(f\left(x_i\right)-y_i\right)^2+\lambda\left\|f\right\|_K^2.$$

Aplicando el Teorema de representación (Scholkopf *et al.* 2001) la solución al problema anterior tiene la siguiente forma:

$$f(x)=\sum_{i=1}^{n}c_iK(x_i,x).$$

Luego, en este trabajo se demuestra un Teorema de representación según (Scholkopf *et al.* 2001).

Que finalmente, en notación matricial podemos expresar como:

$$\left\|Y-Kc\right\|^2+\lambda\left\|f\right\|_K^2$$

Desarrollamos el segundo sumando:

$$\begin{aligned}\left\|f\right\|_K^2 &= <f,f>_K=<\sum_{i=1}^{n}c_ik\left(x_i,.\right),\sum_{i=1}^{n}c_ik\left(x_i,.\right)>_K\\ &= \sum_{i=1}^{n}\sum_{j=1}^{n}c_ic_j<k\left(x_i,.\right),k\left(x_i,.\right)>_K=\sum_{i=1}^{n}\sum_{j=1}^{n}c_ic_jk\left(x_ix_j\right)\\ &= c^tKc.\end{aligned}$$

Obteniéndose la siguiente expresión a minimizar:

$$\|Y - Kc\|^2 + \lambda c^t K c.$$

Esta representación nos permite obtener una solución relativamente sencilla al problema de minimización que planteamos en la Sección 5.2, según se propone en (Rifkin & Lippert 2007).

El enfoque de minimizar la función de pérdida cuadrática, junto con la regularización, permite obtener un modelo que equilibra la precisión en los datos de entrenamiento y la capacidad de generalizar a nuevos datos.

Capítulo 8

Aplicaciones del Aprendizaje automático

Como parte del trabajo, relevamos algunas aplicaciones de la Teoría del Aprendizaje Estadístico que resuelven con éxito una serie de problemas reales de la Ciencias y la Industria.

En particular la utilización de algoritmos de Aprendizaje Automático en muchos campos de la ciencia ha logrado mejorar los resultados de aplicar la técnicas tradicionales de cada disciplina.

En los últimos años la difusión y el éxito de las herramientas de aprendizaje automático aplicadas en las ciencias Naturales se ha "contagiado" a las Ciencias Sociales.

Relevamos el trabajo de Hainmuller y Hazlett del Departamento de Ciencias Políticas del M.I.T. donde proponen en (Hainmueller & Hazlett 2013a) la utilización de regresión con Mínimos Cuadrados Regularizados Kernel[1] para las investigación en Ciencias Sociales.

Esto resulta apropiado porque permite evitar fuertes supuestos paramétricos, permitiendo una interpretación similar a lo de los modelos lineales generalizados, permitiendo ademas interpretaciones mas complejas para analizar no-linealidades, interacciones y efectos heterogéneos.

La Teoría del Aprendizaje (Poggio & Smale 2005) es un camino para la comprensión del cerebro y para construír máquinas inteligentes.

De la misma forma en que Rosenblatt (Vapnik 1995) en los años 60 sugirió el primer modelo de aprendizaje estadístico, llamado perceptrón, inspirado en la literatura de la neurofisiología, hoy en día se desarrollan nuevas arquitecturas

[1] En inglés, KRLS: *Kernel Regularized Least Squares.*

de modelos inspiradas en la biología y sus correspondientes sistemas de visión computarizada.

Finalmente recorremos algunos de los resultados (Smale *et al.* 2009) que permiten establecer fundamentos matemáticos para modelos recientes diseñados sobre la base de datos fisiológicos y anatómicos que describen el cortex visual de los primates. Estos modelos permiten procesar cuantitativamente un conjunto de datos novedosos y proveer de un desempeño similar al humano en la categorización rápida de imágenes complejas.

8.1. Ejemplos de Aplicaciones para la Ciencias y la Industria

Los algoritmos de Aprendizaje Automático son utilizados en las más diversas ramas de la Industria y de la Ciencia.

8.1.1. Clinical survival analysys

La técnica EP-SVM[2] utiliza clasificación, implementando estrategias estocásticas para determinar el mejor kernel y los parametros del kernel para usar con un clasificador binario implementado con SVM.

Esta técnica ha sido desarollada y testeada con éxito en aplicaciones biomédicas tales como la detección de cáncer de senos y el colorectal.

En (Margolis *et al.* 2011) se comparan varios métodos en el cual EP-SVM supera a la regresión logística que a su vez supera a los métodos tradicionalmente aplicados.

Los resultados publicados en la anterior referencia demuestran una clara mejora al usar EP-SVM comparado con otros métodos.

8.1.2. Motion Estimation

El problema de estimación del movimiento se relaciona con la selección de modelo estastadístico en (Wechsler *et al.* 2004), donde el objetivo es seleccionar un modelo de movimiento a partir de varios posibles dado un conjunto finito de muestras con ruido. Se desarrolló una aplicación exitosa para el desafiante problema de estimar el modelo de movimiento a partir de pequeños conjuntos

[2]En inglés *Evolutionary Programming trained Support Vector Machine.*

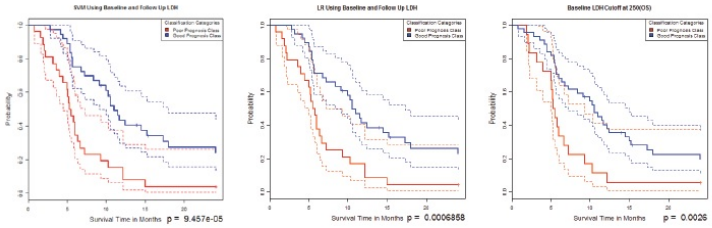

Figura 8.1: Comparación de los resultados contra otras técnicas tomada de (Margolis *et al.* 2011).

de datos de medidas de imágenes. De esta forma se realizan interpolaciones y extrapolaciones de movimientos a partir de secuencias de imágenes.

En (Wechsler *et al.* 2004) se demuestra que la selección de modelos basada en SLT[3], obtiene mejor desempeño frente a métodos alternativos de selección de modelos, como el FPE de Akaike, el criterio de Schwartz (SC), la validación cruzada generalizada (GCV) y el selector de modelos de Shibata (SMS).

Los datos de entrenamiento consisten en desplazamientos de puntos. La problema de estimar el movimiento corresponde a elegir el mejor modelo de movimiento de un conjunto de posibles modelos. Se utilizan métodos supervisados de regresión utilizando una función de perdida cudrática (squared loss) y técnicas de estimación del movimiento para elegir un modelo de flujo que tenga la mejor desempeño predictivo.

Se prueban en (Wechsler *et al.* 2004) la exitosa aplicación de la técnica a datos experimentales resultantes de una secuencia de imágenes de un brazo en movimiento y el correspondiente flujo normal.

8.1.3. Ubicación de dispositivos en redes inalámbricas

Técnicas y algoritmos se aplican al problema de determinar la ubicación de un dispositivo informático midiendo los valores de la fuerza de la señal captada por un conjunto de puntos de acceso.

La medición de la fuerza de la señal es una parte del modo de operación normal del equipamiento WIFI, lo que no implico la utilización de hardware específico para capturar los datos.

La técnica propuesta en (Battiti *et al.* 2002) se basan en SVM, y se implementan y comparan en su aplicación al mismo conjunto de datos con otros enfoques considerados en la literatura científica del tema. Se aplican tests en el

[3]Del inglés *Statistical Learing Theory*:SLT.

Sequence	*Experiments*	*Samples*	*Criteria*				
			vm	*fpe*	*gcv*	*sc*	*sms*
Non-noisy	*Interpolation*	*32*	*92.7%*	*70.8%*	*74.5%*	*67.1%*	*66.5%*
		64	*90.7%*	*66.3%*	*68.3%*	*65.7%*	*59.4%*
	Extrapolation	*32*	*100%*	*90.2%*	*94.3%*	*84.2%*	*80.2%*
		64	*100%*	*85.0%*	*90.0%*	*80.0%*	*84.3%*
	Whole	*32*	*98.8%*	*84.8%*	*86.6%*	*86.0%*	*81.0%*
	Interpolation	*64*	*99.6%*	*85.0%*	*86.2%*	*87.0%*	*82.0%*
Noisy	*Interpolation*	*32*	*92.9%*	*65.5%*	*71.5%*	*62.7%*	*60.0%*
		64	*93.0%*	*50.9%*	*54.2%*	*52.0%*	*46.0%*
	Extrapolation	*32*	*100%*	*76.8%*	*86.4%*	*74.0%*	*74.1%*
		64	*100%*	*80.0%*	*80.3%*	*68.5%*	*70.1%*
	Whole	*32*	*99.5%*	*85.3%*	*87.5%*	*79.2%*	*76.1%*
	Interpolation	*64*	*99.0%*	*86.2%*	*84.8%*	*86.0%*	*86.2%*

Figura 8.2: Comparación de los resultados contra otras técnicas tomada de (Wechsler *et al.* 2004).

entorno real demostrándose que los resultados son similares a los obtenidos con las técnicas tradicionales, con la ventaja de complejidad algorítmica baja en la fase de operación normal. Inclusive, el algoritmo utilizado es particularmente apto para la clasificación, donde supera otras técnicas.

8.1.4. Simuación de Imágenes Infrarojas

Un algoritmo de simulación en tiempo real de imágenes basadas en infrarojos se presenta en (Chaochao *et al.* 2007).

Las resultados de las simulaciones indican que el algoritmo basado en SVR[4] tiene mejor capacidad de generalización que otros métodos, lográndose una precision y calidad en tiempo real satisfactorios.

El principal problema de la simualción infrarroja de escenas es calcular el campo de la temperatura de superficie. Hasta ese momento, el método termodinámico es el utilizado para calcular la temperatura del objetivo y temperatura del background, tiene un modelo complejo y utiliza muchos recursos para realizar los cálculos, asumiéndose importantes simplificaciones y supuestos.

[4]Del inglés: *Support Vector Regression*: SVR.

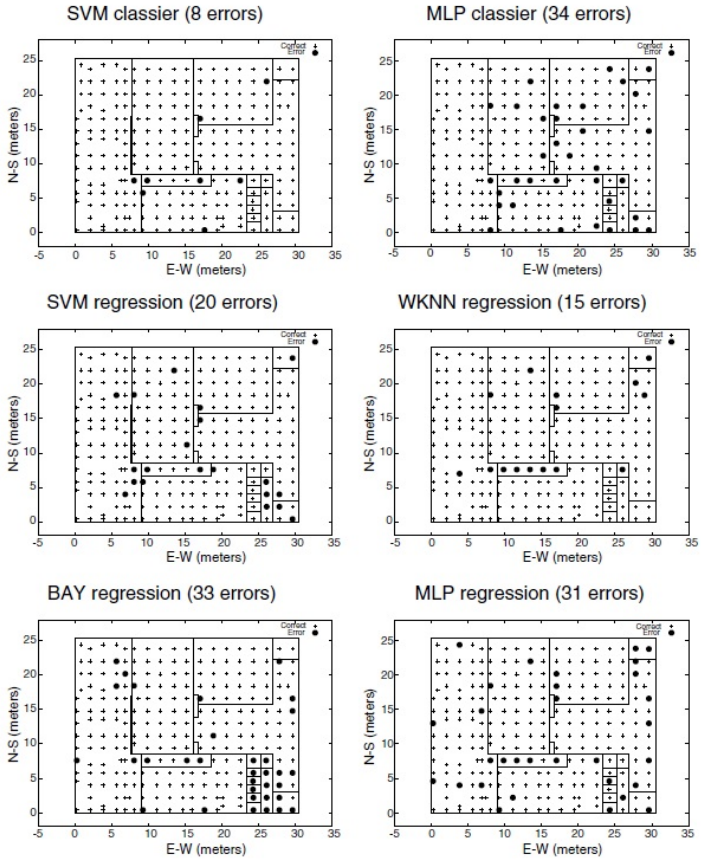

Figura 8.3: Comparación de los resultados contra otras técnicas tomada de (Battiti *et al.* 2002).

Arithmetic	Average temperature difference	Maximum temperature difference	Mean square error
ν - SVR	0.90825	3.2	0.86502
BP	1.57921	13.8	2.81321

Arithmetic	Average temperature difference	Maximum temperature difference	Mean square error
ν - SVR	0.51605	2.75	0.50052
BP	1.48296	5.56	1.31844

Figura 8.4: Comparación de los resultados para la escena del vehículo y del edificio respectivamente contra BP[5] tomada de (Chaochao *et al.* 2007).

Para superar estos problemas, en (Chaochao *et al.* 2007) se propone utilizar técnicas de Aprendizaje Supervisado.

La muestra de entrenamiento comprende de datos de entrada que son los que afectan la temperatura de la superficie de un objeto, y los datos de salida es la correspondiente temperatura.

Los datos de entrada a utilizar son entre otros la longitud, latitud, hora de la medición, temperatura del aire, presión del aire, humedad, coeficiente de nubes, velocidad el viento. En caso de que el objeto sea por ejemplo un vehículo en movimiento, se agregaran los datos de la velocidad y running time del mismo.

8.1.5. Aplicación a las Ciencias Sociales

Investigadores del Departamento de Ciencias Políticas del M.I.T. proponen en (Hainmueller & Hazlett 2013a) la utilización de regresión con Mínimos Cuadrados Regularizados Kernel[6] en las Ciencias Sociales. Argumentan que este metodo es apropiado para las investigación en Ciencias Sociales porque permite evitar fuertes supuestos paramétricos, permitiendo una interpretacion similar a lo de los modelos lineales generalizados, permitiendo ademas interpretaciones mas complejas para analizar nolinealidades, interacciones y efectos heterogéneos.

Los autores extienden el método de forma que sea mas útil para la investigación social:

- Derivan estimadores para los efectos marginales puntuales y sus varianzas.
- Establecen el insesgamiento, consistencia y normalidad asintótica de los estimadores bajo ciertos supuestos sencillos.

[6]En inglés *Kernel Regularized Least Squares*:KRLS.

- Proponen una regla automatizada simple para seleccionar el ancho de banda del kernel.
- Proporcionan software, un paquete de R (Hainmueller & Hazlett 2013b).

8.2. Hacia la comprensión del funcionamiento del cerebro

La Teoría del Aprendizaje es un camino para comprender la inteligencia (Poggio & Smale 2005) tanto en cerebros como en máquinas, permitiendo la construcción de máquinas inteligentes que aprenden de la experiencia y mejoran sus capacidades de la misma forma que lo hacen los niños.

Si se compara las máquinas que aprenden descritas por la teoría del aprendizaje clásica, tales como las máquinas kernel, con el cerebro, una de las diferencias principales es la aparente habilidad de los humanos y animales de aprender utilizando muy pocos ejemplos. La comparación con cerebros reales ofrece otro desafio relacionado a la Teoría de Aprendizaje. Los algoritmos de aprendizaje clásicos corresponden a una arquitectura de una sola capa. El estudio del cortex del cerebro sugiere una arquitectura jerárquica, que permite tratar el problema de la complejidad de la muestra.

Se han establecido los fundamentos matemáticos (Smale *et al.* 2009) para modelos inspirados en la arquitectura fisiológica y anatómica del córtex visual de los primates.

De esta forma la Neurociencia provee de nuevas ideas y enfoques para el aprendizaje estadístico y la visión computarizada.

Los recientes modelos desarrollados logran un desempeño similar al humano en la categorización rápida de imágenes complejas.

Estos modelos se agrupan en una reciente familia de arquitecturas inspiradas en la biología y sus correspondientes sistemas de visión computarizada. Estos modelos son de estructura jerárquica al igual que el cortex.

La definición de la respuesta neural y su kernel derivado están basados en la recursividad que define una jerarquía de los kernels locales y pueden ser interpretados como una arquitectura multicapa en donde las capas son asociadas con el incremento de escalas espaciales. En cada capa los kernels derivados son construidos en forma recursiva, obteniendo el máximo dentro de un conjunto de transformaciones. Este modelo, si bien puramente matemático,

tiene un componente de significado clave: un sistema que permite vincular el desarrollo matemático con problemas del mundo real.

Capítulo 9

Aplicación a Datos Quimiométricos

Presentamos una aplicación práctica de las técnicas de Aprendizaje Automático discutidas en los capítulos anteriores, utilizando un conjunto de datos reales proporcionado por un laboratorio en Montevideo. El objetivo general de este capítulo es desarrollar un sistema capaz de predecir automáticamente valores químicos a partir de datos de espectroscopía de infrarrojo cercano (NIRS).

La motivación para este trabajo se basa en que los métodos analíticos tradicionales pueden ser lentos y costosos. La espectroscopía NIRS ofrece una alternativa más rápida y no destructiva para la medición de componentes químicos, con aplicaciones en diversas industrias como la agricultura, la petroquímica y la farmacéutica. El desafío es utilizar Aprendizaje Automático para construir modelos que relacionen los espectros NIRS con los valores químicos de interés, y para esto se buscará mejorar la precisión de las predicciones y, en lo posible, automatizar el proceso de análisis.

Se hará especial énfasis en las Máquinas de Vectores de Soporte (SVM). El objetivo planteado por los expertos de campo es obtener un coeficiente de determinación R^2 de al menos 0.92 al predecir valores químicos a partir de los datos NIRS. Se explorará cómo, a través de la eliminación del ruido y la reducción de la dimensionalidad, SVM puede lograr valores de R^2 que superen este objetivo. Además, se compararán los resultados obtenidos con SVM con otras técnicas de aprendizaje automático.

9.1. Quimiometría

El término quimiometría (Chau *et al.* 2004) fué introducido por Wold y Kowalski a principios de los 70s. En forma similar los términos como biometría y econometría fueron introducidos, por ejemplo, en los campos de Ciencias Biológicas y la Economía.

Según Wold (Wold 1995) la Quimiometría[1] estudia cómo obtener información química relevante a partir de datos de mediciones, cómo se representa y despliega esa información, y cómo se convierten esos datos en información valiosa. Puede ser considerada como una subdisciplina que provee de teoría básica y metodología a la química analítica moderna.

Dos razones principales (Chau *et al.* 2004) han facilitado la evolución de la Quimiometría:

- Es posible adquirir grandes cantidades de datos a través de instrumentos químicos avanzados.
- El incremento en el poder de cómputo amplió las capacidades en el procesamiento de señales y la interpretación de los datos químicos.

En este contexto, similar al que se produce en otras disciplinas, un conjunto de técnicas basadas en aprendizaje supervisado se constituyen en una tecnología clave para extraer información y darle sentido al océano de bits que nos rodea (Poggio & Smale 2005).

Con estas herramientas (nuevas tecnologías y teoría) el laboratorio químico analítico se ha transformado. En particular, la aplicación de las técnicas NIRS compiten en reducción de costos y velocidad con muchas técnicas analíticas de la tradicional "química húmeda".

9.2. La técnica NIRS

La espectroscopía de infrarrojo cercano[2] se refiere al uso de radiación infrarroja dentro del rango de longitud de onda de 780 a 2526 nm (Nadeem & Heindel 2018). Cuando la radiación NIR pasa a través de la muestra, se absorbe preferentemente dependiendo de los sobretonos y las combinaciones de las frecuencias fundamentales de los enlaces químicos encontrados dentro de la

[1] En inglés *Chemometrics*.
[2] En inglés *Near Infra Red Spectroscopy*: NIRS.

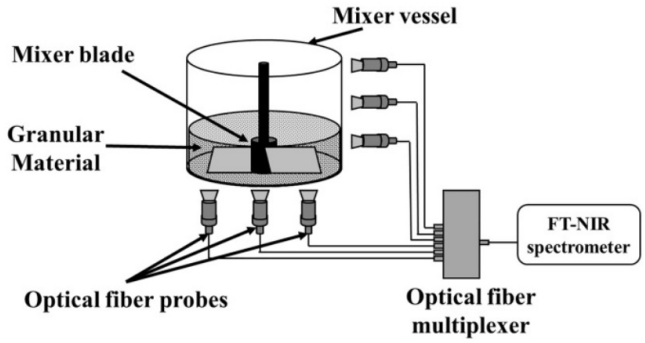

Figura 9.1: Esquema del funcionamiento de un espectrómetro NIRS tomado de (Nadeem & Heindel 2018).

misma. La mayoría de los enlaces covalentes como OH, NH y CH se pueden detectar en el rango cercano al infrarrojo del espectro.

Una configuración del equipamiento NIRS típica, como se ilustra en las figuras 9.1 y 9.2 (Nadeem & Heindel 2018), consta de un espectrómetro NIR que incluye una fuente de luz, un monocromador y un detector que detecta las señales de luz que se transmiten o reflejan desde la muestra.

El NIRS ofrece flexibilidad en el despliegue de las sondas según el tipo de material y proceso que se esté monitoreando. Las sondas se pueden colocar en las paredes del recipiente para permitir mediciones por contacto o se pueden realizar mediciones sin contacto donde la muestra pasa a través de la línea de visión de los sensores sin tocarla. Las sondas normalmente cumplen la función de iluminar el espacio de la muestra y de recoger la señal de luz reflejada. Los datos así recogidos suelen enviarse a un ordenador para su posprocesamiento y análisis estadístico.

Una vez que se adquieren los datos NIR, se aplican diferentes técnicas de procesamiento de datos para interpretar los datos en las mediciones de homogeneidad de la mezcla. La medida de la desviación estándar en la longitud de onda y el dominio del tiempo proporciona información sobre la homogeneidad general y en tiempo real de la mezcla. La observación directa de los espectros IR permite la medición cualitativa de la homogeneidad de la mezcla, mientras que el análisis estadístico de las variaciones de los espectros a lo largo del tiempo proporciona una estimación objetiva de los puntos finales de la mezcla, es decir, los puntos en los que se han alcanzado niveles máximos o aceptables de homogeneidad.

9.3. Descripción del Problema

El problema planteado consiste en obtener una función que permita predecir valores de un componente químico a partir de datos obtenidos a través de espectroscopía de infrarojo cercano[3], como alternativa al análisis químico tradicional, lento y costoso.

9.3.1. La muestra: datos quimiométricos

Se cuenta con una muestra de 832 elementos, que fue dividida al azar en una muestra de entrenamiento (80 % de los datos) y una de test (20 %) por los expertos de campo que nos proponen el problema.

Cada elemento de la muestra cuenta con 2201 características que corresponden a las absorbancias, que es el logaritmo del inverso de las reflectancias (Khoshhesab 2012) correspondientes a cada nivel de banda analizado.

El sistema opera en el modo de reflectancia[4] mediante el uso de un haz de fibras ópticas. Este haz transporta la energía luminosa a la muestra, que se centra en una ventana de vidrio de cuarzo por encima de la punta de las fibras. Al llegar a la muestra, la luz penetra en el material de interés y la energía no absorbida se refleja difusamente desde la muestra. La energía reflejada es recogida por varios detectores de sulfuro de plomo (PbS) montados en ángulos de 45° con respecto al plano del vidrio. La energía reflejada se mide y se utiliza para cuantificar el grado de absorción en cada longitud de onda.

[3]En inglés *Near Infra Red Spectroscopy*:NIRS.

[4]La reflectancia es la medida de la cantidad de luz y calor que refleja o absorbe un determinado color o superficie. Se centra específicamente en la cantidad de luz y calor que se refleja o absorbe, más que en la intensidad o viveza de un color.

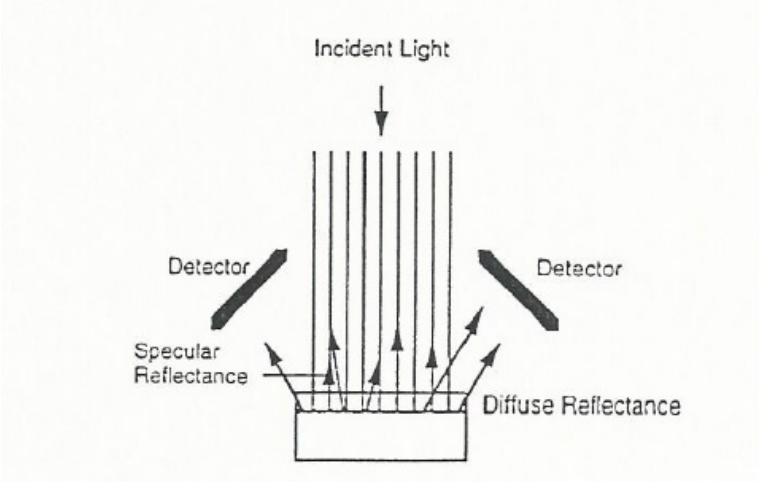

Figura 9.2: Esquema de la técnica NIRS tomado de (Thosar *et al.* 2001).

Es decir x_i, el elemento i-ésimo de la muestra, es un vector de dimensión 2201 $(x_{i,1}, ..., x_{i,2201})$ tal que $x_{i,j} = log\frac{1}{R_{i,j}}$, siendo $R_{i,j}$ la reflectancia en la banda j-ésima.

Esta transformación es conocida como transformación de absorbancia (Zhai *et al.* 2013).

Cada elemento de la muestra está etiquetado de la forma:

{vector de características, valor químico)}.

El valor del componente químico específico es medido por métodos de análisis químico tradicional. En particular, utilizaremos el valor correspondiente al porcentaje de contenido de Nitrógeno Total.

9.3.2. Objetivo

Se desea un sistema capaz de predecir el Valor Químico en forma automática, en función de un nuevo vector de características obtenido por espectroscopía de infra-rojo cercano.

La capacidad de predicción del modelo se mide en la industria, según los expertos de campo, a través del coeficiente de determinación (R^2) entre los valores predichos y los medidos.

El valor objetivo a superar del R^2, propuesto por los expertos de campo, es de 0.92.

9.3.3. Enfoque

Buscaremos entonces ajustar una función a la muestra de entrenamiento compuesta por un conjunto de datos de alta dimensionalidad obtenidos por NIRS y su correspondiente variable de respuesta obtenida mediante análisis químico.

La literatura relevada (Zhang *et al.* 2008) indica que se presentan relaciones no lineales entre los datos del espectro y las variables cuantitativas de interés.

Esta situación y dado que los datos presentan alta dimensionalidad, hace que resulten inadecuadas las técnicas tradicionales de regresión lineal.

Proponemos entonces obtener la función de predicción utilizando métodos de aprendizaje supervisado, en particular métodos kernel de regresión no lineal.

Para el tratamiento de los datos, se evaluará la aplicación de técnicas de reducción de la dimensionalidad (Velliangiri *et al.* 2019) de los datos del espectro, utilizando métodos de análisis funcional considerando que, por su naturaleza, los datos espectrales pueden ser considerados funcionales (Ana M *et al.* 2013).

9.3.4. Alcance

Si bien el objetivo final es proporcionar un sistema de cálculo automático, para esta etapa plantearemos un modelo y su correspondiente función predictora que cumpla con los requisitos señalados.

9.4. NIRS y Aprendizaje Automático

La reflectancia cercana al infra-rojo ha probado ser una herramienta analítica poderosa. Ha sido utilizada ampliamente en las industrias agrícola, petroquímica, textil y farmacéutica.

Se ha desarrollado para aplicaciones tales como la determinación cuantitativa de nutrientes en las industria agrícola y alimenticia. Por ejemplo la determinación de agua, proteínas y grasa dentro de muestras complejas tales como granos y leche (Stenlund *et al.* 2009).

Específicamente, la aplicación de la espectroscopía de infra-rojo cercano para el análisis de muestras farmacéuticas se ha desarrollado en forma significativa en lo que va del siglo (Zhang *et al.* 2008).

Dado que la materia orgánica controla el metabolismo de las plantas, es importante estimar y monitorear sus componentes bioquímicos.

La concentración de estos componentes puede calcularse mediante el análisis de laboratorio de un conjunto de muestras. Si bien estas estimaciones son precisas, estos métodos son costosos, llevan tiempo, son destructivos y complejos (Blackmer *et al.* 1994). En cambio, a través de la espectroscopía NIRS es posible la determinación de la concentración de varios elementos a partir de un solo escaneo no destructivo Rossel & Behrens (2010).

La reflectancia espectral de las hojas de las plantas responde a las acciones combinadas de varios factores químicos y físicos de la planta, como por ejemplo el contenido interno de componentes bioquímicos, distribución y organización de las células y contenido de agua. Es entonces que la reflectancia visible y cercana al infra-rojo tiene el potencial de permitir estimar el nitrógeno, el fósforo, el potasio u otros componentes bioquímicos (Zhai *et al.* 2013).

9.5. Bases Teóricas de Regresión y Métodos Kernel para Análisis Quimiométrico

Dada una muestra de n elementos $\{(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)\}$ buscamos encontrar una función f tal que $f(x) = \hat{y}$ sea un buen predictor de y para un input futuro x.

Para el caso de estudio un elemento (x_i, y_i) de la muestra consiste en:

- x_i es un vector de dimensión d que contiene los valores espectrales.
- y_i representa la variable cuantitativa de interés, el contenido de Nitrógeno de la muestra.

Para el tratamiento de este tipo de problemas, la literatura plantea el uso de métodos lineales y sus variaciones sesgadas, así como sus competidores modernos tales como regresión no paramétrica, redes neuronales y support vectors machines.

Los métodos de regresión sesgados fueron utilizados originalmente en quimiometría (aplicados a la investigación de alimentos, estudios de contaminación ambiental, etc.) donde resulta habitual que el número de variables (dimesión del vector x_i) sea superior a la cantidad de observaciones ($d > n$).

Se han desarrollado entonces métodos sesgados como la Regresión de Componentes Principales, Regresión de Mínimos Cuadrados Parciales y la Regresión Ridge.

En particular, la regresión con mínimos cuadrados parciales (PLSR) ha sido utilizada tradicionalmente en química como método de calibración multivariada y en particular es un algoritmo ampliamente usado para modelar datos NIRS.

9.5.1. Métodos no Lineales: Métodos Kernel

La literatura más reciente indica que la técnica lineal PLSR[5] no es capaz de obtener resultados precisos el modelado NIRS, y que tampoco supera el reto de generar una función predictiva para una muestra limitada (Chaochao *et al.* 2007). En los experimentos realizados, verificamos que dado que la muestra tratada es amplia, se pueden obtener resultados primarios satisfactorios con PLSR.

Las técnicas basadas en Kernels (Aronszajn 1950) son uno de los desarrollos más importantes dentro de los algoritmos de aprendizaje automático.

Las representaciones Kernel ofrecen una solución alternativa proyectando los datos en un nuevo espacio de características de alta dimensión para para mejorar la capacidad de las máquinas lineales.

Máquinas de Vectores de Soporte

Las características especiales y el excelente desempeño empírico de las SVMs[6] en el campo de la química se demuestra en destacadas publicaciones científicas de los últimos años (Li *et al.* 2009).

SVM fue extendido por Vapnik para la regresión (Lipkowitz *et al.* 2007) y se convierte en uno de los métodos aplicados en las publicaciones científicas más recientes sobre modelado NIRS (Li *et al.* 2009) y específicamente para nuestro caso de estudio (Zhang *et al.* 2008).

9.5.2. Métodos de extracción de características para datos fucionales

Wavelets

La Transformación Wavelet (Morettin & Pinheiro 2017) es un método matemático basado en la transformada de Fourier, que puede ser usado en la compresión de datos, filtros para manejar el suavizado y el ruido, validación baseline y análisis de multicomponentes superpuestos de señales (Shao *et al.* 2007).

[5] En inglés *Partial Least Squares Regression.*
[6] En inglés *Support Vector Machines.*

Según (Morettin & Pinheiro 2017) el análisis funcional de datos[7] basado en wavelets es un enfoque moderno para tratar con la inferencia estadística cuando las observaciones son curvas o imágenes. Realizar inferencia (estimación y pruebas) en el dominio de las wavelets es beneficioso en varios aspectos como: reducción de dimensionalidad, decorrelación, localización y regularización.

El análisis Wavelet genera una estimación del contenido de una frecuencia local de una señal representando los datos utilizando una familia de funciones wavelet que varían en escala y posición.

La señal puede ser reconstruida con precisión con una cantidad relativamente pequeña de componentes (Zhang *et al.* 2008).

WT[8] puede ser entonces utilizado para comprimir los datos espectrales obtenidos en NIRS.

Transformación Wavelet Discreta

La Transformación Wavelet Discreta (que abreviaremos como DWT) está basada en un filtro low pass H y un filtro high pass G y una decimaciónn binaria (Liang *et al.* 2009).

La DWT[9] de un vector de datos puede ser calculada rápidamente usando un banco de filtros. La estructura básica del banco de filtros comprende un par de filtros H/G seguidos por una operación de muestreo hacia abajo (down-sampling) que consiste en descartar el resto de los puntos de la salida de los filtros.

Los filtros se seleccionan de forma tal que la transformación sea invertible, preservando entonces la información de la señal.

La salida de los filtros del canal low-pass pueden ser descompuestos a su vez por sucesivos pares de filtros hasta cierto número N_{it} de iteraciones.

Transformada de Fourier Rápida

Las plantillas utilizadas en FFT[10] son ondas de senos y cosenos con diferentes frecuencias. De esta forma las técnicas FFT nos pueden decir fácilmente la información de frecuencias global de una señal. Pero (Chau *et al.* 2004), en algunos casos lo que se desea es encontrar es algunos picos espectrales correspondientes a ciertos químicos en el análisis espectral. También se necesita

[7] En inglés *Functional Data Analysis*: FDA.
[8] En inglés *Wavelet Transformation*: WT.
[9] En inglés *Discrete Wavelet Transformation*: DWT.
[10] En inglés *Fast Fourier Transformation*.

determinar las frecuencias locales, lo que no puede ser realizado sencillamente con la información extraída utilizando la Tranformada de Fourier Rápida. Es entonces cuando se hacen necesarias las técnicas wavelet, que son las que utilizaremos en este trabajo.

Estas técnicas son aplicadas con éxito en (Shao *et al.* 2007) y (Zhang *et al.* 2008) frente a datos similares a los de nuestro problema.

9.5.3. Revisión bibliográfica

A continuación mencionamos las principales referencias prácticas estudiadas, relevantes para este trabajo.

SVM en química y en NIRS genérico

En una primera etapa, relevamos trabajos sobre técnicas estándar y SVM aplicados en la química, para luego buscar en concreto su aplicación en el modelado NIRS.

Máquinas de Vectores de Soporte y su aplicación en química

En Li *et al.* (2009) se presenta a SVM como un método adecuado tanto para clasificación como para regresión aplicados a problemas de la química. Específicamente se trata un problema de regresión en el campo de la quimiometría, donde se plantea predecir el punto de ebullición a partir del espectro de infrarrojo cercano de muestras de diesel. Los resultados del modelo SVMr se comparan con los de un modelo PLS (partial least squares). Se destaca que al modelo PLS, usado como referencia, se le hace imposible tener en cuenta las relaciones no lineales (James *et al.* 2014).

Se dispone de una muestra de 246 elementos. Se dividen los datos al azar en una muestra de entrenamiento y otra de test. El rango del espectro utilizado para ajustar el modelo es 760-1100 nm.

Puntos destacables:

- En su introducción teórica, en el punto 5.2 se explica como se transforma el problema de clasificación SVM en uno de regresion.
- En el punto 6.3 se plantea la extracción de outliers de la muestra de entrenamiento antes de generar el modelo.
- Se transforman los datos NIRS con diferenciación de primer órden.
- Para SVMr se utiliza el método ν-SVR con kernel RBF

- Se utiliza el parámetro γ sugerido por defecto por el paquete *libsvm* Se seleccionan los parámetros C y ν a través de un algorimo genético.
- En 7.3 se indica que se le hace dificil explicar los resultados al investigador.

Un método SVM con espectroscopía NIRS utilizando wavelets

En Liang *et al.* (2009) se propone utiliza la regresión SVM para predecir la concentración de ciertos elementos a partir de datos NIRS de tabletas de medicamentos.

Se dispone de una muestra de 36 elementos. 24 se usan como datos de entrenamiento y 12 de test. El espectro se registra en el rango 760-1100 nm.

Destacamos:

- En la introducción se indica que se usa la transformación wavelet para preprocesar los datos NIRS, de forma de eliminar parte del ruido. Para ello se descartan los coeficientes pequeños obtenidos antes de realizar la transformación wavelet inversa para obtener la señal sin ruido.
- En 3.1 se indica que se utiliza la wavelet de Daubechies db8
- Se utiliza PLS como modelo de referencia
- En el registro de los datos se realiza un promedio de cuatro escaneos.
- En 4 se menciona el preprocesamiento SNV.
- En 4 se menciona que se utiliza ν-SVM. Kernel RBF con validación cruzada 5 para determinar los parámetros del modelo.(ν, σ, C).

Utilizando minería de datos para modelar e interpretar el espectro de reflectancia difusa del suelo

En Rossel & Behrens (2010) se presenta un abundante comparativo de distintos algoritmos para calibrar espectros de refectancia vis-NIR para para predecir el contenido SOC, CC y PH de muestras de suelo.

Se utilizan métodos PLSR, MARS, RF, BT, SVM. Se plantea el preprocesamiento de los datos utilizando DWT. Se utilizan métodos de selección de atributos o procedimientos de ranqueo en el dominio del espectro y también en el dominio wavelet.

Se dispone de una muestra de 1104 elementos. El espectro se registra en el rango 350-2500 nm en 876 anchos de banda. Al usar técnicas de reducción de

características se utilizan entre 72 y 137 coeficientes wavelet y entre 11 y 31 componentes principales.

Tomamos nota especialmente de:

- Se destaca el poder de la espectroscopía NIR que permite la determinación de la concentración de varios elementos a partir de un solo escaneo no destructivo.
- En 3.2.1.2 al modelar con los coeficientes wavelet se ordenan los coeficientes por su varianza como criterio de selección.
- El modelo SVM es el que obtiene las mejores predicciones, que inclusive son mejoradas con el tratamiento DWT.

NIRS aplicado a hojas de plantas

Observamos que el modelado NIRS cambia cualitativamente dependiendo del objeto de análisis. Es decir, son muy distintos los tipo de relaciones entre los valores del espectro y la variable cuantitativa a predecir, según sea el objeto de estudio el suelo, carne u hojas de vegetales.

Para aproximarnos a nuestro objeto de estudio, el tabaco, revisamos trabajos relativos a técnicas de modelado NIRS sobre hojas de vegetales en general y para el tabaco en particular.

Estimación de los contenidos de nitrógeno, fósforo y potasio en las hojas de diferentes plantas mediante espectroscopia de reflectancia del infrarrojo cercano y visible en laboratorio: comparación de los métodos de regresión de mínimos cuadrados parciales y regresión con Máquinas de Vectores de Soporte

En Zhai *et al.* (2013) se buscan modelos para estimar los algunos de los componentes de la materia orgánica de las plantas: Nitrógeno, fósforo y potasio.

Se compara modelos PLSR y SVM para predecir los componentes mencionados.

Se utiliza una muestra de 95 hojas de muy diversas especies.

Se aplican varias técnicas de preprocesamiento.

Tomamos notas especialmente de:

- Técnicas de preprocesamiento.
- Técnicas de análisis de correlación.

- Los R^2 nunca son superiores a 0.7, considerando que son variedades de hojas de distintas especies.

En Shao *et al.* (2007) se plantea un modelo de clasificación para predecir la marca de cigarrillos en base a datos NIRS.

En la investigación se usa el rango de 340 a 1000 nm. La muestra es de 100 elementos con 661 datos de espectro. Luego del proceso WT se usan 21 datos. La muestra se divide al azar en 80 elementos para entrenamiento y 20 para testing.

Se destacan:

- Se aplican dos tipos de preprocesamiento. Suavizado Savizky-Golay con un gap de tres puntos. Mutiplicative scatter correction (MSC).
- Se utiliza la transformación wavelet Daubechies.
- Se utilizan modelos de BP-redes neurales.

En Zhang *et al.* (2008), se trata un problema muy similar al nuestro, pero para una muestra reducida.

Se detacan:

- Utilización de DWT para compresión del espectro.
- Diversas técnicas aplicadas al pre-procesamiento de los datos.
- Suavizado previo al procesamiento de los datos.

Hemos hecho foco en el marco teórico de la aplicación de técnicas de aprendizaje automático a datos quimiométricos, específicamente en el contexto de la espectroscopia NIRS. Se exploran los métodos kernel como una alternativa a los métodos lineales, destacando su capacidad para modelar relaciones no lineales en los datos. Dentro de estos métodos, se enfatiza el uso de Máquinas de Vectores de Soporte (SVM), tanto para regresión como para clasificación, debido a su buen desempeño empírico en química. Además, se revisan métodos de extracción de características, como la Transformación Wavelet Discreta (DWT), que permite reducir la dimensionalidad de los datos espectrales, considerándolos funcionales, y eliminar el ruido. En resumen, el marco teórico establece la necesidad de utilizar métodos no lineales, como SVM, para modelar las relaciones complejas entre los espectros NIRS y las propiedades químicas. También, se destaca la importancia del preprocesamiento de los datos, mediante técnicas de reducción de dimensionalidad y eliminación de ruido, para mejorar el

rendimiento de los modelos de aprendizaje automático. La combinación de estos métodos, en particular SVM con coeficientes wavelet, es presentada como un enfoque prometedor para abordar el problema de la predicción de componentes químicos a partir de datos NIRS.

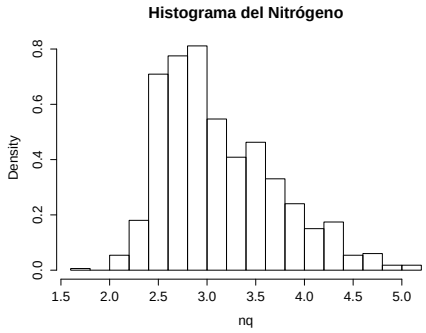

Figura 9.3: Histograma de los valores de N, nuestra variable de respuesta, $Min = 1,7$ $Media = 3,1$ $Max = 5,1$. Elaboración propia.

9.6. Estadística Descriptiva

9.6.1. Valor químico

El valor químico que utilizaremos como variable de respuesta en nuestro trabajo es el porcentaje de contenido de Nitrógeno total.

En la figura 9.3 presentamos un histograma de los valores registrados en la muestra.

9.6.2. Espectros y su relación con el Nitrógeno

Primero intentamos sin éxito visualizar alguna relación entre los espectros y el nivel de nitrógeno medido. Para ello coloreamos los espectros según su valor de Nitrógeno correspondiente.

La figura 9.4 muestra las curvas de los resultados del espectrómetro coloreadas según los distintos niveles de porcentaje de Nitrógeno total. Coloreamos en rojo/azul/verde según el nivel de nitrógeno. A simple vista no se observa una relación entre, por ejemplo, la altura de las curvas (su distancia del eje x) y el valor de nitrógeno. En la figura 9.5 dibujamos el conjunto de curvas como una superficie, ordenando las curvas por su valor de nitrógeno correspondiente.

Son 832 líneas, en las que se superponen colores: como era de prever no observamos ninguna relación.

Luego en la figura 9.5 procedimos a ordenar los espectros según su etiqueta de valor químico y los graficamos como una superficie. A simple vista no fuimos capaces de encontrar relación alguna.

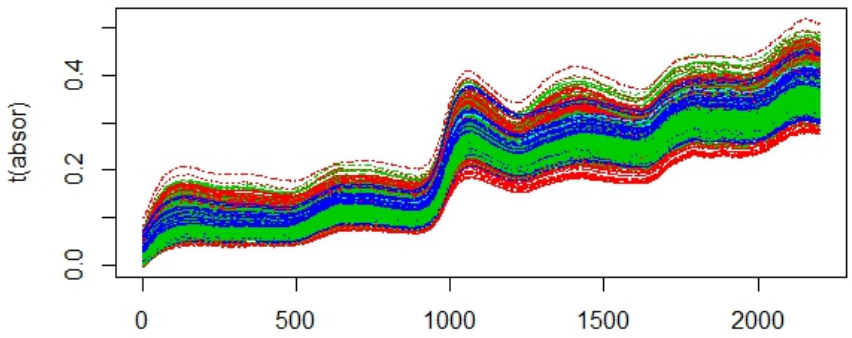

Figura 9.4: Gráfico de los espectros coloreados según valor químico.

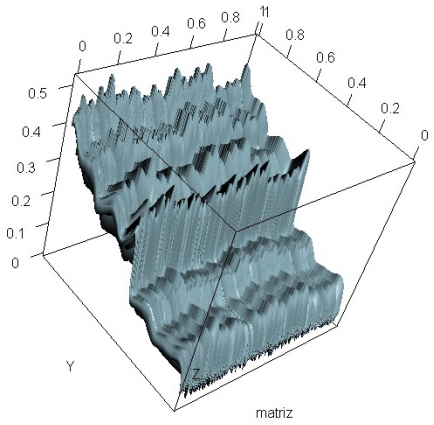

Figura 9.5: Superficie de los espectros.

9.6.3. Correlación de los componentes del espectro

Calculamos el coeficiente de correlación de cada variable predictora contra la variable de respuesta (Espectro Original) y lo presentamos en el gráfico de la figura 9.6.

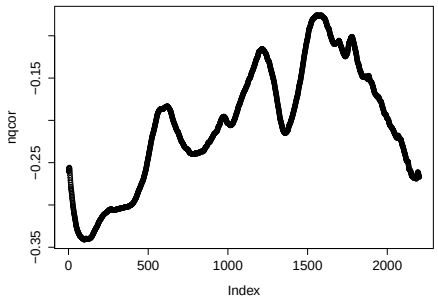

Figura 9.6: Correlación de las variables predictoras con el valor químico.

9.6.4. Datos atípicos (outliers)

Outliers funcionales Utilizaremos el concepto de *profundidad estadística* para datos funcionales (Cuevas *et al.* 2007). Estas medidas de profundidad son útiles para definir medias de posición y dispersión, para clasificación y para detección de outliers. A través del paquete fda.usc (Febrero-Bande & Oviedo de la Fuente 2012) del software R (R Core Team 2014) es posible detectar valores atípicos a partir de la profundidad estadística.

En particular, utilizamos una nueva herramienta visual denominada Outliergram (Arribas-Gil & Romo 2014)

El outliergram es una técnica para visualizar y detectar valores atípicos en datos funcionales, específicamente en curvas. Se basa en la relación entre dos medidas de profundidad: la profundidad de banda modificada (MBD) y el índice de epígrafe modificado (MEI). La MBD mide la proporción de tiempo que una curva pasa dentro de las bandas formadas por otras curvas, mientras que el MEI mide la proporción de tiempo que una curva está por debajo de las demás. Al graficar los valores de MBD versus MEI para cada curva, se observa que las curvas con formas similares tienden a formar una parábola, mientras que las curvas con formas atípicas se alejan de esta.

Para la detección de valores atípicos, el outliergram (para nuestros datos, graficado en la figura 9.7) utiliza la distancia vertical de cada punto (MEI,

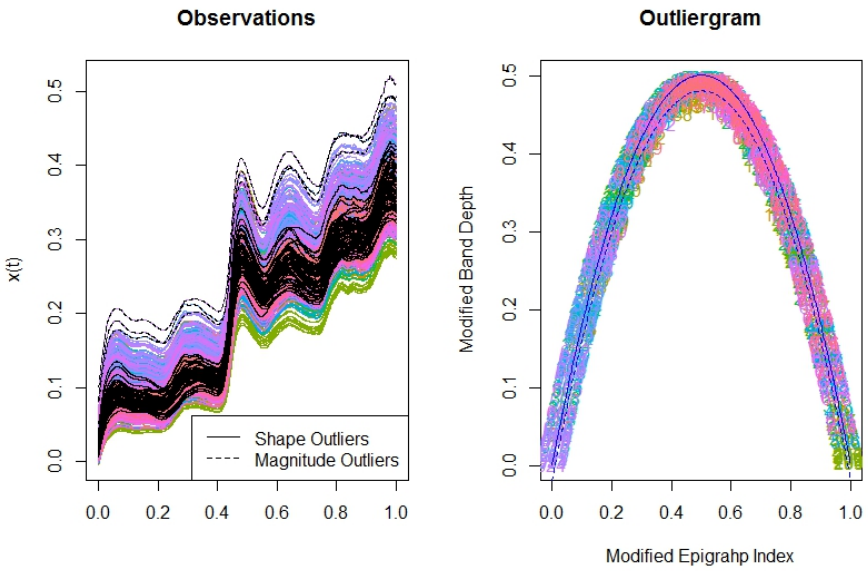

Figura 9.7: Salida del Outliergram.

MBD) a la parábola que se forma. Aquellas curvas que se encuentren a una distancia significativamente grande de esta parábola son consideradas valores atípicos de forma. Además, el algoritmo considera un paso adicional para curvas que se encuentran por encima o por debajo de la mayoría de las curvas, donde estas son desplazadas verticalmente hacia el centro de la muestra para evaluar si su forma es atípica.

Los outliers de forma detectados son los siguientes: 101 140 159 158 139 71 115 735 116 147 824 138 143 160 287 823 635 130 229 72 289 829 477 141 230 634 104 632 117 642 100 388 633 137 641 640 134 636 133 99 119 285 663 132 630 290 150 157 687 661 219 220 112 221 3 664

Los outliers de magnitud detectados son los siguientes: 554 623 624

Outliers de datos expresados como coeficientes wavelet

Se utilizan algoritmos aplicables a la detección de outliers en datos multivariados, utilizando los coeficientes wavelet lo que presumimos que puede ampliar el alcance de los datos atípicos detectados a través del outliergram.

En particular utilizamos el algoritmo LOF (Breunig *et al.* 2000) y (Torgo 2010).

El método LOF[11] es un enfoque para la detección de valores atípicos en conjuntos de datos multidimensionales. A diferencia de otros métodos que con-

[11]Factor de Outlier Local.

sideran si un objeto es o no un valor atípico como una propiedad binaria, LOF asigna a cada objeto un grado de atipicidad, indicando cuán aislado está un objeto con respecto a su vecindad local. Este enfoque es particularmente útil en escenarios complejos donde las estructuras de los datos pueden variar en densidad, ya que el factor LOF de un objeto se calcula en relación a la densidad de su entorno inmediato, y no a una vista global del conjunto de datos.

Para calcular el LOF de un objeto, se consideran varios conceptos: la distancia k de un objeto, que es la distancia a su k-ésimo vecino más cercano, y la vecindad k-distancia, que incluye todos los objetos que están a una distancia no mayor a la k-distancia. También se define la distancia de alcanzabilidad de un objeto con respecto a otro, que es la mayor entre la distancia real entre los objetos y la k-distancia del segundo objeto. A partir de estas definiciones, se calcula la densidad de alcanzabilidad local de un objeto, que es el inverso del promedio de las distancias de alcanzabilidad con respecto a sus vecinos más cercanos. Finalmente, el LOF se calcula como el promedio del ratio entre la densidad de alcanzabilidad local de un objeto y las densidades de sus vecinos. Un LOF cercano a 1 indica que el objeto está en una región con densidad similar a sus vecinos, mientras que un LOF mucho mayor que 1 indica que es un valor atípico local.

Los primeros outliers detectados con este criterio son los siguientes elementos de la muestra: 665 666 164 556 72.

9.7. Metodología

9.7.1. Pre-procesamiento

Las variables predictoras son vectores cuyos elementos están muy correlacionados (se puede decir que son datos funcionales). En este contexto se hará necesaria la utilización de prácticas de reducción de la dimensionalidad aplicada a datos funcionales.

Como problema adicional, en los datos originales, es mayor la dimensión de los datos que el tamaño de la muestra.

Tomamos dos caminos para tratar estos problema:

- Utilizar métodos sesgados aplicables cuando $n < d$, que son estándares en la literatura NIRS, que adicionalmente reducen el ruido.
- Reducción de la dimensionalidad funcional.

Por las características de los aparatos de captura de datos, es razonable considerar que los datos tienen ruido. Esto puede ser observado si se mira en detalle las curvas de los espectros originales.

PLS al seleccionar componentes, reduce ruido y además reduce cantidad de variables.

Como alternativa para reducir dimensionalidad utilizaremos la transformación discreta wavelet (DWT).

Para mejorar los resultados de los métodos Kernel, realizaremos reducción del ruido aplicando técnicas de suavizado que se detallan más adelante.

9.7.2. Evaluación y Selección de Modelos

A continuación detallamos las metodología para evaluar y seleccionar los modelos. Buscamos un correcto ajuste de la función predictora, pero sin caer en el sobreajuste.

Selección de modelo sobre datos de entrenamiento Sobre la muestra de entrenamiento evaluaremos los distintos modelos, a través de la validación cruzada de 10 pasos[12]. Usaremos el indicador MSE CV10 , es decir el error

[12]La validación cruzada de 10 pasos es un método para evaluar el rendimiento de un modelo dividiendo aleatoriamente un conjunto de datos en 10 partes y utilizando cada parte para entrenamiento y prueba. El proceso se repite 10 veces y se calcula la precisión promedio.

cuadrático medio que resulta del promedio de las 10 evaluaciones realizadas[13]. El modelo que tenga el menor indicador es seleccionado para pasar a la siguiente etapa de evaluación, seleccionándose entonces los parámetros del modelo (Sheather 2009).

Modelo	Parámetros a determinar
PLSR	cantidad de componentes
SVM R	parámetro de regularización C, parámetro del kernel γ
KRLS	parámetro de regularización λ, parámetro del kernel σ

Tabla 9.1: Parámetros a determinar en el modelo.

Evaluación sobre datos de entrenamiento Una vez definido el modelo en la etapa anterior, se evalúa el desempeño del mismo calculando el MSE sobre el total de la muestra de entrenamiento y el R^2 entre los valores predichos y los medidos.

Procedimiento de evaluación:

- Se utiliza el modelo seleccionado para predecir los valores correspondientes para todos los elementos de la muestra de entrenamiento.
- Se ajusta un modelo lineal entre los valores predichos y los medidos.
- Se calculan MSE y R^2.

Evaluación sobre datos de Test Se procede en forma similar que en el modelo evaluado sobre datos de entrenamiento. Se realizan las predicciones con los datos de Test.

[13]para KRLS usaremos LOOE

9.8. Experimentos realizados

A continuación se procede a evaluar en la práctica el poder predictivo de los distintos modelos estudiados.

- Extracción supervisada de características con PLSR (Duda *et al.* 2000).
- Regresión SVM de los datos originales del espectro.
- Regresión SVM de coeficientes wavelet de espectros originales.
- KRLS de coeficientes wavelet de espectros originales (Hainmueller & Hazlett 2013a).
- Regresión SVM de coeficientes wavelet de espectros suavizados.

9.8.1. Extracción supervisada de características con PLSR

Es un método supervisado pues, en forma similar a LDA (utilizado en problemas de clasificación (Duda *et al.* 2000)), para la generar los componentes se utilizan las variables de respuesta. No solo se proyectan los predictores (como en PCA) sino que también las variables de respuesta.

Extracción de características La muestra es proyectada en ese nuevo espacio, obteniendo una nueva muestra con la misma cantidad de características. Podemos ordenar las nuevas características según su puntaje.

El procedimiento consiste en evaluar modelos lineales con mínimos cuadrados ordinarios para distintos conjuntos de componentes. Los conjuntos de componentes son seleccionados en forma incremental uno a uno según su puntaje descendente.

Encontramos que el menor error se encuentra para el conjunto de las 12 primeras características. Este error es evaluado según el criterio 10 fold cross validation.

En la figura 9.8 observamos que se minimiza el error para el modelo de 12 componentes.

Seleccionamos entonces el modelo de doce componentes.

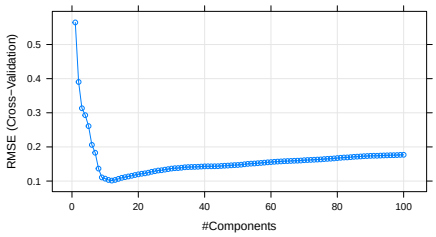

Figura 9.8: El error se minimiza al utilizar doce componentes, MSECVentr = 0,010201.

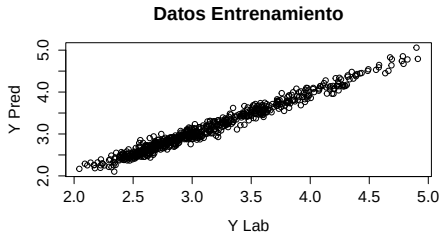

Figura 9.9: Datos entrenamiento, $MSE = 0,0086$ $R^2 = 0,9748$.

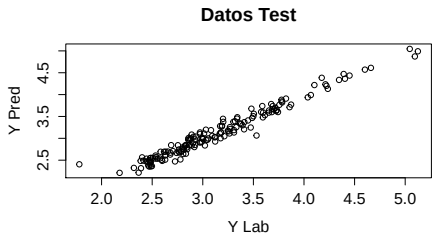

Figura 9.10: Datos test, $MSE = 0,0137$ $R^2 = 0,9624$.

Procedemos a realizar una predicción con el modelo seleccionado de doce componentes con los datos de entrenamiento. Evaluamos la predicción realizando la regresión lineal simple entre los valores predichos y los medidos. El error de mínimos cuadrados es de 0.0086. Es levemente menor, pues el error anterior de la selección era resultado del promedio de los 10 modelos evaluados. El R^2 ajustado obtenido es de 0.9748 que ilustramos en el gráfico de la figura 9.9 donde se grafican el valor de respuesta observado contra el predicho por el modelo.

Realizamos la predicción con el modelo seleccionado para los datos de test. Evaluamos en forma similar. Obtenemos un error levemente mayor, con un $R^2 = 0.9624$ que nos permite estimar la capacidad predictiva del modelo (no hubo sobreajuste) que se observa en el gráfico 9.10 para los datos de test.

El modelo PLS funciona en forma satisfactoria superando el 0.92 del objetivo. Esto se debe a que, según la bibliografía, el PLS no funciona tan bien con muestras pequeñas (50 elementos) pero si con una muestra grande como es nuestro caso (más de 600 elementos).

9.8.2. Regresión SVM de los datos originales del espectro

Procedemos a ajustar un modelo SVM utilizando un kernel Gaussiano.

C es un parámetro de regularización que controla el equilibrio entre el error de entrenamiento y la complejidad del modelo.

La función kernel gausiana está definida por:

$$K(x, x') = \exp(-\gamma \|x - x'\|^2)$$

Por lo tanto debemos elegir el parámetro de regularización C y el γ del kernel, que minimicen el error cuadrático medio.

Para seleccionar los parámetros del modelo realizamos una búsqueda en una grilla de valores. Primero con una grilla que abarca un rango grande, para luego ir reduciendo el tamaño de la misma.

Primero utilizamos, luego de varias aproximaciones, una grilla evaluada entre valores de C entre 140 y 25 y γ entre $0,05$ y $0,08$ obteniendo con los datos de entrenamiento un MSECV $= 0,2166$

Evaluamos una grilla mas ajustada que nos permite seleccionar los parámetros $\gamma = 0,062$ y $C = 217$. Obteniendo un MSE de validación cruzada de $0,2092$, mejorando levemente este valor al hacer más precisa la grilla.

Realizamos predicciónn con los datos de entrenamiento. Para evaluar el modelohacemos una regresión de los valores predichos contra los valores medidos obteniendo un $R^2 = 0,5652$.

Los resultados están lejos de ser satisfactorios.

9.8.3. Regresión SVM de coeficientes wavelet de espectros originales

Procedemos a reducir la dimensionalidad de los datos del espectro.

Las técnicas con mejores resultados en este campo son los coeficientes wavelet. En particular, utilizamos un wavelet HAAR nivel 5 cuya función y algoritmo de calculo observamos graficados en las figuras 9.11 y 9.12 respectivamente (Chuma *et al.* 2017).

La transformada wavelet discreta (DWT) es un tipo muy popular de la transformada wavelet para el procesamiento de señales y el procesamiento y compresión de imágenes, incluso forma parte del estándar JPEG2000 (Group).

La DWT se calcula utilizando el algoritmo que se muestra en la Figura 9.12. En cada etapa de descomposición, la DWT produce coeficientes de wavelet que

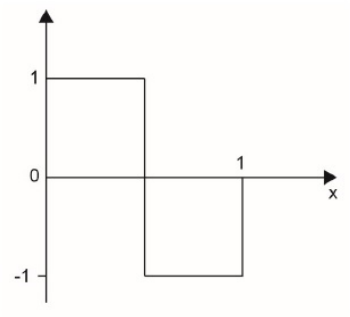

Figura 9.11: Gráfico de la función wavelet tomado de (Chuma *et al.* 2017).

corresponden respectivamente a la mitad superior e inferior del espectro de la señal de entrada. Este algoritmo fue desarrollado por Mallat (Mallat 1989) y es conocida como Transformada Wavelet Ortogonal Rápida.

Utilizamos entonces, como variables predictoras a los 69 coeficientes obtenidos.

Selección del modelo Buscamos en una grilla de valores, la que vamos afinando paso a paso. En la última grilla evaluada obtenemos los siguientes resultados que observamos en la grilla ilustrada el gráfico de la figura 9.13.

El menor valor del error obtenido es para los parámtros $\gamma = 0.003$ y $C = 5$. El error calculado es mucho menor que cuando tratamos los datos de los espectros originales.

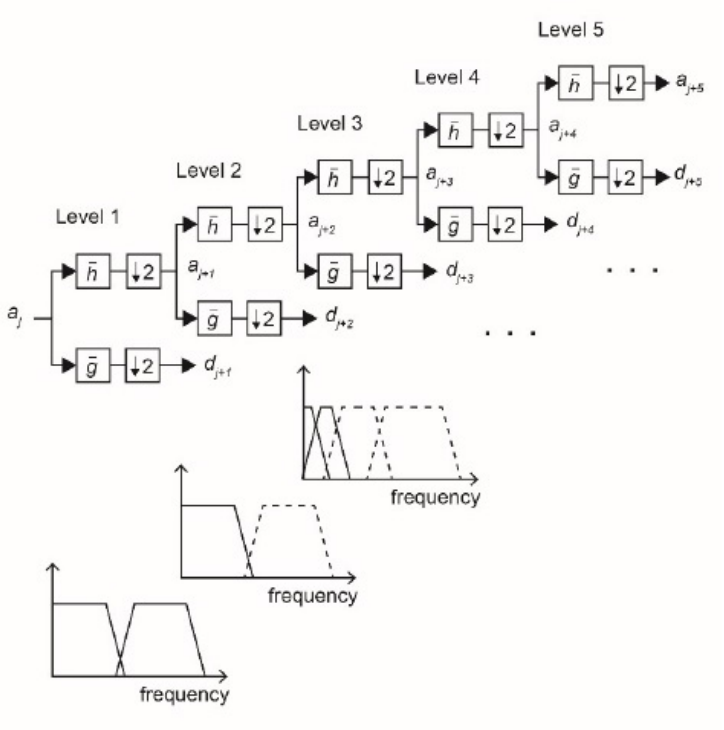

Figura 9.12: El algoritmo para calcular la DWT de cinco niveles tomado de (Chuma *et al.* 2017).

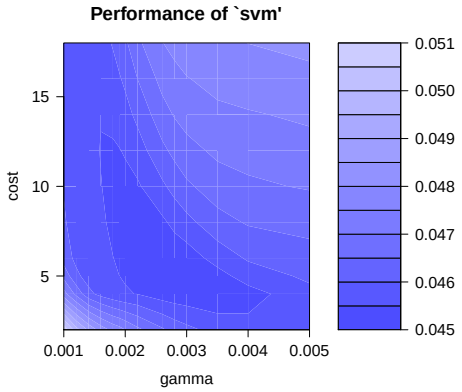

Figura 9.13: Gridsearch: parámetros $\gamma = 0.003$ y $C = 5$, $\text{MSECV}_{\text{entr}} = 0{,}0448$.

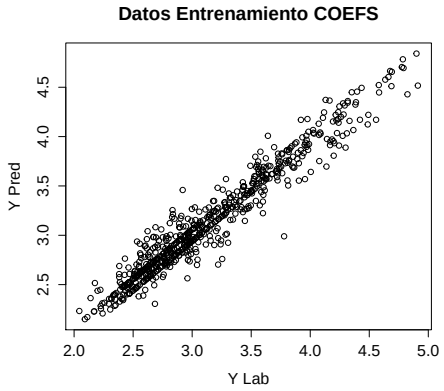

Figura 9.14: Datos Entrenamiento $MSE = 0,0206$ $R^2 = 0,9408$.

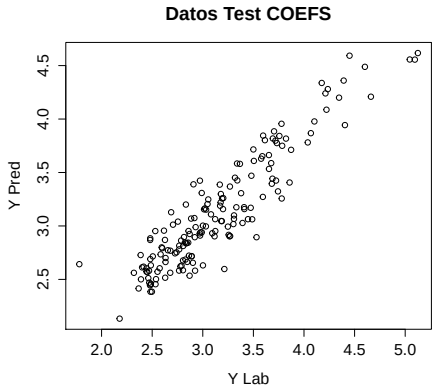

Figura 9.15: Datos Test $MSE = 0,0553$ $R^2 = 0,8484$.

Evaluación del modelo con los datos de entrenamiento (figura 9.14).

Evaluación del modelo con los datos de test (figura 9.15).

Se ha logrado obtener un modelo con mejores resultados que con los datos de los espectros originales. No alcanzamos el R^2 objetivo con los datos de test (0.84), pero si con los datos de entrenamiento (0.94).

9.8.4. KRLS de coeficientes wavelet de espectros originales

Hainmuller y Hazlett del Departamento de Ciencias Políticas del M.I.T. proponen en (Hainmueller & Hazlett 2013a) la utilización de regresion con minimos cuadrados regularizados kernel en las ciencias sociales. Los autores han desarrollado un paquete de R: KRLS (Hainmueller & Hazlett 2013b).

Este paquete nos permite, para un σ dado obtener el λ óptimo correspondiente.

Por lo tanto recorreremos una grilla de valores de σ.

Para cada modelo resultante de una combinación de σ y su λ óptimo evaluamos, en forma similar a lo ya realizado con SVM, en la muestra de entrenamiento y test.

El R^2 para los datos de entrenamiento se maximiza para $\sigma = 69$

En nuestro código para KRLS incluímos en la iteración el cálculo de R^2 para los datos de entrenamiento y de test.

A modo de ejemplo mostramos una parte de la tabla de datos obtenidos:

	sigmas	lambdas	R2func	R2entr	R2test
[17,]	69.0	0.11727456	0.9901312	0.9912999	0.8711333*
[18,]	70.0	0.11716759	0.9898935	0.9910678	0.8712389
[19,]	71.0	0.11706063	0.9896534	0.9908333	0.8713369
[20,]	72.0	0.16694832	0.9835318	0.9853243	0.8708421
[21,]	73.0	0.16684136	0.9832031	0.9849975	0.8709383

Tabla 9.2: Subconjunto de los datos obtenidos optimizando el modelo SVM.

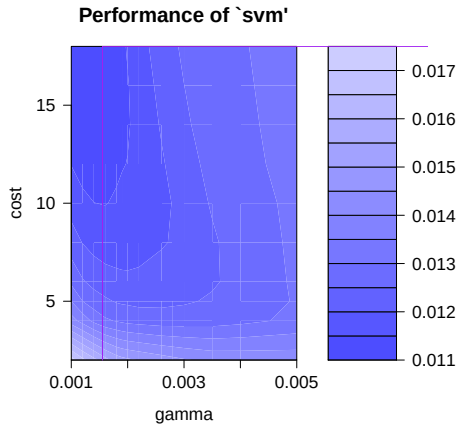

Figura 9.16: Gridsearch: parámetros $\gamma = 0.003$ y $C = 5$, $\text{MSECV}_{\text{entr}} = 0{,}0448$.

Los resultados son mejores que los de SVM según observamos en el cuadro 9.2. Alcanzamos el objetivo para los datos de entrenamiento (0.99), pero no para los datos de test (0.87).

9.8.5. Regresión SVM de coeficientes wavelet de espectros suavizados

Dado que SVM es sensible al ruido, procedemos a utilizar distintas técnicas del extracción del ruido.

En este caso, aplicamos wavelet denoising a los datos del espectro, con un wavelet DB8 (Daubechies de vanishing-monent 8).

Reprocesamos luego los espectros suavizados, disminuyendo la dimensionalidad a través de los coeficientes wavelet en forma similar a lo realizado con los espectros originales.

Proseguimos luego con el procedimiento similar al realizado para los espectros originales.

Busqueda en grilla de valores graficada en la figura 9.16.

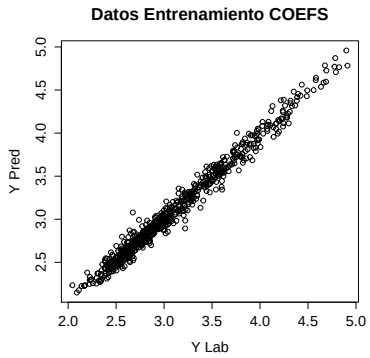

Figura 9.17: Datos Entrenamiento $MSE = 0,006$ $R^2 = 0,9805$.

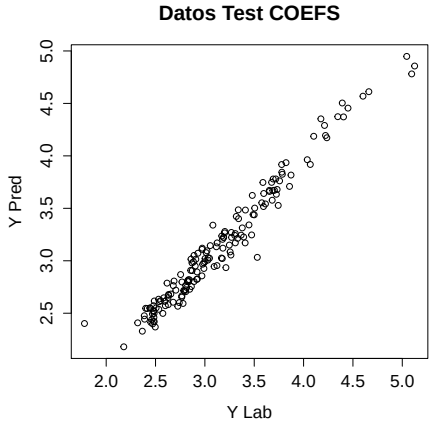

Figura 9.18: Datos Test $MSE = 0,0141$ $R^2 = 0,9612$.

Evaluación del modelo con los datos de entrenamiento graficada en la figura 9.17.

Evaluación del modelo con los datos de test graficada en la figura 9.18.

Al eliminar el ruido, registramos resultados notoriamente mejores. Logramos superar el R^2 objetivo de $0,92$ tanto para los datos de entrenamiento $(0,98)$ como para los datos de test $(0,96)$.

9.9. Resumen de los resultados obtenidos

- Se superó ampliamente el objetivo de 0.92 planteado por los expertos de campo
- Se logró mejorar los resultados de las técnicas SVM a través de la eliminación de ruido y reducción de dimensionalidad funcional.
- Aunque la muestra es grande, lo que favorece a PLSR, se obtuvieron resultados similares con SVM. Para muestras futuras más pequeñas hemos logrado una técnica con resultados prometedores.
- Si bien los tiempos de ajuste del modelo son largos (media hora en el peor de los casos), los tiempos de predicción son casi instantáneos.

9.9.1. Resultados obtenidos

Cuadro de resumen resultados obtenidos:

Modelo	MSE CV	Entr MSE	Entr R^2	Test MSE	Test R^2
PLS Espectro Original, 12 componentes	0.0102	0.0086	0.9748	0.0137	0.9624
SVM Espectro Coeficientes DWT	0.0448	0.0206	0.9408	0.0553	0.8484
KRLS Espectro Coeficientes DWT	N/A	N/A	0.9908	N/A	0.8713
SVM Espectro Suavizado Coeficientes DWT	0.011	0.006	0.9805	0.0141	0.9612

Tabla 9.3: Resumen de resultado para regresión con coeficientes wavelet.

Salvo para el tratamiento directo con SVM de los datos originales del espectro, en todos los casos logramos superar el R^2 objetivo para los datos de entrenamiento.

Para los datos de test, superamos el objetivo con SVM aplicado a los coeficientes de los espectros suavizados (0.9612).

9.9.2. Cumplimiento del objetivo cuantitativo

Se logró cumplir en forma holgada con el objetivo planteado. Para medir los resultados utilizamos el R^2 obtenido en el set de datos de test, si bien en la publicación de referencia se usan solamente datos de entrenamiento.

Modelo	R^2
Valor Objetivo planteado por Expertos de Campo	0.9200
Valor Obtenido con PLS en publicación (Zhang *et al.* 2008)	0.9411
Valor Obtenido con SVM en publicación (Zhang *et al.* 2008)	0.9724
PLS Espectro Original, 12 componentes	0.9624
SVM Espectro Suavizado Coeficientes DWT	0.9612

Tabla 9.4: Cumplimiento del objetivo cuantitativo.

Observación. Cabe destacar que la contrastación de la métrica R^2 frente a referencias externas es metodológicamente válida, pese a evaluarse sobre conjuntos de datos distintos. Esta viabilidad se fundamenta en la profunda equivalencia estructural y fisicoquímica de las matrices espectroscópicas analizadas (NIRS). Lejos de buscar una correspondencia numérica exacta, la comparación demuestra que la capacidad de generalización y la estabilidad del algoritmo alcanzan estándares congruentes con la literatura de vanguardia en quimiometría. Para una justificación detallada sobre la homogeneidad del dominio de estos datos y su respaldo bibliográfico, véase el Apéndice E.

9.9.3. Mejora de resultados eliminando ruido y reduciendo dimensionalidad

Utilizamos con éxito la reducción de la dimensionalidad de los datos funcionales a través de la utilización de los coeficientes wavelet. Esos nos permitió ajustar una función más efectiva con SVM.

Procedimos a utilizar técnicas de eliminación de ruido que mejoraron aún más los resultados, permitiéndonos superar el R^2 objetivo y aproximarnos al R^2 publicado en (Zhang *et al.* 2008).

9.9.4. Regresión PLS y SVM

En las publicaciones relevadas se utiliza como modelo de referencia el PLS. PLS no tiene los mejores resultados, dado que fracasa en obtener una buena función predictora cuando las muestras son peqeñas ($n = 50$).

En nuestro caso, para una muestra muy grande de mas de 600 en la muestra de entrenamiento, PLS funciona con resultados razonables tratando directamente los datos originales del espectro. Esto se produce como ya dijimos por

el tamaño de la muestra, así como por la selección de características, método con el cual de hecho estamos eliminando ruido implicitamente.

SVM alcanza resultados competitivos con PLS recién al eliminar el ruido de la muestra y reducir las dimensiones utilizando los coeficientes wavelet.

Los buenos resultados de SVM sobre datos suavizados y coeficientes DWT, nos permiten disponer un método que potencialmente puede dar resultados satisfactorios cuando se traten muestras pequeñas.

En un escenario de recalibración del sistema, dónde el costo de cada análisis químico tradicional de una muestra es muy alto, será crucial poder disponer de esta técnica.

9.9.5. Tiempos de cómputo

Las distintas técnicas implicaron la utilización de un alto poder de cómputo, registrándose tiempos largos de procesamiento.

Esta situación no genera problemas en la utilización de estos modelos, dado que el cómputo efectivo de la función predictora aplicado a una nueva muestra son muy pequeños.

9.9.6. Robustez de SVM

Si bien encontramos outliers en los datos, no fue necesario descartar estos datos para cumplir con el objetivo. Nos planteamos en un futuro inmediato, reprocesar los modelos obtenidos descartando los outliers y verificar con los expertos los registros de las muestras etiquetadas como valores atípicos.

Capítulo 10

Conclusiones Finales y Trabajo Futuro

El aprendizaje automático, y en particular la Teoría del Aprendizaje Estadístico, constituyen un marco de trabajo sólido para la comprensión y construcción de sistemas con la capacidad de aprender a partir de datos. El propósito primordial es la inferencia de reglas generales a partir de las muestras observadas, con la capacidad de extrapolar estas reglas a datos no previamente analizados. En este contexto, los métodos basados en kernels ofrecen un enfoque potente y adaptable para la construcción de sistemas de aprendizaje no lineal, al integrar de manera implícita los datos en un espacio de características de alta dimensión.

10.1. Conclusiones Generales

- El aprendizaje automático, especialmente la teoría del aprendizaje estadístico, proporciona un marco sólido para comprender y construir sistemas que pueden aprender de los datos. El objetivo es inferir reglas generales a partir de muestras observadas y generalizar sobre datos no vistos. Esto es especialmente útil para problemas que son demasiado complejos para ser resueltos con soluciones explícitas preprogramadas.

- La elección del espacio de hipótesis, la función de pérdida y el método de regularización son decisiones de diseño cruciales en la construcción de un sistema de aprendizaje exitoso. El espacio de hipótesis determina el conjunto de funciones posibles que el sistema puede aprender, la función de pérdida cuantifica el error de una función dada y el método de regular-

ización ayuda a prevenir el sobreajuste penalizando funciones demasiado complejas.

- Los métodos de kernel, que incrustan implícitamente datos en un espacio de características de dimensión alta, ofrecen una forma poderosa y flexible de construir sistemas de aprendizaje no lineal. Este enfoque aprovecha la teoría y los algoritmos bien establecidos para encontrar patrones lineales, mientras que la función de kernel proporciona una forma computacionalmente eficiente de trabajar con representaciones de datos de alta dimensión.

10.2. Conclusiones de la Aplicación Práctica

- Las técnicas de aprendizaje automático, particularmente los métodos basados en kernel como la Regresión de Vectores de Soporte (SVM) y los Mínimos Cuadrados Regularizados de Kernel (KRLS), se pueden aplicar eficazmente al análisis de datos quimiométricos. Proporcionan una alternativa a los análisis químicos tradicionales, que consumen mucho tiempo y son costosos, como la predicción del contenido de nitrógeno en las hojas de tabaco a partir de datos de espectroscopia de infrarrojo cercano (NIRS).

- El preprocesamiento de los datos espectrales, incluidas las técnicas de reducción de ruido y de dimensionalidad, es esencial para mejorar el rendimiento de los modelos SVM. Por ejemplo, suavizar los datos espectrales y aplicar la Transformada de Wavelet Discreta (DWT) pueden mejorar significativamente la precisión predictiva de los modelos SVM.

- Si bien tanto PLSR como SVM lograron una buena precisión predictiva en el conjunto de datos dado, SVM, especialmente cuando se aplica a datos preprocesados, es prometedor para escenarios con tamaños de muestra más pequeños. Esto convierte a SVM en una herramienta potencialmente valiosa para recalibrar sistemas donde obtener análisis químicos adicionales para el entrenamiento es costoso.

Las fuentes enfatizan la transición de los enfoques tradicionales de Minimización de Riesgo Empírico (ERM) a técnicas de regularización más sofisticadas como SVM y KRLS para abordar el sobreajuste y mejorar el rendimiento

de la generalización. Destacan la importancia de comprender los fundamentos teóricos de estos métodos para aplicarlos eficazmente a problemas del mundo real como el análisis quimiométrico. dimensión, ofrecen una forma poderosa y flexible de construir sistemas de aprendizaje no lineal. Este enfoque aprovecha la teoría y los algoritmos bien establecidos para encontrar patrones lineales, mientras que la función de kernel proporciona una forma computacionalmente eficiente de trabajar con representaciones de datos de alta dimensión.

10.3. Trabajo futuro

Tareas siguientes a la culminación de este trabajo:

Luego de la etapa que intentamos cumplir con este trabajo, continuaremos trabajando sobre el problema en cuestión. Algunas de las tareas que se considera relevante abordar a corto plazo son:

- Simular una muestra pequeña a partir de la muestra de entrenamiento.
- Generar modelos SVMr y KRLS independientes para predecir valores de otras 3 substancias disponibles en el conjunto de datos.
- Probar otras técnicas de suavizado/eliminación de ruido.
- Aplicar KRLS a datos suavizados: en datos no suavizados supera el R^2 de SVM.
- Introducir outliers artificiales y reprocesar.
- Extraer outliers y reprocesar.

El objetivo de este trabajo es proporcionar un sistema de cálculo automático, obteniendo un modelo ajustado y su correspondiente función predictora que cumpla con los requisitos señalados. Una tarea futura podría ser generar un paquete de R para que la comunidad científica disponga de las herramientas utilizadas.

Lineas de investigación tentativas:

- Generar un modelo que busque la función $f : \mathbb{R}^{2201} \to \mathbb{R}^4$ o para para el caso de los coeficientes wavelet $f : \mathbb{R}^{69} \to \mathbb{R}^4$ que permita predecir los valores de las cuatro substancias simultáneamente (Lue 2009). Esta tarea está en el contexto de regresión aplicada a funciones de salida vectorial (Micchelli & Pontil 2005).

- Es decir, generar un modelo de regresión que prediga los valores químicos de los cuatro substancias. Es decir $\mathcal{Y} \in \mathbb{R}^4$.

Apéndice A

Código R: Clasificación Lineal con SVM

En este apartado se documenta en R utilizable para el procedimiento de clasificación lineal mediante Máquinas de Vectores de Soporte (SVM).

svm_clasif_lineal_a_v01.r

```
n <- 150 # number of data points
p <- 2 # dimension
sigma <- 1 # variance of the distribution
meanpos <- 0 # centre of the distribution of positive examples
meanneg <- 3 # centre of the distribution of negative examples
npos <- round(n/2) # number of positive examples
nneg <- n-npos # number of negative examples
# Generate the positive and negative examples
xpos <- matrix(rnorm(npos*p,mean=meanpos,sd=sigma),npos,p)
xneg <- matrix(rnorm(nneg*p,mean=meanneg,sd=sigma),npos,p)
x <- rbind(xpos,xneg)
# Generate the labels
y <- matrix(c(rep(1,npos),rep(-1,nneg)))
# Visualize the data
plot(x,col=ifelse(y>0,1,2))
legend("topleft",c('Positive','Negative'),col=seq(2),pch=1,text
   .col=seq(2))

## Prepare a training and a test set ##
ntrain <- round(n*0.8) # number of training examples
tindex <- sample(n,ntrain) # indices of training samples
xtrain <- x[tindex,]
xtest <- x[-tindex,]
ytrain <- y[tindex]
```

```
ytest <- y[-tindex]
istrain=rep(0,n)
istrain[tindex]=1
# Visualize
plot(x,col=ifelse(y>0,1,2),pch=ifelse(istrain==1,1,2))
legend("topleft",c('Positive␣Train','Positive␣Test','Negative␣
   Train','Negative␣Test'),
       col=c(1,1,2,2),pch=c(1,2,1,2),text.col=c(1,1,2,2))
#1.2 Train a SVM
#Now we train a linear SVM with parameter C=100 on the training
    set
# load the kernlab package
library(kernlab)
# train the SVM
svp <- ksvm(xtrain,ytrain,type="C-svc",kernel='vanilladot',C
   =100,scaled=c())
#Look and understand what svp contains
# General summary
svp
# Attributes that you can access
attributes(svp)
# For example, the support vectors
alpha(svp)
alphaindex(svp)
b(svp)
# Use the built-in function to pretty-plot the classifier
plot(svp,data=xtrain)
```

Apéndice B

Código R: Clasificación No Lineal con SVM

En este apartado se detalla un script en R, a modo de ejemplo, para la evaluación y selección del modelo de clasificación no lineal con Máquinas de Vectores de Soporte.

svm_clasif_nolineal_a_v01.r

```
#clasificacion no lineal
https://stat.ethz.ch/pipermail/r-help/2000-April/006140.html

n <- 150 # number of data points
p <- 2 # dimension
sigma <- 1 # variance of the distribution
meanpos <- 0 # centre of the distribution of positive examples
meanneg <- 3 # centre of the distribution of negative examples
npos <- round(n/2) # number of positive examples
nneg <- n-npos # number of negative examples
# Generate the positive and negative examples
xpos <- matrix(rnorm(npos*p,mean=meanpos,sd=sigma),npos,p)
xneg <- matrix(rnorm(nneg*p,mean=meanneg,sd=sigma),npos,p)
x <- rbind(xpos,xneg)
# Generate the labels
y <- matrix(c(rep(1,npos),rep(-1,nneg)))
# Visualize the data
plot(x,col=ifelse(y>0,1,2))
legend("topleft",c('Positive','Negative'),col=seq(2),pch=1,text
   .col=seq(2))

## Prepare a training and a test set ##
ntrain <- round(n*0.8) # number of training examples
```

```
tindex <- sample(n,ntrain) # indices of training samples
xtrain <- x[tindex,]
xtest <- x[-tindex,]
ytrain <- y[tindex]

ytest <- y[-tindex]
istrain=rep(0,n)
istrain[tindex]=1
# Visualize
plot(x,col=ifelse(y>0,1,2),pch=ifelse(istrain==1,1,2))
legend("topleft",c('Positive Train','Positive Test','Negative
   Train','Negative Test'),
       col=c(1,1,2,2),pch=c(1,2,1,2),text.col=c(1,1,2,2))
#1.2 Train a SVM
#Now we train a linear SVM with parameter C=100 on the training
    set
# load the kernlab package
library(kernlab)
# train the SVM
svp <- ksvm(xtrain,ytrain,type="C-svc",kernel='vanilladot',C
   =100,scaled=c())
#Look and understand what svp contains
# General summary
svp
# Attributes that you can access
attributes(svp)
# For example, the support vectors
alpha(svp)
alphaindex(svp)
b(svp)
# Use the built-in function to pretty-plot the classifier
plot(svp,data=xtrain)
```

Apéndice C

Pseudocódigo de la transformación DWT

algoritmo$_{dwt_5niveles.m}$

```
% --- PROCEDIMIENTO: Transformada Wavelet Discreta (DWT) de 5
    niveles ---

% 1. Carga de datos de la senal (ej. espectro NIR)
% Supongamos que X_original es nuestra matriz de datos
senal = X_original(1, :);
% Tomamos un espectro (fila) de ejemplo

% 2. Definicion de parametros de la transformada
niveles = 5;
% Numero de niveles de descomposicion
familia = 'haar';      % Familia wavelet a utilizar (Haar)

% 3. Aplicacion de la DWT unidimensional
% C: vector que contiene todos los coeficientes concatenados
% L: vector con las longitudes de cada componente (A5, D5, D4,
    D3, D2, D1)
[C, L] = wavedec(senal, niveles, familia);
% 4. Extraccion de los coeficientes de Aproximacion del nivel 5
    (A5)
% Estos coeficientes representan la senal con su
    dimensionalidad reducida
coef_A5 = appcoef(C, L, familia, niveles);
% 5. (Opcional) Extraccion de los coeficientes de Detalle (D1 a
    D5)
% coef_D1 = detcoef(C, L, 1);
% ...
```

```
% coef_D5 = detcoef(C, L, 5);

% 6. Seleccion final de caracteristicas
% El vector reducido se pasa al modelo de clasificacion SVM
caracteristicas_svm = coef_A5;
% Comprobacion de la reduccion de dimensionalidad
disp('Dimension original de la senal:');
disp(length(senal));
disp('Dimension reducida (Coeficientes A5):');
disp(length(caracteristicas_svm));
```

Apéndice D

Ejemplo Práctico: Mínimos Cuadrados Regularizados por Núcleos (KRLS)

En el marco del Aprendizaje Estadístico Supervisado basado en Núcleos abordado en este trabajo, los Mínimos Cuadrados Regularizados por Núcleos (KRLS, por sus siglas en inglés) representan una técnica fundamental, especialmente útil para problemas de regresión y para la inferencia de efectos marginales. A diferencia de las Máquinas de Vectores de Soporte (SVM) clásicas orientadas a la clasificación mediante la maximización del margen, KRLS Hainmueller & Hazlett (2013a) minimiza una función de pérdida cuadrática penalizada (regularización de Tikhonov). Esto permite obtener superficies de respuesta suaves sin necesidad de especificar analíticamente la forma funcional de las relaciones no lineales. Sin embargo, cabe mencionar que en la aplicación de este modelo dentro de la presente monografía no se obtuvieron los resultados positivos esperados. Tal como se expone en las líneas de trabajo futuro, queda pendiente un análisis más profundo y el ajuste de sus parámetros para intentar mejorar su nivel de desempeño y precisión. A fin de operacionalizar este enfoque, se expone seguidamente un *script* en el entorno y lenguaje computacional R R Core Team (2014) basado en el paquete `KRLS` Hainmueller & Hazlett (2013b). Este bloque se adjunta con el objetivo de implementar una aplicación empírica del estimador. La sintaxis aborda cuatro escenarios de ajuste que ponen de manifiesto la versatilidad del «truco del núcleo» (*kernel trick*):

- **Ejemplo Lineal:** Demuestra cómo el método ajusta una relación subyacente que es estrictamente lineal, validando que la técnica generaliza bien

incluso en casos simples.

- **Ejemplo No Lineal:** Ilustra la capacidad de capturar relaciones polinómicas complejas de forma automática, sin indicarle al modelo qué variables transformar.
- **Ejemplo 2D:** Muestra la aproximación de una función trigonométrica con ruido unidimensional. Además, destaca una de las mayores ventajas de KRLS: la estimación analítica de las derivadas parciales (efectos marginales continuos).
- **Ejemplo 3D:** Ajuste de una superficie tridimensional compleja a partir de predictores bivariados.

ejemplos$_k rls.r$

```
library(KRLS)

# === 1. EJEMPLO LINEAL ===
# Configuracion de los datos
N <- 200
x1 <- rnorm(N)
x2 <- rbinom(N, size=1, prob=.2)
y <- x1 + .5*x2 + rnorm(N, 0, .15)
X <- cbind(x1, x2)

# Ajuste del modelo KRLS
krlsout <- krls(X=X, y=y)
# Resumen de efectos marginales y contribucion de cada variable
summary(krlsout)
# Graficos de efectos marginales y esperanzas condicionales
plot(krlsout)

# === 2. EJEMPLO NO LINEAL ===
# Configuracion de los datos (relacion cubica para x1)
N <- 200
x1 <- rnorm(N)
x2 <- rbinom(N, size=1, prob=.2)
y <- x1^3 + .5*x2 + rnorm(N, 0, .15)
X <- cbind(x1, x2)

# Ajuste del modelo
krlsout <- krls(X=X, y=y)
summary(krlsout)
```

```
plot(krlsout)

# === 3. EJEMPLO 2D (Aproximacion de funciones y derivadas) ===
# Datos predictores
X <- matrix(seq(-3, 3, .1))
# Funcion real subyacente
Ytrue <- sin(X)
# Adicion de ruido aleatorio
Y <- sin(X) + rnorm(length(X), sd=.3)

# Aproximacion de la funcion usando KRLS
out <- krls(y=Y, X=X)
# Obtencion de valores ajustados y errores estandar (se)
fit <- predict(out, newdata=X, se.fit=TRUE)

# Visualizacion de resultados: Ajuste de f(x)
par(mfrow=c(2,1))
plot(y=Ytrue, x=X, type="l", col="red", ylim=c(-1.2,1.2), lwd
   =2, main="f(x)")
points(y=fit$fit, X, col="blue", pch=19)
arrows(y1=fit$fit+1.96*fit$se.fit, y0=fit$fit-1.96*fit$se.fit,
       x1=X, x0=X, col="blue", length=0)
legend("bottomright", legend=c("true␣f(x)=sin(x)", "KRLS␣fitted
   ␣f(x)"),
       lty=c(1,NA), pch=c(NA,19), lwd=c(2,NA), col=c("red","
          blue"), cex=.8)

# Visualizacion de derivadas df(x)/dx
plot(y=cos(X), x=X, type="l", col="red", ylim=c(-1.2,1.2), lwd
   =2, main="df(x)/dx")
points(y=out$derivatives, X, col="blue", pch=19)
legend("bottomright", legend=c("true␣df(x)/dx=cos(x)", "KRLS␣
   fitted␣df(x)/dx"),
       lty=c(1,NA), pch=c(NA,19), lwd=c(2,NA), col=c("red","
          blue"), cex=.8)

# === 4. EJEMPLO 3D (Superficies de respuesta complejas) ===
# Grafico de la funcion verdadera
par(mfrow=c(1,2))
f <- function(x1, x2){ sin(x1)*cos(x2) }
x1 <- x2 <- seq(0, 2*pi, .2)
z <- outer(x1, x2, f)
persp(x1, x2, z, theta=30, main="true␣f(x1,x2)=sin(x1)cos(x2)")
```

```
# Aproximacion de la superficie con KRLS
# Datos y resultados (incorporando ruido)
X <- cbind(sample(x1, 200, replace=TRUE), sample(x2, 200,
   replace=TRUE))
y <- f(X[,1], X[,2]) + runif(nrow(X))

# Ajuste de la superficie
krlsout <- krls(X=X, y=y)

# Grafico de la superficie ajustada
ff <- function(x1i, x2i, krlsout) {
    predict(object=krlsout, newdata=cbind(x1i, x2i))$fit
}
z <- outer(x1, x2, ff, krlsout=krlsout)
persp(x1, x2, z, theta=30, main="KRLS␣fitted␣f(x1,x2)")
```

Apéndice E

Consideraciones sobre la Comparativa de Performance mediante R^2

En el presente trabajo, la evaluación de la capacidad predictiva de los modelos basados en núcleos se ha realizado utilizando el coeficiente de determinación (R^2), cumpliendo con los estándares de validación requeridos por el cliente. Es fundamental precisar que, si bien los valores comparativos presentados en el la Tabla 9.4 provienen de la aplicación de algoritmos sobre conjuntos de datos distintos a los de las referencias externas, existe una equivalencia metodológica y estructural que justifica dicha comparación.

E.1. Naturaleza de los Datasets y Homogeneidad de Dominio

La validez de contrastar métricas de desempeño entre estudios independientes radica en la **similitud de la naturaleza fisicoquímica de los datos**. En ambos casos, se trabaja con matrices espectroscópicas de alta dimensionalidad obtenidas mediante NIRS (*Near-Infrared Spectroscopy*). La estructura de covarianza, la autocorrelación de las bandas de longitud de onda y la presencia de ruido instrumental en estos conjuntos de datos presentan desafíos analíticos análogos. Esto permite que el R^2 actúe como un indicador de robustez del algoritmo, permitiendo una evaluación de la eficiencia del modelo más allá de la muestra específica utilizada.

E.2. Justificación Metodológica según zhang2008qua

Este enfoque de validación y comparación de performance se sustenta en investigaciones precedentes de alta relevancia en el área de la quimiometría. Específicamente, el trabajo de zhang2008qua constituye un antecedente crítico, al demostrar que la combinación de espectroscopía infrarroja cercana y máquinas de vectores de soporte (SVM) es altamente eficaz para el análisis cuantitativo de constituyentes químicos en matrices biológicas complejas.

En su investigación, los autores establecen que el éxito de la generalización del modelo se mide por su capacidad de mantener un R^2 elevado frente a la variabilidad inherente de las muestras. Al utilizar conjuntos de datos de naturaleza similar a los analizados en esta monografía, los hallazgos de zhang2008qua validan el uso de estas métricas como un estándar de comparación para determinar la optimización de los hiperparámetros y la selección de la función de núcleo.

Por lo tanto, la comparativa presentada no pretende establecer una identidad numérica entre muestras, sino demostrar que los algoritmos aquí implementados alcanzan niveles de precisión y estabilidad consistentes con los reportados en la literatura científica de vanguardia, para el momento en que se elaboró este trabajo, para problemas de regresión espectroscópica.

Apéndice F

Teoremas y demostraciones

F.1. Teorema Generalizado de Representación

Este resultado es útil para una amplia clase de problemas de optimización, donde los regularizadores RKHS tienen soluciones que se pueden expresar como expansiones kernel en términos de los datos de entrenamiento. A continuación presentamos el planteo y la correpondiente demostración propuesta en (Scholkopf *et al.* 2001) del Teorema No paramétrico de Representación.

Enunciado del Teorema

Sean:

- $\mathcal{X}$ un conjunto no vacío.
- Un kernel positivo definido k en $\mathcal{X} \times \mathcal{X}$.
- Una muestra de entrenamiento $\{(x_1, y_1), \ldots, (x_n, y_n)\} \in (\mathcal{X} \times \mathbb{R})^m$.
- Una función g estricta y monótonamente creciente en valores reales $[0, \infty)$.
- Una fución de costo arbitraria $c : (\mathcal{X} \times \mathbb{R}^2)^m \to \mathbb{R} \cup \{\infty\}$.
- Una clase de funciones $\mathcal{F} = \left\{ f \in \mathbb{R}^{\mathcal{X}} | f(\cdot) = \sum_{i=1}^{\infty} \beta_i k(\cdot, z_i), \beta_i \in \mathbb{R}, z_i \in \mathcal{X}, ||f|| < \infty \right\}$.

$||\cdot||$ es la norma en el $RKHS\ \mathcal{H}_k$ asociada al kernel k. Es decir, para cualquier $z_i \in \mathcal{X}, \beta_i \in \mathbb{R}, i \in \mathbb{N}$,

$$||\sum \beta_i k(\cdot, z_i)||^2 = \sum_{i=1}^{\infty} \sum_{j=1}^{\infty} \beta_i \beta_j k(z_i, z_j).$$

Luego cualquier $f \in \mathcal{F}$ minimiza el funcional de riesgo regularizado

$$c((x_1, y_1, f(x_1)), \cdots, (x_m, y_m, f(x_m)))) + g(||f||),$$

admite una representación de la forma:

$$f(\cdot) = \sum_{i=1}^{m} \alpha_i k(\cdot, x_i).$$

Demostración

Dado que asumimos que k mapea en $\mathbb{R}$ consideramos:

$$\phi : X \to \mathbb{R}^{\mathcal{X}}, x \mapsto k(\cdot, x).$$

Dado que k es un reproducing kernel, evaluar la función $\phi(x)$ en el punto x' implica:

$$(\phi(x))(x') = k(x, x') = \langle \phi(x), \phi(x') \rangle,$$

para todo $x, x' \in X$ Aquí el producto interno es en H_k. Dado $x_1, ..., x_m$, cualquier $f \in \mathcal{F}$ puede ser descompuesta en una parte que vive en el span de las funciones $\phi(x_i)$ y una parte ortogonal a ellas, es decir:

$$f = \sum_{i=1}^{m} \alpha_i \phi(x_i) + v,$$

para algún $\alpha \in \mathbb{R}^m$ y $v \in F$ que satisface, para todo j,

$$\langle v, \phi(x_j) \rangle = 0.$$

Evaluar f en un punto de la muestra de entrenamiento x_j, es proyectar f sobre $\phi(x_i)$

$$f(x_j) = \langle f, \phi(x_j) \rangle = \langle \sum_{i=1}^{m} \alpha_i \phi(x_i) + v, \phi(x_i) \rangle,$$

$$= \sum_{i=1}^{m} \alpha_i \phi(x_i)\phi(x_j) + \langle v, \phi(x_j) \rangle = \sum_{i=1}^{m} \alpha_i \phi(x_i)\phi(x_j),$$

que es independiente del v elegido. Por lo tanto, el resultado de evaluar f en cualquier punto de la muestra no depende de v, el primer sumando del riesgo a minimizar $c((x_1, y_1, f(x_1)), ..., (x_m, y_m, f(x_m)))$ tampoco depende del v elegido.

Por lo tanto podemos elegir $v = 0$.

Procedamos ahora con el segundo sumando del riesgo,

$$g(\|f\|) = g\left(\|\sum_{i=1}^{m}\alpha_i\phi(x_i) + v\|\right) = g(\sqrt{\|\sum_{i=1}^{m}\alpha_i\phi(x_i)\|^2 + \|v\|^2}\Bigg).$$

Dado que g es estrictamente creciente, si $v = 0$ entonces su valor será menor. Por lo tanto un minimizador del riesgo necesariamente debe tener $v = 0$.

Luego

$$f = \sum_{i=1}^{m}\alpha_i\phi(x_i),$$

es decir

$$f(\cdot) = \sum_{i=1}^{m}\alpha_i k(\cdot, x_i).$$

F.2. Demostración de una cota para el error

Lema

Estos resultados ya los habíamos planteado en este trabajo cuando tratamos inecuaciones exponenciales. Sea ξ una variable aleatoria en un espacio de probabilidad Z con media μ y varianza σ^2 satisfaciendo $|\xi - \mu| \leq M$ casi seguramente. Entonces, para todo $\epsilon > 0$, tenemos:

Bernstein:

$$\mathbb{P}_{z\in Z^m}\left\{\left|\frac{1}{m}\sum_{i=1}^{m}\xi(z_i) - \mu\right| > \epsilon\right\} \leq \exp\left\{\frac{-m\epsilon^2}{2(\sigma^2 + \frac{1}{3}M\epsilon)}\right\}.$$

Hoeffding:

$$\mathbb{P}_{z\in Z^m}\left\{\left|\frac{1}{m}\sum_{i=1}^{m}\xi(z_i) - \mu\right| > \epsilon\right\} \leq 2\exp\left\{\frac{-m\epsilon^2}{2M^2}\right\}.$$

Enunciado del Teorema

En (Wu *et al.* 2006) se presenta una cota para una función ψ genérica, que cumpla las condiciones que siguen. Sea $\mathcal{H}$ un subconjunto de $\mathcal{C}(\mathcal{X})$ tal que para toda $f \in \mathcal{H}$ se cumple que $|f(x) - y| \leq M$ casi seguramente. Si $\psi : \mathbb{R} \to \mathbb{R}$ es

una función de pérdida regresiva que satisface:

$$|\psi(t) - \psi(t')| \leq C|t - t'|^3, \quad \forall t, t' \in [-M, M],$$

luego para todo $\epsilon > 0$ se cumple que:

$$\text{Prob}\{\mathcal{E}_{\mathcal{H}}(f_{\mathbf{z}}) \leq \epsilon\} \geq 1 - \mathcal{N}\left(\mathcal{H}, \left(\frac{\epsilon}{8C}\right)^{1/s}\right) 2\text{exp}\left\{-\frac{m\epsilon^2}{128C^2M^{2s}}\right\}.$$

Demostración del Teorema

Sea $\{f_j\}_{j=1}^N \subset H$ con $N = N(H, \eta)$ y $\eta = (\epsilon/(4C))^{1/8}$ tal que para cada $f \in H$ existe algún $j \in \{1, ..., N\}$ satisfaciendo $\|f - f_j\|_\infty < \eta$. Entonces, de acuerdo a la condición de Lipschitz dada en nuestra suposición, vemos de $|y - f(x)| \leq M$ que

$$|\mathcal{E}(f) - \mathcal{E}(f_j)| \leq \int_Z |\psi(y - f(x)) - \psi(y - f_j(x))| d\rho \leq C \int_Z |f(x) - f_j(x)|^8 d\rho \leq C\eta^8,$$

y que casi seguramente

$$|\mathcal{E}_m(f) - \mathcal{E}_m(f_j)| \leq \frac{1}{m} \sum_{i=1}^m |\psi(y_i - f(x_i)) - \psi(y_i - f_j(x_i))| \leq C\eta^8.$$

Por lo tanto

$$|\mathcal{E}(f) - \mathcal{E}_m(f)| \leq |\mathcal{E}(f) - \mathcal{E}(f_j)| + |\mathcal{E}(f_j) - \mathcal{E}_m(f_j)| + 2C\eta^8.$$

Por la elección de η, encontramos $2C\eta^8 = \epsilon/2$. Por lo tanto, el evento $\{z \in Z^m : |\mathcal{E}(f) - \mathcal{E}_m(f)| > \epsilon \text{ para algún } f \in H\}$ está contenido en el evento $\bigcup_{j=1}^N \{z \in Z^m : |\mathcal{E}(f_j) - \mathcal{E}_m(f_j)| > \epsilon/2\}$. Es entonces que:

$$\mathbb{P}_{z \in Z^m}\left(\sup_{f \in H} |\mathcal{E}(f) - \mathcal{E}_m(f)| > \epsilon\right) \leq \sum_{j=1}^N \mathbb{P}_{z \in Z^m}\left(|\mathcal{E}(f_j) - \mathcal{E}_m(f_j)| > \epsilon/2\right).$$

Para cada j, aplicamos el Lema a la variable aleatoria $\xi(z) = \psi(y - f_j(x))$ en (Z, ρ). Se satisface $\mu = \mathcal{E}(f_j)$ y $\frac{1}{m}\sum_{i=1}^m \xi(z_i) = \mathcal{E}_m(f_j)$. Además, $|\xi(z) - \mu|$

puede ser expresado como:

$$|\psi(y-f(x))-\int \psi(y-f_j(x'))d\rho(z')| = |\int \psi(y-f_j(x))-\psi(y-f_j(x'))d\rho(z')|,$$

que está acotado por $\int_Z C|y-f_j(x)-(y'-f_j(x'))|^8 d\rho(z') \leq 2CM^8$. Entonces, tenemos de la desigualdad de Hoeffding en el Lema antes enunciado que

$$\mathbb{P}_{z\in Z^m}\left(|\mathcal{E}_m(f_j)-\mathcal{E}(f_j)|>\epsilon/2\right) \leq 2\exp\left(\frac{-m(\epsilon/2)^2}{2(2CM^8)^2}\right).$$

Entonces

$$\mathbb{P}_{z\in Z^m}\left(\sup_{f\in H}|\mathcal{E}(f)-\mathcal{E}_m(f)|>\epsilon\right) \leq N(H,\eta)2\exp\left(\frac{-m\epsilon^2}{32C^2M^8}\right).$$

Dado que f y f_u están ambas en H, tenemos $\mathcal{E}(f)-\mathcal{E}(f_u) \leq 2\sup_{f\in H}|\mathcal{E}(f)-\mathcal{E}_m(f)| \leq 2\epsilon$ con la confianza anterior. Entonces llegamos a la demostración reemplazando ϵ por $\epsilon/2$. □

F.3. Demostración de una solución alternativa para el problema de clasificación

Si nuestro conjunto de datos Z es separable, es posible obtener clasificador resolviendo la ecuación que más abajo transcribimos y luego, si w^* es la solución calculada, el clasificador asigna a cada punto x la señal de $\langle w^*, x\rangle - c(w^*)$. O sea,

$$x \mapsto \text{sgn}\langle w^*, x\rangle - c(w^*)).$$

Es posible resolver una forma equivalente a esta ecuación:

$$\max_{||w||=1}\frac{1}{2}\left[\min_{y_i=1}\langle w, x_i\rangle - \max_{y_i=-1}\langle w, x_i\rangle\right].$$

En (Cucker & Zhou 2007) se plantea el siguiente Teorema que nos permite llegar al planteo de la solución alternativa. Teorema: Supongamos que el problema planteado tiene una solución w^* tal que $\Delta(w^*)\rangle 0$. Entonces, $w^* = \frac{w}{||w||}$, donde w es una solución del siguiente problema de optimización:

$$\min_{w\in\mathbb{R}^n, b\in\mathbb{R}} ||w||^2$$
$$\text{sujeto a} y_i(w^T x_i - b) \geq 1, \quad i = 1, ..., m.$$

Además, $\Delta(w^*) = \frac{1}{||w||}$ representa el margen.

Demostración

Consideremos un minimizador $(\tilde{w}, \tilde{b})$ de la función cuadrática $||w||^2$ sujeta a las restricciones lineales dadas. Recordemos que:

$$\Delta(w) = \frac{1}{2}\left(\min_{y_i=1} w^T x_i - \max_{y_i=-1} w^T x_i\right).$$

Entonces,

$$\Delta\left(\frac{\tilde{w}}{||\tilde{w}||}\right) = \frac{1}{2}\left(\min_{y_i=1} \frac{\tilde{w}^T x_i - \tilde{b}}{||\tilde{w}||} - \max_{y_i=-1} \frac{\tilde{w}^T x_i - \tilde{b}}{\|\tilde{w}\|}\right)$$
$$\geq \frac{1}{||\tilde{w}||},$$

ya que $\langle \tilde{w}, x_i\rangle - \tilde{b} \geq 1$ cuando $y_i = 1$ y $\langle \tilde{w}, x_j\rangle - \tilde{b} \leq -1$ cuando $y_j = -1$.

Afirmamos que $\Delta(w_0) \leq \frac{1}{||w_0||}$ para cualquier vector unitario w_0. Si esto es cierto, entonces podemos concluir que $\Delta\left(\frac{w}{||w||}\right) = \frac{1}{||w||}$ y por lo tanto $w^* = \frac{w}{||w||}$. Supongamos, por contradicción, que para algún vector unitario $w_0 \in \mathbb{R}^n$ se cumple $\Delta(w_0) > \frac{1}{||w_0||}$. Consideremos el vector $w = \frac{w_0}{\Delta(w_0)}$ junto con

$$b = \frac{1}{2\Delta(w_0)}\left(\min_{y_i=1}\langle w_0, x_i\rangle + \max_{y_i=-1}\langle w_0, x_i\rangle\right).$$

Se verifica que:

$$\langle w, x_i\rangle - b = \frac{\langle w_0, x_i\rangle - \frac{1}{2}\left(\min_{y_i=1}\langle w_0, x_i\rangle + \max_{y_i=-1}\langle w_0, x_i\rangle\right)}{\Delta(w_0)} \geq 1 \quad \text{si } y_i = 1,$$

y

$$\langle w, x_i\rangle - b = \frac{\langle w_0, x_i\rangle - \frac{1}{2}\left(\min_{y_i=1}\langle w_0, x_i\rangle + \max_{y_i=-1}\langle w_0, x_i\rangle\right)}{\Delta(w_0)} \leq -1 \quad \text{si } y_i = -1.$$

Pero $||w||^2 = \frac{||w_0||^2}{\Delta(w_0)^2} = \frac{1}{\Delta(w_0)^2} < ||w||^2$, lo cual es una contradicción con el hecho de que w sea un minimizador.

Apéndice G

Conceptos Matemáticos Fundamentales

G.1. Algunos conceptos de Análisis Funcional

Procedemos a resumir los estudios realizados por el autor de este trabajo sobre algunos conceptos de análisis funcional necesarios para acercarnos a comprender los resultados y teoremas aplicados. Este resumen fue preparado en base a al mathcamp del curso (Rifkin *et al.* 2003) y literatura de referencia.

Presentamos a continuación algunas de las definiciones básicas de análisis funcional . Primero, se presentan espacios vectoriales de fimension finita y estructuras adicionales que manejaremos sobre los anteriores. Se introducen luego los espacios de Hilbert, que juegan un rol rundamental en este trabajo, y se comentaran algunas peculiaridades al trabajar en dimension infinita. Finalmente, comentaremos de matrices y operadores lineales, en el contexto de mapeos lineales entre espacios vectoriales. Veremos como las matrices representan funciones lineales entre espacios vectoriales de dimension finita, y se desarrollara una teoria paralela sobre operadores lineales entre espacios de Hilbert en general. Asumimos que estamos trabajando con un campo base $\mathbb{R}$.

Estructuras en Espacios Vectoriales

Un **Espacio Vectorial** V es un conjunto con estructura lineal. Esto quiere decir que podemos sumar elementos espacio vectorial o multiplicar sus elimentos por escalares para obtener otro elemento. Un ejemplo conocido de un espacio vectorial es $\mathbb{R}n$. Dados $x = (x_1, ..., x_n)$ y $y = (y_1, ..., y_n)$ en $\mathbb{R}^n$, podemos formar un nuevo vector $x + y = (x_1 + y_1, ..., x_n + y_n) \in \mathbb{R}^n$. En forma similar podemos

formar $rx = (rx_1, ..., rx_n) \in \mathbb{R}^n$ (Rifkin *et al.* 2003).

Definido en forma más rigurosa un **Espacio Vectorial** o K-espacio vectorial (Steinwart & Christmann 2008) es una terna $(E, +, \cdot)$, donde E es un conjunto no vacío y $+ : E \times E \to E$ y $\cdot : K \times E \to E$ son mapeos que satisfacen:

- $(x + y) + z = x + (y + z)$ para todo $x, y, z \in E$.
- $x + y = y + x$ para todo $x, y \in E$.
- Existe un elemento $0 \in E$ tal que $x + 0 = x$ para todo $x \in E$.
- Para todo $x \in E$ existe un elemento $-x \in E$ tal que $x + (-x) = 0$.
- $(\alpha\beta) \cdot x = \alpha \cdot (\beta \cdot x)$ para todo $\alpha, \beta \in K$, $x \in E$.
- $1 \cdot x = x$ para todo $x \in E$.
- $(\alpha + \beta) \cdot x = \alpha \cdot x + \beta \cdot x$ para todo $\alpha, \beta \in K$ y $x \in E$.
- $\alpha \cdot (x + y) = \alpha \cdot x + \alpha \cdot y$ para todo $\alpha \in K$ y $x, y \in E$.

Cada espacio vectorial tiene una **base**. Un subconjunto $B = \{v_1, ..., v_m\}$ de V es llamado una **base** si cada vector $v \in V$ puede ser expresado únicamente como una combinacion lineal $v = c_1 v_1 + \cdots + c_n v_m$ para ciertas constantes $c_1, ..., c_m \in \mathbb{R}$. La cardinalidad de B es conocida como la **dimensión** de V. Es de destacar que dos espacios vectoriales finito dimensionales sobre $\mathbb{R}$ son isomórficos, dado que una biyeccción entre las bases puede ser extendida linealmente para obtener un isomorfismo entre los dos espacios vectoriales. De esta forma, salvo un isomorfismo, para cada $n \in N$ existe un solo espacio vectorial n-dimensional que es $\mathbb{R}^n$. Una **métrica** ρ en un conjunto $\mathcal{X}$ es una función que asigna a cada par de puntos en $\mathcal{X}$ un número real no negativo $\rho(x, y)$ tal que se satisfacen los siguientes axiomas:

1. $\rho(x, y) = 0 \Leftrightarrow x = y$, para todo $x, y \in \mathcal{X}$;
2. $\rho(x, y) = \rho(y, x)$ para todo $x, y \in \mathcal{X}$;
3. $\rho(x, z) \leq \rho(x, y) + \rho(y, z)$ para todo $x, y, z \in \mathcal{X}$.

La pareja $(\mathcal{X}, \rho)$ se denomina entonces **espacio métrico** (Arkhangelskiĭ 1990) y al número $\rho(x, y)$ se llama la **distancia** distancia entre los puntos x

y y. Los axiomas métricos expresan en forma abstracta las propiedades bien conocidas de la distancia usual en el plano y en el espacio tridimensional. La condición 3 se llama axioma del triángulo y la condición 2 el axioma de simetría. Los espacios vectoriales pueden tener estructuras adicionales que los distinguen a unos de otros, que a contiunación mencionaremos. Siguiendo con el razonamiento anterior, lo formulamos de otra forma para repasar otros conceptos de la **distancia** (métrica) sobre V: es una función $d : V \times V \to \mathbb{R}$ que satisface

- Positividad: $d(v, w) \geq$ para todo $v, w \in V$, y $d(v, w) = 0$ si y solo si $v = w$.
- Simestría: $d(v, w) = d(w, v)$ para todo $v, w \in V$.
- Desigualdad triangular: $d(v, w) \leq d(v, x) + d(x, w)$ para todo $v, w, x \in V$.

La más común en $\mathbb{R}^n$ esta dada por $d(x, y) = \sqrt{(x_1 - y_1)^2 + \cdots + (x_n, y_n)^2}$. Es de destacar que esta noción de métrica no requiere una estructura lineal ni ninguna otra sobre V, una métrica puede ser definida en cualquier conjunto. Un concepto similar que sí requiere una estructura lineal en $\mathbb{R}^n$ sobre V es la **norma**, que mide el "largo"de los vectores en V. Formalmente una norma es una funcion $\| \cdot \| : V \to \mathbb{R}$ que satisface las siguientes propiedades:

- Positividad: $\|v\| \geq 0$ para todo $v \in V$, y $\|v\| = 0$ si y solo si $v = 0$.
- Homogeneidad: $\|rv\| = |r|\|v\|$ para todo $r \in \mathbb{R}$ y $v \in V$.
- Subaditivdad: $\|v + w\| \leq \|v\| + \|w\|$ para todo $v, w \in V$.

Por ejemplo, la norma euclidiana en $\mathbb{R}^n$ es $\|x\|_2 = \sqrt{x_1^2 + \cdots + x_n^2}$ que también es llamada la norma l_2. También es relevante la norma l_1 $\|x\|_1 = |x_1| + \cdots + |x_n|$. Podemos generalizar estos ejemplos con la norma l_p. Dado un espacio vectorial con norma $(V, \| \cdot \|)$, podemos definir la **función distancia** sobre V como $d(v, w) = \|v - w\|$. Por ejemplo a la norma l_2 en $\mathbb{R}^n$ brinda la función distancia euclidiana:

$$d(x, y) = \|x - y\|_2 = \sqrt{(x_1 - y_1)^2 + \cdots + (x_n - y_n)^2},$$

mientras la norma l_1 en $\mathbb{R}^n$ brinda la función distancia Mahattan o Taxi.

$$d(x, y) = \|x - y\|_1 = \|x_1 - y_1\| + \cdots + \|x_n - y_n\|,$$

Es importante observar que todas las normas es un espacio vectorial finito dimensional V son **equivalentes**. Esto quiere decir que para dos normas cualquiera μ y ν sobre V, existen dos constantes positivas C_1 y C_2 tales que para todo $v \in V, C_1\mu(v) \leq \nu(v) \leq C_2\mu(v)$. En particular la continuidad o convergencia con respecto a una norma implica continuidad o convergencia con respecto a cualquier otra norma en un espacio vectorial finito dimensional. Por ejemplo, sobre $\mathbb{R}^n$ tenemos la inecuación $\|x\|_1/\sqrt{n} \leq \|x\|_2 \leq \|x\|\|_1$.

Otra estructura que podemos incorporar a un espacio vectorial es el producto interno. Un **producto interno** sobre V es una función $\langle \cdot, \cdot \rangle : V \times V \to \mathbb{R}$ que satisface las siguientes propiedades:

- Simetria: $\langle v, w \rangle = \langle w, v \rangle$ para todo $v, w \in V$
- Linealidad: $\langle r_1 v_1 + r_2 v_2, w \rangle = r_1 \langle v_1, w \rangle + r_2 \langle v_2, w \rangle$ para todo $r_1, r_2 \in V$
- Positivodefinido: $\langle v, v \rangle \geq 0$ para todo $v \in V$, y $\langle v, v \rangle = 0$ si y solo si $v = 0$

Por ejemplo, el producto interno estándar en $\mathbb{R}^n =$ es $\langle x, y \rangle = x_1 y_1 + \cdots + x_n y_n$, que también se escribe como $x \cdot y$. Dado un espacio con producto interno $(V, \langle \cdot, \cdot \rangle)$, podemos definir la norma de $v \in V$ como $\|v\| = \sqrt{\langle v, v \rangle}$. Se puede comprobar que esta definición cumple con las propiedades antes mencionadas. Por otra parte no toda norma proviene de un producto interno. La condición necesaria y suficiente para que una norma sea inducida por un producto interno es la **ley del paralelogramo**:

$$\|v + w\|^2 + \|v - w\|^2 = 2\|v\|^2 + 2\|w\|^2.$$

Por ejemplo podemos verificar que la norma l_2 en $\mathbb{R}^n$ es inducida por el producto interno estandar, mientras que la norma l_1 no es inducida por un producto interno dado que no satisface la ley del paralelogramo.

Un resultado muy importante que involucra los productos internos es la **desigualdad de Cauchy-Schwarz**:

$$\langle v, w \rangle \leq \|v|\ ||w\| \text{para todo } v, w, \in V.$$

Un **Espacio de producto interno** es un espacio vectorial V con un producto interno $<\cdot, \cdot> : V \times V \to \mathbb{R}$. Este producto interno generaliza el concepto

de producto interno en un espacio euclidiano y nos permite definir conceptos como ortogonalidad y normas. Un **funcional lineal**[1] es una función lineal de un espacio vectorial V a su campo subyacente de escalares. Si V es un espacio vectorial sobre un cuerpo k, el conjunto de todos los funcionales lineales de V a k es en sí mismo un espacio vectorial sobre k con la adición y la multiplicación por escalares definidas puntualmente. Este espacio se denomina **espacio dual** de V. Se denota a menudo por $\mathrm{Hom}(V, k)$, o, cuando el cuerpo k se sobreentiende, V^*; también se usan otras notaciones. El espacio dual desempeña un papel clave en el análisis funcional, particularmente en la optimización y el estudio de la convergencia débil.

Espacios de Hilbert

Un objeto central del desarrollo de la teoría del aprendizaje en general, es un **Espacio de Hilbert**: un espacio con producto interno completo. Es complejo trabajar en dimensiones infinitas, por ejemplo, cuando tratamos con espacios de funciones. La mayor parte de la discusión en la Sección 1 se traslada fácilmente al caso de espacios vectoriales de dimensión infinita, pero debemos ser un poco cuidadosos con la noción de base, ya que ahora tenemos que lidiar con sumas infinitas. En particular, solo nos preocuparemos por los espacios de Hilbert $\mathcal{H}$ que tienen una base ortonormal contable $(v_n)_{n=1}^{\infty}$, de modo que podamos escribir cada elemento $x \in \mathcal{H}$ como

$$v = \sum_{n=1}^{\infty} (v, v_n) v_n.$$

Repasaremos ahora de algunos conceptos que nos permitirán dar sentido a estas propiedades. Primero discutimos sucesiones de Cauchy y completitud. Recordemos que una sucesión $(v_n)_{n\in\mathbb{N}}$ en un espacio normado V converge a $v \in V$ si $||v_n - v|| \to 0$ cuando $n \to \infty$, o equivalentemente, si para todo $\epsilon > 0$ existe $N \in \mathbb{N}$ tal que $||v_n - v|| < \epsilon$ siempre que $n \geq N$. Intuitivamente, esto significa que v_n se acerca arbitrariamente a v a medida que avanzamos en la sucesión. Una condición similar para una sucesión es ser de Cauchy. Una sucesión $(v_n)_{n\in\mathbb{N}}$ en V es de Cauchy si la distancia entre cualquier par de elementos de la sucesión se vuelve arbitrariamente pequeña a medida que avanzamos en la sucesión. Más formalmente, $(v_n)_{n\in\mathbb{N}}$ es de Cauchy si para todo $\epsilon > 0$ existe $N \in \mathbb{N}$ tal que $||v_m - v_n|| < \epsilon$ siempre que $m, n \geq N$. Claramente, toda sucesión

[1] También llamado **forma lineal**.

convergente es de Cauchy, por la desigualdad triangular, pero el recíproco no es cierto. Un espacio vectorial normado es**completo** si toda sucesión de Cauchy converge. Intuitivamente, esto significa que no hay "huecos.[en] el espacio. Por ejemplo, el conjunto de los números racionales ($\mathbb{Q}$) no es completo ya que le faltan los irracionales. Más concretamente, la sucesión 1.4142, 1.41421, 1.414213, ... converge a $\sqrt{2}$ en los números reales ($\mathbb{R}$), pero $\sqrt{2} \notin \mathbb{Q}$. Por otro lado, $\mathbb{R}$ es completo por definición (de hecho, $\mathbb{R}$ es la completación de $\mathbb{Q}$), y se puede demostrar que $\mathbb{R}^n$ es completo para todo número natural n. Además, todo espacio vectorial normado de dimensión finita (sobre $\mathbb{R}$) es completo. Esto se debe a que, como vimos en la Sección 1, todo espacio vectorial real de dimensión n es isomorfo a $\mathbb{R}^n$, y cualquier par de normas en $\mathbb{R}^n$ son equivalentes. Por lo tanto, V es completo si y solo si $\mathbb{R}^n$ es completo bajo la norma estándar, lo cual es cierto. Ahora estamos en condiciones de definir los espacios de Hilbert: un **espacio de Hilbert** es un espacio con producto interno completo. La observación al final del párrafo anterior muestra que $\mathbb{R}^n$ y cualquier espacio con producto interno de dimensión finita son ejemplos de espacios de Hilbert. El ejemplo arquetípico de un espacio de Hilbert de dimensión infinita es el espacio de sucesiones de cuadrado sumable,

$$l_2 = \{(a_n)_{n=1}^{\infty} | a_n \in \mathbb{R}, \sum_{n=1}^{\infty} a_n^2 < \infty\},$$

donde la suma y la multiplicación por escalares se definen componente a componente, y el producto interno se define por $\langle a, b\rangle = \sum_{n=1}^{\infty} a_n b_n$. También podemos considerar el análogo continuo de l_2, es decir, el espacio de funciones de cuadrado integrable

$$L_2([0,1]) = \{f : [0,1] \to \mathbb{R} | \int_0^1 f(x)^2 dx < \infty\},$$

donde la integral es la integral de Lebesgue con respecto a la medida de Lebesgue dx en $[0,1]$. La suma y la multiplicación por escalares en este espacio se definen puntualmente, y el producto interno está dado por $\langle f, g\rangle_{L_2} = \int_0^1 f(x)g(x)dx$. Con esta estructura, $L_2([0,1])$ también es un espacio de Hilbert de dimensión infinita. Un espacio de Hilbert siempre tiene una base ortonormal, pero esta podría ser no numerable. Típicamente, solo nos interesan los espacios de Hilbert con una base ortonormal numerable, y una condición natural que podemos imponer para asegurar esta propiedad es la separabilidad. Intuitivamente, un espacio es separable si puede ser aproximado por un subconjunto

numerable de él mismo, en un sentido que precisaremos pronto. Por ejemplo, los números reales ($\mathbb{R}$) pueden ser aproximados por los números racionales ($\mathbb{Q}$), que son numerables, por lo que $\mathbb{R}$ es separable. Un espacio lineal R se dice que está o es **normado** (Kolmogorov & Fomin 2012) si a cada elemento $x \in R$ le corresponde un número no negativo $\|x\|$, llamado la norma de x, tal que:

- $\|x\| = 0$ si y solo si $x = 0$,
- $\|\alpha x\| = |\alpha|\|x\|$, para todo escalar α,
- $\|x + y\| \leq \|x\| + \|y\|$.

  La noción de una sucesión de Cauchy en $\mathbb{R}$ puede generalizarse a espacios vectoriales normados. Decimos que una sucesión (x_n) en un espacio vectorial normado E es una **sucesión de Cauchy** si satisface la siguiente propiedad:

  Para todo $\epsilon > 0$, existe un $N(\epsilon) \in \mathbb{N}$ tal que $||x_m - x_n|| < \epsilon$, si $m, n \geq N(\epsilon)$.

  Decimos que un espacio vectorial normado E es completo, o un **espacio de Banach**, si toda sucesión de Cauchy en E converge. Teorema: El espacio vectorial normado (Coleman 2012) $(\mathbb{R}^n, ||\cdot||_\infty)$ es un espacio de Banach. Demostración. Sea (x_k) una sucesión de Cauchy en $\mathbb{R}^n$. Usando superíndices para las coordenadas de los elementos de la sucesión, tenemos $x_k = (x_k^1, x_k^2, \ldots, x_k^n)$. Para $i = 1, 2, \ldots, n$, la sucesión (x_k^i) es una sucesión de Cauchy en $\mathbb{R}$. Como las sucesiones de Cauchy en $\mathbb{R}$ convergen, para cada i existe un x^i tal que

  $$\lim_{k\to\infty} x_k^i = x^i.$$

  Si definimos $x = (x^1, x^2, \ldots, x^n)$, entonces es fácil ver que

  $$\lim_{k\to\infty} x_k = x.$$

  □

  Según (Wibisono 2010) un **espacio de Banach** es un espacio normado completo. Por ejemplo, $\mathcal{C}([0, 1])$ con la norma del supremo es un espacio de Banach. Otro ejemplo puede ser el espacio de sucesiones absolutamente

sumables,

$$\ell_1 = \left\{ (a_n)_{n=1}^{\infty} \mid a_n \in \mathbb{R}, \sum_{n=1}^{\infty} |a_n| < \infty \right\},$$

donde la suma y la multiplicación por escalares se definen componente a componente, y la norma está dada por

$$\|a\|_1 = \sum_{n=1}^{\infty} |a_n|.$$

Teorema. Un espacio vectorial normado de dimensión finita es un espacio de Banach (Coleman 2012). En particular, los espacios vectoriales normados $(\mathbb{R}^n, ||\cdot||_p)$, para $1 \leq p \leq \infty$, son espacios de Banach.

Demostración. Sea $(u_i)_{i=1}^n$ una base del espacio vectorial normado de dimensión n, $(E, ||\cdot||)$. Si $x = \sum_{i=1}^n x_i u_i$ y definimos $\phi(x) = (x_1, x_2, \ldots, x_n)$, entonces ϕ es un isomorfismo lineal de E sobre $\mathbb{R}^n$. Ahora, definiendo $||\phi^{-1}(y)||_* = ||y||_\infty$ para $y \in \mathbb{R}^n$, obtenemos una norma en E. Como $(\mathbb{R}^n, ||\cdot||_\infty)$ es completo, también lo es $(E, ||\cdot||_*)$. La equivalencia de normas en E implica que $(E, ||\cdot||)$ también es completo.

Matrices y Operadores

Además de hablar de espacios vectoriales, también podemos hablar de operadores en esos espacios. Un **operador lineal** es una función $L : V \to W$ entre dos espacios vectoriales que preserva la estructura lineal. En dimensión finita, todo operador lineal puede ser representado por una matriz al elegir una base tanto en el dominio como en el rango, es decir, trabajando en coordenadas. Por esta razón, enfocaremos la primera parte de nuestra discusión en matrices. Si V es n-dimensional y W es m-dimensional, entonces una transformación lineal $L : V \to W$ está representada por una matriz A cuyas columnas son los valores de L aplicados a la base de V. El rango de A es la dimensión de la imagen de A, y la nulidad de A es la dimensión del núcleo de A. El **Teorema de rango-nulidad** establece que $rango(A) + nulidad(A) = m$, la dimensión del dominio de A.

También nótese que la transpuesta de A es una matriz n × m $A^T quesatisface$

$$\langle Av, w \rangle_{R^m} = (Av)^T w = v^T A^T w = \langle v, A^T w \rangle_{R^n},$$

para todo $v \in \mathbb{R}^n$ y $w \in \mathbb{R}^m$. Sea A una matriz $n \times n$ con entradas reales. Recordemos que un *autovalor* $\lambda \in \mathbb{R}$ de A es una solución a la ecuación $Av = \lambda v$ para algún vector no nulo $v \in \mathbb{R}^n$, y v es el autovector de A correspondiente a λ. Si A es simétrica, es decir, $A = A^T$, entonces los autovalores de A son reales. Además, en este caso el Teorema espectral nos dice que existe una base ortonormal de $\mathbb{R}^n$ que consiste en los autovectores de A. Sean $v_1, v_2, ..., v_n$ esta base ortonormal de autovectores y $\lambda_1, \lambda_2, ..., \lambda_n$ los autovalores correspondientes. Es posible entonces escribir:

$$A = \sum_{i=1}^{n} \lambda_i v_i v_i^T,$$

lo cual se denomina **descomposici[on espectral** de A. También podemos escribir esto como

$$A = V \Lambda V^T$$

donde V es la matriz $n \times n$ cuyas columnas son v_i y Λ es la matriz diagonal $n \times n$ con entradas λ_i. La ortonormalidad de $v_1, v_2, ..., v_n$ hace que V sea una matriz ortogonal, es decir, $V^{-1} = V^T$.

G.2. Algunos conceptos de Análisis Convexo

Estudiamos el apéndice A6 de (Steinwart & Christmann 2008) donde se desarrollan distintos temas de análisis convexo entre ellos:

- Propiedades de las funciones convexas.
- Cálculo subdiferencial para funciones convexas.
- Conceptos adicionales de convexidad.

Propiedades de las Funciones Convexas

Decimos que un subconjunto A de un espacio de Banach E se denomina **convexo** si, para todo $x_1, x_2 \in A$ y todo $\alpha \in (0, 1)$, tenemos

$$\alpha x_1 + (1 - \alpha) x_2 \in A.$$

En este caso, una función $f : A \to \mathbb{R} \cup \{\infty\}$ se denomina **convexa** si, para todo $x_1, x_2 \in A$ y todo $\alpha \in (0, 1)$, tenemos

$$f(\alpha x_1 + (1 - \alpha)x_2) \leq \alpha f(x_1) + (1 - \alpha)f(x_2).$$

Además, f se denomina cóncava si $-f$ es convexa. En el apéndice mencionado de (Steinwart & Christmann 2008) se cita a (Rockafellar 1970) haciendo referencia al resultado que a continuación planteamos. Lema: Matriz Hessiana de funciones convexas. Sea $O \subset \mathbb{R}^n$ un conjunto abierto y convexo y sea $g : O \to \mathbb{R}$ una función dos veces continuamente diferenciable. Entonces, g es convexa si y solo si su matriz Hessiana $Q_x = (q_{ij}(x))_{i,j}$ definida por

$$q_{ij}(x) = \frac{\partial^2 g}{\partial x_i \partial x_j}(x_1, ..., x_n)$$

es semidefinida positiva para todo $x \in O$. Una consecuencia inmediata de este resultado es el siguiente ejemplo para una función convexa de $\mathbb{R}^n$ en $\mathbb{R}$. Sea $K \in \mathbb{R}^{n \times n}$ una matriz simétrica, $b \in \mathbb{R}^n$, y $c \in \mathbb{R}$. La función cuadrática>

$$g(a) := a^T K a + b^T a + c, \quad a \in \mathbb{R}^n,$$

es convexa en $\mathbb{R}^n$ si y solo si K es semidefinida positiva, es decir, si $a^T K a \geq 0$ para todo $a \in \mathbb{R}^n$. Lema: Continuidad de funciones convexas. Sea $f : \mathbb{R} \to \mathbb{R} \cup \{\infty\}$ una función convexa y sea Dom $f = \{t \in \mathbb{R} : f(t) < \infty\}$. Entonces tenemos:

- f es continua en todo $t \in$ Int Dom f.
- Si f es semicontinua inferiormente, entonces $\frac{f}{\text{Dom } f}$ es continua.

El lema anterior no es cierto en general para espacios de Banach. De hecho, todo espacio de Banach E de dimensión infinita tiene un funcional lineal, y por lo tanto convexo, $x^* : E \to \mathbb{R}$ que no es continuo en ningún punto. **Teorema Fundamental del Cálculo**. Sea $f : [a, b] \to \mathbb{R}$ una función Lipschitz continua. Entonces f es diferenciable en casi todo punto $t \in [a, b]$. Además, la derivada f' de f, definida casi en todas partes, es integrable en el sentido de Lebesgue y satisface

$$f(x) = f(a) + \int_a^x f'(t)dt, \quad x \in [a, b].$$

Este Teorema puede ser utilizado para definir un test de convexidad y una formula para calcular las constantes locales de Lipschitz. Dados dos espacios de Banach E y F y un subconjunto $A \subset E$, llamamos a una función $f : A \to F$ **Lipschitz continua** si existe una constante $c > 0$ tal que

$$\|f(x) - f(x')\|_F \leq c\|x - x'\|_E \quad \text{para todo } x, x' \in A.$$

En este caso, la menor constante c se denota por $\|f\|_{Lip}$. Además, una función $f : \mathbb{R} \to \mathbb{R}$ se llama "localmente Lipschitz continua"si para todo $t > 0$ la restricción $f|_{[-t,t]}$ de f al intervalo $[-t, t]$ es Lipschitz continua. Lema: Continuidad de Lipschitz local y convexidad. Toda función convexa $f : \mathbb{R} \to \mathbb{R}$ es localmente Lipschitz continua, y para $t > 0$ tenemos

$$\|f|_{[-t,t]}\|_{Lip} \leq \frac{2}{t}\|f\|_{[-2t,2t]_\infty}.$$

Si además $f(0) = 0$, entonces $s \mapsto \frac{\|f\|_{[-s,s]\infty}}{s}$ es creciente en $(0, \infty)$ y tenemos

$$\|f|_{[-t,t]}\|_{Lip} \leq \frac{1}{t}\|f|_{[-2t,2t]_\infty}, \quad t > 0.$$

Lema: Prueba de convexidad. Sea $f : [a, b] \to \mathbb{R}$ una función Lipschitz continua y sea $N \subset [a, b]$ un conjunto de medida de Lebesgue nula tal que f es diferenciable en todo $t \in [a, b] \setminus N$. Supongamos que, para todo $s, t \in [a, b] \setminus N$ con $s \leq t$, tenemos $f'(s) \leq f'(t)$. Entonces, f es convexa.

Lema: Cálculo de la constante de Lipschitz. Sea $f : [a, b] \to \mathbb{R}$ una función Lipschitz continua y sea $N \subset [a, b]$ un conjunto de medida de Lebesgue nula tal que f es diferenciable en todo $t \in [a, b] \setminus N$. Entonces tenemos

$$\|f\|_{Lip} = \sup_{t \in [a,b] \setminus N} |f'(t)|.$$

Teorema: Existencia de minimizadores. Sea E un espacio de Banach reflexivo y sea $f : E \to \mathbb{R} \cup \{\infty\}$ una función convexa y semicontinua inferiormente. Si existe un $M > 0$ tal que el conjunto $\{x \in E : f(x) \leq M\}$ es no vacío y acotado, entonces f tiene un mínimo global, es decir, existe un $x_0 \in E$ tal que

$$f(x_0) \leq f(x), \quad \forall x \in E.$$

Además, si f es estrictamente convexa, entonces x_0 es el único elemento que minimiza f.

Cálculo subdiferencial para funciones convexas

Recopilamos algunas propiedades importantes de los subdiferenciales. E y F denotan espacios de Banach sobre $\mathbb{R}$. Definición. Sea E un espacio de Banach, $f : E \to \mathbb{R} \cup \{\infty\}$ una función convexa, y $w \in E$ con $f(w) < \infty$. Entonces, el **subdiferencial** de f en w se define como

$$\partial f(w) := \{w' \in E' : \langle w', v - w\rangle \leq f(v) - f(w) \text{ para todo } v \in E\}.$$

Proposición. Sea $f : E \to \mathbb{R} \cup \{\infty\}$ una función convexa y $w \in E$ tal que $f(w) < \infty$. Si f es continua en w, entonces el subdiferencial $\partial f(w)$ es un subconjunto no vacío, convexo y débil*-compacto de E'. Además, si $c \geq 0$ y $\delta > 0$ son constantes que satisfacen

$$|f(v) - f(w)| \leq c\|v - w\|, \quad v \in w + \delta B_E,$$

entonces tenemos $\|w'\| \leq c$ para todo $w' \in \partial f(w)$. Proposición. Cálculo subdiferencial. Sean $f, g : E \to \mathbb{R} \cup \{\infty\}$ funciones convexas, $\lambda \geq 0$, y $A : F \to E$ un operador lineal acotado. Entonces, se cumplen las siguientes reglas:

- **Homogeneidad:**para todo $w \in E$ con $f(w) < \infty$, tenemos $\partial(\lambda f)(w) = \lambda \partial f(w)$.
- **Aditividad:** si existe un $w_0 \in E$ en el que tanto f como g son continuas, entonces, para todo $w \in E$ tal que $f(w) < \infty$ y $g(w) < \infty$, tenemos $\partial(f + g)(w) = \partial f(w) + \partial g(w)$.
- **Regla de la cadena:**si existe un $v_0 \in F$ tal que f es finita y continua en Av_0, entonces, para todo $v \in F$ tal que $f(Av) < \infty$, tenemos $\partial(f \circ A)(v) = A^* \partial f(Av)$, donde $A^* : E' \to F'$ denota el operador adjunto de A.
- **Mínimos:** la función f tiene un mínimo global en $w \in E$ si y solo si $0 \in \partial f(w)$.
- **Diferenciabilidad:** si f es finita y continua en $w \in E$, entonces f es diferenciable en el sentido de Gâteaux en w si y solo si $\partial f(w)$ es un singleton, y en este caso, tenemos $\partial f(w) = \{f'(w)\}$.

- **Monotonía:** Si f es finita y continua en todos los $w \in E$, entonces ∂f es un operador monótono, es decir, para todo $v, w \in E$, $v' \in \partial f(v)$, y $w' \in \partial f(w)$, tenemos $\langle v' - w', v - w \rangle \geq 0$.

Conceptos adicionales de convexidad

Introducimos algunas nociones más fuertes de convexidad y discutimos sus relaciones entre sí.
Una función $f : A \to \mathbb{R}$ en un subconjunto convexo $A \subset E$ de un espacio de Banach E se denomina **estrictamente convexa** si, para todo $x_1, x_2 \in A$ con $x_1 \neq x_2$ y todo $\alpha \in (0, 1)$, tenemos:

$$f(\alpha x_1 + (1 - \alpha)x_2) < \alpha f(x_1) + (1 - \alpha)f(x_2).$$

El **módulo de convexidad** de f se define como:

$$\delta_f(\varepsilon) := \text{ínf} \left\{ \frac{f(x_1) + f(x_2)}{2} - f\left(\frac{x_1 + x_2}{2}\right) : x_1, x_2 \in A \text{ con } \|x_1 - x_2\| \geq \varepsilon \right\},$$

para todo $\varepsilon > 0$, y decimos que f es **uniformemente convexa** si $\delta_f(\varepsilon) > 0$ para todo $\varepsilon > 0$. El siguiente lema describe algunas relaciones menos triviales entre las diferentes nociones de convexidad. Lema: Dado un intervalo I y una función $f : I \to \mathbb{R}$, tenemos:

- Si f es convexa y satisface $f(\alpha_0 x_1 + (1 - \alpha_0)x_2) = \alpha_0 f(x_1) + (1 - \alpha_0)f(x_2)$ para algún $x_1, x_2 \in I$, $\alpha_0 \in [0, 1]$, entonces, para todo $\alpha \in [0, 1]$, tenemos

$$f(\alpha x_1 + (1 - \alpha)x_2) = \alpha f(x_1) + (1 - \alpha)f(x_2).$$

- Si f es continua, entonces f es convexa si y solo si, para todo $x_1, x_2 \in I$, tenemos

$$f\left(\frac{x_1 + x_2}{2}\right) \leq \frac{1}{2}f(x_1) + \frac{1}{2}f(x_2).$$

- Si f es continua, entonces f es estrictamente convexa si y solo si, para todo $x_1, x_2 \in I$ con $x_1 \neq x_2$, tenemos

$$f\left(\frac{x_1 + x_2}{2}\right) < \frac{1}{2}f(x_1) + \frac{1}{2}f(x_2).$$

- Si f es uniformemente convexa y continua, entonces es estrictamente convexa. Recíprocamente, si I es compacto y f es estrictamente convexa y continua, entonces es en realidad uniformemente convexa.

Lema: Sea $I \subset \mathbb{R}$ un intervalo no vacío, $f : I \to \mathbb{R}$ una función estrictamente convexa, y $\varepsilon > 0$. Entonces se tiene

$$\delta_f(2\varepsilon) = \text{ínf} \left\{ \frac{f(x-\varepsilon) + f(x+\varepsilon)}{2} - f(x) : x \text{ satisface } x - \varepsilon \in I \text{ y } x + \varepsilon \in I \right\}.$$

Apéndice H

Historia de la Investigación del Problema del Aprendizaje

Vapnik en la introducción de (Vapnik 1995) destaca cuatro períodos en la hisotria de la investigación del problema de aprendizaje.

- Construcción de las primeras maquinas que aprenden.
- Construcción de los fundamentos de la teoria.
- Construcción de las redes neuronales.
- Construcción de alternativas a las redes neuronales.

Años 60: El perceptron de Rosenblatt

Rosenblatt en el año 1957 propuso el primer modelo de Aprendizaje Estadístico llamado Perceptron. La idea ya habia sido discutida en la literatura de la neurofisiología, pero Rosenblatt hizo algo inusual: describió el modelo como un programa para computadoras y demostró mediante experimentos que el modelo podía ser generalizado. El perceptron fue contruído para resolver problemas de clasificación binaria. Del punto de vista geométrico, cada "neurona"dividía el espacio X en dos regiones: una región en la cual la salida y tomaba el valor 1 y otra en la cual tomaba el valor -1. Ambas regiones eran separadas por un hiperplano. Rosenblatt consideró un modelo compuesto de varias neuronas: la salidas de las neuronas del nivel anterior son las entradas del siguiente nivel. El último nivel solo contiene una neurona. De esta forma el perceptron dividía el espacio X en dos partes separadas por una superficie lineal a trozos. En los años

60 no estaba claro como seleccionar los coeficientes de las neuronas en forma simultanea (la solución fue obtenida 20 años después). Siguiendo los conceptos fisiológicos tradicionales Rosenblatt propuso un algoritmo simple para iterativamente encontrar los coeficientes, transformando el espacio X en uno nuevo Z eligiendo coeficientes para todas las neuronas exceptuando la última. En 1962 Novikoff demostro el primer Teorema sobre el Perceptron, comenzando la Teoria del Aprendizaje. Este Teorema jugó un rol extremadamente importante en la teoria del aprendizaje, vinculando la causa de la capacidad de generalización del modelo con el principio de minimizar la cantidad de errores en la muestra de entrenamiento.

Años 60 y 70: Fundamentos de la Teoría del Aprendizaje

A medida que los experimentos con el Perceptron se hicieron populares otros tipos de algoritmos de Aprendizaje estadístico fueron sugeridos. Pero, a diferencia del Perceptron, estos algoritmos o "maquinas´´ fueron considerados desde su origen como herramientas para resolver problemas de la vida real antes que un modelo general para el fenómeno del aprendizaje.

Luego de los años 60, nada realmente extraordinario sucedió en el contexto del análisis aplicado: hasta 1986 cuando se propone un algoritmo general llamado *back-propagation* para obtener, en forma simultánea, los coeficientes de varias neuronas.

Sin embargo, estos años fueron fructíferos para el desarollo de la Teoría del Aprendizaje. En 1968 una filosofía de la Teoría del Aprendizaje fue desarrrollada.

Usando los conceptos de entropía VC y dimensión VC fue formulada la ley de los grandes numeros en espacios funcionales, su relación con los preocesos de aprendizaje fueron descritos, y las cotas no asintóticas para la tasa de convergencia fueron obtenidas.

Las cotas obtenidas introdujeron un novedoso principio inductivo: la minimizacion del riesto esstructural, completando el desarrollo de la Teoria del Aprendizaje. El nuevo paradigma de la teoria es presentado en el libro Theory of Pattern Recognition de Vapnik y Chervonenkis en 1974.

Entre 1976 y 1981 los resultados originalmente obtenidos para el conjunto de funciones indicatrices fueron generalizados para las funciones reales. Estos resultados fueron presentados en 1979 en otra obra de Vapnik. Algunas de las ideas desarrolladas en este periodo son las siguientes:

- La Teoría de Regularizacion para la solución de problemas ill-posed: Tykhonov y otros.
- La estimación no paramétrica de densidades: Rosenblatt, Parze y otros.
- La idea de la complejidad algorítmica: Solomonoff, Kolmogorov y Chaitin.

Años 80: Redes Neuronales

En 1986, varios autores propusieron independientemente un método para obtener simultáneamente los coeficientes vectoriales de todas las neuronas de un Perceptrón utilizando el llamado método de retropropagación (LeCun 1986) (Rumelhart *et al.* 1986). La idea de este método es extremadamente simple: si en lugar del modelo de neurona de McCulloch-Pitts se considera un modelo ligeramente modificado donde la función discontinua signo $(w \cdot x - b)$ se reemplaza por la continua denominada aproximación sigmoide,

$$y = S(w \cdot x - b),$$

donde $S(u)$ es una función monótona con las propiedades

$$S(-\infty) = -1, \quad S(+\infty) = 1,$$

por ejemplo, $S(u) = \tanh(u)$, entonces la composición de las nuevas neuronas es una función continua que para cualquier x fijo tiene un gradiente con respecto a todos los coeficientes de todas las neuronas. En 1986 se encontró el método para evaluar este gradiente. Utilizando el gradiente evaluado se puede aplicar cualquier técnica basada en gradiente para construir una función que aproxime la función deseada. Simplificación de los Objetivos del Análisis Teórico

El descubrimiento de la técnica de retropropagación puede considerarse como el segundo nacimiento del Perceptrón. Sin embargo, este nacimiento ocurrió en

una situación completamente diferente. Desde 1960 aparecieron computadoras poderosas, además, nuevas ramas de la ciencia se involucraron en la investigación sobre el problema del aprendizaje. Esto cambió esencialmente la escala y el estilo de la investigación. A pesar del hecho de que no se puede afirmar con certeza que las propiedades de generalización del Perceptron con muchos neuronas ajustables sean mejores que las propiedades de generalización del Perceptron con una sola neurona ajustable y aproximadamente el mismo número de parámetros libres, la comunidad científica estaba mucho más entusiasmada con este nuevo método debido a la escala de los experimentos. Los primeros experimentos de Rosenblatt se realizaron para el problema del reconocimiento de dígitos. Para demostrar la capacidad de generalización del perceptrón, Rosenblatt utilizó datos de entrenamiento que consistían en varios cientos de vectores, conteniendo varias docenas de coordenadas. En los años 80 e incluso ahora en los 90, el problema del aprendizaje del reconocimiento de dígitos sigue siendo importante. Hoy en día para obtener buenas reglas de decisión se utilizan decenas (incluso cientos) de miles de observaciones sobre vectores con varios cientos de coordenadas. Esto requirió una organización especial de los procesos computacionales. Por lo tanto, en los años 80 los investigadores en Inteligencia Artificial se convirtieron en los principales actores en el juego del Aprendizaje Automático. Entre los investigadores de Inteligencia Artificial los partidarios de enfoques simples tenían una influencia considerable. Precisamente ellos declararon que "Las teorías complejas no funcionan, los algoritmos simples sí´´.

Los partidarios de enfoques simples en inteligencia artificial abordaron el problema del aprendizaje con una gran experiencia en la construcción de .ªlgoritmos simples´´ para problemas donde la teoría es muy complicada. A finales de los años 60 se prometieron traductores automáticos de lenguaje natural en un par de años (incluso ahora este problema extremadamente complicado está lejos de resolverse); el siguiente proyecto fue construir un solucionador general de problemas; después de esto vino el proyecto de construir un controlador automático de grandes sistemas, y así sucesivamente. Todos estos proyectos tuvieron poco éxito. El siguiente gran problema a investigar fue la creación de una tecnología de Aprendizaje Computacional o Automático. Primero, los partidarios de enfoques simples cambiaron la terminología. En particular el Perceptrón fue renombrado como Red Neuronal. Luego se implentó un programa de investigación conjunto con fisiólogos, y el estudio del problema del aprendizaje

se volvió menos general, más orientado a temas específicos. En los años 60 y 70, el objetivo principal de la investigación era encontrar la mejor manera de hacer inferencia inductiva a partir de muestras pequeñas. En los años 80, el objetivo se convirtió en construir un modelo de generalización que utilizara el cerebro. Uno de los primeros intentos de introducir teoría en la comunidad de Inteligencia Artificial se hizo en 1984 cuando se sugirió el modelo probablemente aproximadamente correcto (PAC). Este modelo se define por un caso particular del concepto de consistencia comunmente utilizado en estadística en el que se incorporaron algunos requisitos sobre la complejidad computacional.

Los 90: Volviendo a los orígenes

En la década del 90 algo cambió en relación con las Redes Neuronales: se comienza a prestar más atención a técnicas alternativas (Powell 1992), dedicando gran esfuerzo al estudio del método de funciones de base radia. Al igual que en los años 60, las Redes Neuronales vuelven a cambiar su nombre pasándose a denominar "perceptrones multicapa´´. Las partes mas avanzadas de la Teoría del Aprendizaje Estadístico ahora atraen a más investigadores. En particular, en los últimos años, tanto el principio de minimización del riesgo estructural como el principio de longitud de descripción mínima se han convertido en temas populares de análisis. Las discusiones sobre la teoría de muestras pequeñas, en contraste con la asintótica, se han generalizado.

Parece que todo está volviendo a sus fundamentos.

Además, la teoría del aprendizaje estadístico ahora juega un papel más activo: después de completar el análisis general de los procesos de aprendizaje, se inició la investigación en el área de la síntesis de algoritmos óptimos que poseen el nivel más alto de capacidad de generalización para cualquier número de observaciones.

Bibliografía

Adams, R. A. (1975): *Sobolev spaces*. Academic Press [A subsidiary of Harcourt Brace Jovanovich, Publishers], New York-London. Pure and Applied Mathematics, Vol. 65.

Alon, N., S. Ben-David, N. Cesa-Bianchi, & D. Haussler (1997): "Scale-sensitive dimensions, uniform convergence, and learnability." *J. ACM* **44(4)**: p. 615–631.

Ana M, A., E. Manuel, V. Mariano J, & A.-M. M Carmen (2013): "Functional analysis of chemometric data." *Open Journal of Statistics* **2013**.

Anthony, M. & P. L. Bartlett (2009): *Neural Network Learning: Theoretical Foundations*. New York, NY, USA: Cambridge University Press, 1st edition.

Arak, T. V. (1982): "On the convergence rate in kolmogorov's uniform limit theorem. i." *Theory of Probability & Its Applications* **26(2)**: pp. 219–239.

Arkhangelskiĭ, A. V. (1990): *General topology*. Encyclopaedia of mathematical sciences. Berlin: Springer-Verlag.

Aronszajn, N. (1950): "Theory of reproducing kernels." *Transactions of the American Mathematical Society* **68(3)**: pp. 337–404.

Arribas-Gil, A. & J. Romo (2014): "Shape outlier detection and visualization for functional data: the outliergram." *Biostatistics* .

Battiti, R., M. Brunato, & A. Villani (2002): "Statistical learning theory for location fingerprinting in wireless lans." *Technical report*, Universita di Trento, Dipartimento di Informatica e Telecomunicazioni.

Blackmer, T. M., J. S. Schepers, & G. E. Varvel (1994): "Light reflectance compared with other nitrogen stress measurements in corn leaves." *Agronomy Journal* **86(6)**: pp. 934–938.

Boucheron, S., G. Lugosi, & O. Bousquet (2004a): "Concentration inequalities." In O. Bousquet, U. Luxburg, & G. Rotsch (editors), "Advanced Lectures on Machine Learning," volume 3176 of *Lecture Notes in Computer Science*, pp. 208–240. Springer Berlin Heidelberg.

Boucheron, S., G. Lugosi, & O. Bousquet (2004b): "Introduction to statistical learning theory." In O. Bousquet, U. Luxburg, & G. Rotsch (editors), "Advanced Lectures on Machine Learning," volume 3176 of *Lecture Notes in Computer Science*, pp. 169–207. Springer Berlin Heidelberg.

Breiman, L. (1996): "Bagging predictors." *Machine Learning* **24(2)**: pp. 123–140.

Breunig, M. M., H.-P. Kriegel, R. T. Ng, & J. Sander (2000): "Lof: Identifying density-based local outliers." *SIGMOD Rec.* **29(2)**: pp. 93–104.

Chaochao, H., W. Xiaodi, & T. Wuqin (2007): "Infrared image simulation based on statistical learning theory." *International Journal of Infrared and Millimeter Waves* **28(12)**: pp. 1143–1153.

Chapelle, O., B. Schölkopf, & A. Zien (editors) (2006): *Semi-Supervised Learning*. Cambridge, MA: MIT Press.

Chau, F., Y. Liang, J. Gao, & X. Shao (2004): *Chemometrics: From Basics to Wavelet Transform*. Chemical Analysis: A Series of Monographs on Analytical Chemistry and Its Applications. Wiley.

Chuma, E., L. Meloni, Y. Iano, & L. Bravo-Roger (2017): "Fpga implementation of a de-noising using haar level 5 wavelet transform."

Coleman, R. (2012): *Calculus on Normed Vector Spaces*.

Cortes, C. & V. Vapnik (2009): "Support-vector networks." *Chem. Biol. Drug Des.* **297**: pp. 273–297.

Cox, D. D. (1988): "Approximation of Least Squares Regression on Nested Subspaces." *The Annals of Statistics* **16(2)**: pp. 713 – 732.

Cristianini, N. & J. Shawe-Taylor (2000): *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods.* Cambridge University Press.

Cucker, F. & S. Smale (2001): "On the mathematical foundations of learning." *Bulletin of the American Mathematical Society* **39**: pp. 1–49.

Cucker, F. & S. Smale (2002): "Best choices for regularization parameters in learning theory: On the bias-variance problem." *Foundations of Computational Mathematics* **2**: pp. 413–428.

Cucker, F. & D. Zhou (2007): *Learning Theory: An Approximation Theory Viewpoint.* Cambridge Monographs on Applied and Computational Mathematics. Cambridge University Press.

Cuevas, A., M. Febrero, & R. Fraiman (2007): "Robust estimation and classification for functional data via projection-based depth notions." *Computational Statistics* **22(3)**: pp. 481–496.

De Vito, E., A. Caponnetto, & L. Rosasco (2005): "Model selection for regularized least-squares algorithm in learning theory." *Foundations of Computational Mathematics* **5(1)**: pp. 59–85.

Devroye, L., L. Györfi, & G. Lugosi (1996): *A probabilistic theory of pattern recognition.* Applications of mathematics. Springer.

Drucker, H. (1997): "Improving regressors using boosting techniques." *Proceedings of the 14th International Conference on Machine Learning* .

Duda, R. O., P. E. Hart, & D. G. Stork (2000): *Pattern Classification (2Nd Edition).* Wiley-Interscience.

Evgeniou, T., T. Poggio, M. Pontil, & A. Verri (2002): "Regularization and statistical learning theory for data analysis." *Comput. Stat. Data Anal.* **38(4)**: pp. 421–432.

Evgeniou, T. & M. Pontil (2000): "Statistical learning theory: A primer." *International Journal of Computer Vision* **38**: pp. 9–13.

Evgeniou, T., M. Pontil, & T. Poggio (2000): "Regularization networks and support vector machines." *Adv. Comput. Math.* **13**: pp. 1–50.

Fariña, R. A. (2026): "Relativity for the realm of the living: A proposal for an extended general theory." *BioSystems* **262**: p. 105727.

Febrero-Bande, M. & M. Oviedo de la Fuente (2012): "Statistical computing in functional data analysis: The r package fda.usc." *Journal of Statistical Software* **51(4)**: pp. 1–28.

Fisher, R. A. (1992): *Statistical Methods for Research Workers*, pp. 66–70. New York, NY: Springer New York.

Freund, Y. & R. E. Schapire (1997): "A decision-theoretic generalization of on-line learning and an application to boosting." *Journal of Computer and System Sciences* **55(1)**: pp. 119–139.

Friedman, J. (1991): "Multi-variate adaptive regression splines (with discussion)." *The Annals of Statistics* **19**.

Gallagher, J., C.-K. Lai, & E. Weber (2022): "On a topological erdos similarity problem."

García, D., M. Jung, M. Maestre, G. A. Muñoz-Fernández, & J. B. Seoane-Sepúlveda (2024): "Geometry of homogeneous polynomials in $\setminus^2$."

Girosi, F., M. Jones, & T. Poggio (1998): "Regularization theory and neural networks architectures." *Neural Comput* **7**.

Group, J. P. E. (????): "Jpeg 200." `https://jpeg.org/jpeg2000/index.html`. [Accessed 06-02-2025].

Hainmueller, J. & C. Hazlett (2013a): "Kernel regularized least squares: Reducing misspecification bias with a flexible and interpretable machine learning approach." *Political Analysis* .

Hainmueller, J. & C. Hazlett (2013b): *KRLS: Kernel-based Regularized Least Squares (KRLS)*. R package version 0.3-2.

Hastie, T., R. Tibshirani, & J. Friedman (2009): *The elements of statistical learning: data mining, inference and prediction*. Springer, 2 edition.

Hocking, R. (1996): *Methods and Applications of Linear Models: Regression and the Analysis of Variance*. Wiley Series in Probability and Statistics. Wiley.

Izenman, A. (2008): *Modern Multivariate Statistical Techniques: Regression, Classification, and Manifold Learning*. Springer Texts in Statistics. Springer-Verlag New York.

James, G., D. Witten, T. Hastie, & R. Tibshirani (2014): *An Introduction to Statistical Learning: With Applications in R*. Springer Publishing Company, Incorporated.

Joachims, T. (2002): *Learning to Classify Text Using Support Vector Machines: Methods, Theory, and Algorithms*, volume 29.

Kearns, M. J. & R. E. Schapire (1994): "Efficient distribution-free learning of probabilistic concepts." *Journal of Computer and System Sciences* **48(3)**: pp. 464–497.

Khoshhesab, Z. M. (2012): *Infrared Spectroscopy - Materials Science, Engineering and Technology*. Intech.

Kolmogorov, A. & S. Fomin (2012): *Elements of the Theory of Functions and Functional Analysis [Two Volumes in One]*. Martino Fine Books.

Kolountzakis, M. N. (2023): "Sets of full measure avoiding cantor sets."

Kowalczyk, S. (2014): "On operations in c(x) determined by continuous functions." *Acta Mathematica Hungarica* **142**.

LeCun, Y. (1986): "Learning processes in an asymmetric threshold network." In E. Bienenstock, F. Fogelman-Soulié, & G. Weisbuch (editors), "Disordered systems and biological organization," pp. 233–240. Les Houches, France: Springer-Verlag.

Li, H., Y. Liang, & Q. Xu (2009): "Support vector machines and its applications in chemistry." *Chemometrics and Intelligent Laboratory Systems* **95(2)**: pp. 188 – 198.

Liang, L., B. Wang, Y. Guo, H. Ni, & Y. Ren (2009): "A support vector machine-based analysis method with wavelet denoised near-infrared spectroscopy." *Vibrational Spectroscopy* **49(2)**: pp. 274 – 277.

Lin, Z. & Z. Bai (2011): *Probability Inequalities*. Springer.

Lipkowitz, K., T. Cundari, & D. Boyd (2007): *Reviews in Computational Chemistry*. Number v. 23 in Reviews in Computational Chemistry. Wiley.

Lue, H.-H. (2009): "Sliced inverse regression for multivariate response regression." *Journal of Statistical Planning and Inference* **139(8)**: pp. 2656 – 2664.

von Luxburg, U. & B. Schölkopf (2008): "Statistical Learning Theory: Models, Concepts, and Results." Note.

Mallat, S. (1989): "A theory for multiresolution signal decomposition: The wavelet representation." *IEEE Trans. Pattern Anal. Mach. Intell.* **11**: pp. 674–693.

Margolis, D., W. H. L. Jr., R. Gottlieb, & X. Qiao (2011): "A complex adaptive system using statistical learning theory as an inline preprocess for clinical survival analysis." *Procedia Computer Science* **6(0)**: pp. 279 – 284. Complex adaptive systems.

Micchelli, C. A. & M. Pontil (2005): "On learning vector-valued functions." *Neural Computation* **17(1)**: pp. 177–204.

Morettin, P. & A. Pinheiro (2017): *Wavelets in Functional Data Analysis*.

Nadeem, H. & T. J. Heindel (2018): "Review of noninvasive methods to characterize granular mixing." *Powder Technology* **332**: pp. 331–350.

Niyogi, P. & F. Girosi (1996): "On the relationship between generalization error, hypothesis complexity, and sample complexity for radial basis functions." *Neural Computation* **8(4)**: pp. 819–842.

Niyogi, P. & F. Girosi (1999): "Generalization bounds for function approximation from scattered noisy data." *Adv. Comput. Math.* **10**: pp. 51–80.

Poggio, T. & S. Smale (2005): "The mathematics of learning: Dealing with data *." In W. Chu & T. Lin (editors), "Foundations and Advances in Data Mining," volume 180 of *Studies in Fuzziness and Soft Computing*, pp. 1–19. Springer Berlin Heidelberg.

Pontil, M. (2003): "Learning with reproducing kernel hilbert spaces: a guide tour." *Bulletin of the Italian Artificial Intelligence Association, AI* IA Notizie* .

Powell, M. J. (1992): "The theory of radial basis function approximation in 1990." *Advances in numerical analysis* pp. 105–210.

Puntanen, S. (2008): "Linear models in statistics, second edition by alvin c. rencher, g. bruce schaalje." *International Statistical Review* **76**: pp. 445–445.

R Core Team (2014): *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rifkin, R., S. Mukherjee, T. Poggio, & A. Rakhlin (2003): "9.520 statistical learning theory and applications, spring 2003." http://ocw.mit.edu (Accedido 25 Jul, 2015).

Rifkin, R. M. & R. A. Lippert (2007): "Notes on regularized least-squares." *Technical report*, Motorola Institute.

Rockafellar, R. T. (1970): *Convex analysis*. Princeton Mathematical Series. Princeton, N. J.: Princeton University Press.

Rossel, R. & T. Behrens (2010): "Using data mining to model and interpret soil diffuse reflectance spectra." *Geoderma* **158(1)**: pp. 46–54.

Rumelhart, D. E., G. E. Hinton, & R. J. Williams (1986): "Learning internal representations by error propagation, parallel distributed processing, explorations in the microstructure of cognition, ed. de rumelhart and j. mcclelland. vol. 1. 1986." *Biometrika* **71(599-607)**: p. 6.

Salnikov, D. (2021): "A constructive proof of the glivenko-cantelli theorem."

Scholkopf, B., R. Herbrich, & A. J. Smola (2001): "A generalized representer theorem." In "In Proceedings of the Annual Conference on Computational Learning Theory," pp. 416–426.

Schölkopf, B. & A. Smola (2001): *Smola, A.: Learning with Kernels - Support Vector Machines, Regularization, Optimization and Beyond. MIT Press, Cambridge, MA*, volume 98.

Schölkopf, B., A. Smola, & K.-R. Müller (1998): "Nonlinear component analysis as a kernel eigenvalue problem." *Neural Computation* **10**: pp. 1299–1319.

Schölkopf, B., K. Tsuda, & J. Vert (2003): *Kernel Methods in Computational Biology.*

Shao, Y., Y. He, & Y. Wang (2007): "A new approach to discriminate varieties of tobacco using vis/near infrared spectra." *European Food Research and Technology* **224(5)**: pp. 591–596.

Shawe-Taylor, J. & N. Cristianini (2004a): *Kernel Methods for Pattern Analysis.* New York, NY, USA: Cambridge University Press.

Shawe-Taylor, J. & N. Cristianini (2004b): *Kernel Methods for Pattern Analysis.*

Sheather, S. (2009): *A Modern Approach to Regression with R.* Springer Texts in Statistics. Springer New York.

Simon, H. A. (1983): "Why should machines learn?" In R. Michalski, J. Carbonnel, & T. Mitchell (editors), "Machine Learning: An Artificial Intelligence Approach," pp. 25–37. Palo Alto, CA: Tioga.

Smale, S., L. Rosasco, J. Bouvrie, A. Caponnetto, & T. Poggio (2009): "Mathematics of the neural response."

Steinwart, I. & A. Christmann (2008): *Support Vector Machines.* Springer Publishing Company, Incorporated, 1st edition.

Stenlund, H., E. Johansson, J. Gottfries, & J. Trygg (2009): "Unlocking interpretation in near infrared multivariate calibrations by orthogonal partial least squares." *Analytical Chemistry* **81(1)**: pp. 203–209. PMID: 19117451.

Thosar, S., R. Forbess, N. Ebube, Y. Chen, R. Rubinovitz, M. Kemper, G. Reier, T. Wheatley, & A. Shukla (2001): "A comparison of reflectance and transmittance near-infrared spectroscopic techniques in determining drug content in intact tablets." *Pharmaceutical development and technology* **6**: pp. 19–29.

Tikhonov, A. N. & V. Y. Arsenin (1977): *Solutions of ill-posed problems*. Washington, D.C.: John Wiley & Sons, New York: V. H. Winston & Sons. Translated from the Russian, Preface by translation editor Fritz John, Scripta Series in Mathematics.

Torgo, L. (2010): *Data Mining with R, learning with case studies*. Chapman and Hall/CRC.

Tsypkin, I. Z. (1973): *Foundations of the theory of learning systems [by] Ya. Z. Tsypkin. Translated by Z. J. Nikolic*. Academic Press New York.

Vapnik, V. (1982): *Estimation of Dependences Based on Empirical Data: Springer Series in Statistics (Springer Series in Statistics)*. Secaucus, NJ, USA: Springer-Verlag New York, Inc.

Vapnik, V. (1995): *The Nature of Statistical Learning Theory*. New York, NY, USA: Springer-Verlag New York, Inc.

Vapnik, V. (1998): *Statistical learning theory*. Adaptive and learning systems for signal processing, communications, and control. Wiley.

Vapnik, V. N. (1999): "An overview of statistical learning theory." *Transactions on Neural Networks* **10(5)**: pp. 988–999.

Vapnik, V. N. & A. Y. Chervonenkis (1971): "On the uniform convergence of relative frequencies of events to their probabilities." *Theory of Probability & Its Applications* **16(2)**: pp. 264–280.

Velliangiri, S., S. Alagumuthukrishnan, & S. I. Thankumar joseph (2019): "A review of dimensionality reduction techniques for efficient computation." *Procedia Computer Science* **165**: pp. 104–111. 2nd International Conference on Recent Trends in Advanced Computing ICRTAC -DISRUP - TIV INNOVATION , 2019 November 11-12, 2019.

Wahba, G. (1990): *Spline Models for Observational Data.* CBMS-NSF Regional Conference Series in Applied Mathematics. Society for Industrial and Applied Mathematics.

Wechsler, H., Z. Duric, F. Li, & V. Cherkassky (2004): "V.: Motion estimation using statistical learning theory." *PAMI* **26**: pp. 466–478.

Weisberg, S. (2013): *Applied Linear Regression.* Wiley Series in Probability and Statistics. Wiley.

Wibisono, A. (2010): "Functional analysis review."

Wold, S. (1995): "Chemometrics; what do we mean with it, and what do we want from it?" *Chemometrics and Intelligent Laboratory Systems* **30(1)**: pp. 109 – 115. InCINC '94 Selected papers from the First International Chemometrics Internet Conference.

Wu, Q., Y. Ying, & D.-X. Zhou (2006): "Learning theory: from regression to classification." *Studies in Computational Mathematics* **12**: pp. 257–290.

Zhai, Y., L. Cui, X. Zhou, Y. Gao, T. Fei, & W. Gao (2013): "Estimation of nitrogen, phosphorus, and potassium contents in the leaves of different plants using laboratory-based visible and near-infrared reflectance spectroscopy: Comparison of partial least-square regression and support vector machine regression methods." *Int. J. Remote Sens.* **34(7)**: pp. 2502–2518.

Zhang, Y., Q. Cong, Y. Xie, J. Yang, & B. Zhao (2008): "Quantitative analysis of routine chemical constituents in tobacco by near-infrared spectroscopy and support vector machine." *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy* **71(4)**: pp. 1408–1413.