Facultad de Matemáticas

Grado de Matemáticas
Trabajo Fin de Grado

La dimensión Vapnik-Chervonenkis en Ciencia de los datos

Realizado por:
Miguel Vázquez Caraballo

Dirigido por:
D. Joaquín Borrego Díaz

Departamento:
Ciencias de la Computación e Inteligencia Artificial

Sevilla, Junio 2022

Índice general

Resumen

En esta memoria se expone la dimensión de VC, un concepto originalmente definido por V. N. Vapnik y A. Ya. Chervonenkis que es una medida de la capacidad que tiene una familia de funciones para realizar clasificación binaria de un conjunto de datos, de gran importancia en la teoría de Aprendizaje Estadístico. Se aporta su definición y algunas propiedades importantes, así como una colección de ejemplos de su cálculo.

Para mostrar su relación con la Ciencia de los datos, se desarrolla el marco de aprendizaje PAC como modelo fundacional de la noción de *aprendibilidad* automatizada de un problema. Así, se identifican las clases de funciones aprendibles a partir de su dimensión VC mediante el Teorema Fundamental del Aprendizaje Estadístico.

Abstract

In this report we present the VC dimension, a concept originally defined by V. N. Vapnik and A. Ya. Chervonenkis, which is a measure of the ability of a family of functions to perform binary classification of a data set, of great importance in Statistical Learning theory. Its definition and some important properties are provided, as well as a collection of examples of its computation.

To show its relation to Data Science, the PAC learning framework is developed as a foundational model of the notion of automated problem learning. Thus, the classes of learnable functions are identified from their VC dimension by means of the Fundamental Theorem of Statistical Learning.

Introducción

La teoría del aprendizaje estadístico ha desarrollado diferentes métodos y herramientas para cuantificar la capacidad y riqueza de un modelo de clasificación binaria. Entre éstos una de las más famosas es la dimensión de Vapnik-Chervonenkis.

El problema del aprendizaje estadístico se plantea como sigue. Asumiendo que los ejemplos están distribuidos según una probabilidad desconocida $\mathcal{D}$ sobre la población, el algoritmo considera un *conjunto hipótesis* $\mathcal{H}$ de posibles *conceptos*, que son funciones que relacionan los datos con el valor objetivo a estudiar, recibe un conjunto de entrenamiento S de datos escogidos según $\mathcal{D}$ y devuelve una hipótesis $h_S \in \mathcal{H}$ a partir de S. El objetivo es que a partir de los datos en S la hipótesis se asemeje al concepto objetivo c a aprender que relaciona los datos con el valor objetivo real de la población total.

Para estudiar el problema teóricamente, Leslie Valiant expone en 1984 el marco de aprendizaje Correcto Probablemente Aproximado (PAC por sus siglas en inglés). En el marco de aprendizaje PAC, se considera que una clase de conceptos es aprendible si existe un algoritmo de aprendizaje que al tomar un conjunto de datos i.i.d. según cualquier distribución sobre la población, es capaz de devolver, con alta probabilidad, una hipótesis con pequeño error, provisto de una muestra de tamaño suficiente.

En este trabajo veremos los conceptos teóricos del marco PAC y como esta noción de aprendizaje PAC está íntimamente relacionada con la finitud de la dimensión Vapnik-Chervonenkis de una clase de funciones.

Objetivo

El objetivo principal del trabajo es presentar y comprender el concepto de dimensión de Vapnik-Chervonenkis. Para ello desarrollamos conceptos teóricos

de la ciencia de datos y estadística y presentamos el marco PAC y el paradigma de Reducción del Riesgo Empírico.

Estructura de la memoria

En el CAPÍTULO 1 se presenta la definición formal de la dimensión VC así como un número de propiedades importantes.

Para profundizar la comprensión de la idea en el CAPÍTULO 2 se da una colección de ejemplos de cálculo de la dimensión VC de distintos conjuntos y funciones.

A lo largo del CAPÍTULO 3 desarrollaremos los conceptos teóricos básicos del aprendizaje estadístico, definiremos formalmente el problema anteriormente descrito, desarrollaremos la idea de las clases que son aprendibles PAC con propiedades como la convergencia uniforme, que establece que una clase se comportará de forma similar tanto en el conjunto de datos aportados como en la población total y por lo tanto generaliza bien y veremos algunas relaciones entre estas ideas y la dimensión VC.

En la CAPÍTULO 4 presentamos el teorema No Free Lunch, que muestra que por cada par de algoritmos de aprendizaje, existen tantos problemas en el que el primer algoritmo es mejor que el segundo como problemas en el que el segundo algoritmo es mejor que el primero. La conclusión es que no un existe un aprendiz universal pues todos se comportarán igual de bien en promedio. Como conclusión describimos el Teorema Fundamental Aprendizaje Estadístico que demuestra la equivalencia entre la finitud de la dimensión de Vapnik-Chervonenkis de la clase en cuestión, la propiedad de convergencia uniforme de esta clase de funciones y su aprendibilidad PAC.

Capítulo 1

La dimensión Vapnik-Chervonenkis

Para la redacción de este capítulo, se han utilizado como referencia [1] y [8]. Se puede encontrar una traducción del artículo original de Vapnik y CHervonenkis en [11]

En este capítulo se introduce el concepto principal del trabajo, *Dimensión Vapnik-Chernovenkis* o dimensión VC. Se presentará la definición y se discutirá sus propiedades para dar una intuición sobre la importancia y el uso de la dimensión VC en la ciencia de datos y otras ramas de las matemáticas.

1.1. Definición de la dimensión VC

La dimensión Vapink-Chervonenkis es una medida de la complejidad que tiene una familia de funciones que puede ser aprendida por un algoritmo de aprendizaje estadístico de clasificación binaria. Se define como la cardinalidad del mayor conjunto de puntos que puede ser desmenuzado por alguna función de la familia. La definición formal en [3] es la siguiente.

Definición 1.1.1 ***(Dimensión VC de una familia de funciones)*** *Sea $U \subseteq \mathbb{R}^d$ y $\mathcal{F}$ una familia de funciones de U en $\mathbb{R}$. Al par $(U, \mathcal{F})$ lo denominaremos* espacio de aprendizaje.

- *Diremos que una nube de puntos $X \subseteq U$ es* desmenuzada *por $\mathcal{F}$ si todo subconjunto de X es aprendido por $\mathcal{F}$, i.e. si para todo subconjunto $A \subseteq X$ $\exists f \in \mathcal{F}$ tal que $\forall x \in X \ (x \in A \Leftrightarrow f(x) > 0)$.*

- *Se define la* dimensión de Vapnik-Chervonenkis *de* $(U, \mathcal{F})$ *como*

$$dim_{VC}(U, \mathcal{F}) = sup\,\{n \mid \exists X \subseteq U \text{ con } |X| = n \text{ desmenuzado por } \mathcal{F}\}$$

En aprendizaje automático $\mathcal{F}$ suele ser una familia de funciones parametrizadas $\mathcal{F} = \{f(x, \alpha) \mid \alpha \in \mathbb{R}^n\}$.

Consideraremos también la dimensión VC para pares $(U, \mathcal{F})$ donde $\mathcal{F}$ es una familia de subconjuntos de U identificando cada subconjunto con su función característica.

Se puede considerar desde un punto de vista geométrico como el tamaño del mayor conjunto de puntos que la familia de funciones puede separar de todas las formas posibles. Esto lo veremos de forma más clara con ejemplos en el siguiente tema.

La dimensión VC es un concepto fundamental en aprendizaje automático, ya que mide la capacidad que tiene un algoritmo de clasificación binaria de aprender un conjunto de datos utilizando las funciones que puede construir el modelo.

1.2. Propiedades de la dimensión VC

Veamos algunas propiedades importantes de la dimensión VC. Consideremos un espacio de aprendizaje $(U, \mathcal{F})$ con $dim_{VC}(U, \mathcal{F}) = d < \infty$. Para $n \leq d$ existe un conjunto de puntos $A \subseteq U$, $|A| = n$ tal que los 2^n subconjuntos distintos de A pueden ser aprendidos por $\mathcal{F}$. Surge la pregunta del máximo número de subconjuntos que pueden ser aprendidos si $n > d$.

Definición 1.2.1 *Dado* $X \subseteq U$*, se define* la restricción de $\mathcal{F}$ a X *como el conjunto de subconjuntos de* X *que son aprendidos por* $\mathcal{F}$*.*

$$\mathcal{F}_X = \{Y \subseteq X \mid Y \text{ es aprendido por } \mathcal{F}\}$$

y se define la $\mathcal{F}$-granularidad *de* X *como la cardinalidad de la restricción de* $\mathcal{F}$ *a* X*,*

$$|X|_{\mathcal{F}} = |\{Y \subseteq X \mid Y \text{ es aprendido por } \mathcal{F}\}| = |\mathcal{F}_X|$$

Se tiene que $0 \leq |X|_{\mathcal{F}} \leq 2^{|X|}$, alcanzando el máximo cuando $\mathcal{F}$ desmenuza a X.

Definición 1.2.2 *La* función granularidad máxima *de un espacio de aprendizaje $(U, \mathcal{F})$ para tamaño n se define como el mayor número de subconjuntos de algún $X \subseteq U$ de tamaño n que pueden ser aprendidos.*

$$\pi_{(U,\mathcal{F})}(n) = max\{|X|_{\mathcal{F}} \mid X \subseteq U \ y \ |X| = n\}$$

Para valores pequeños de n, $\pi_{(U,\mathcal{F})}(n)$ crecerá como 2^n. Si $dim_{VC}(U, \mathcal{F}) = d$ es finita, una vez que n alcance a d la función crecerá de forma polinómica. La definición de dimensión Vapnik-Chervonenkis puede ser reformulada como $dim_{VC}(U, \mathcal{F}) = max\{n \mid \pi_{(U,\mathcal{F})}(n) = 2^n\}$.

El siguiente resultado muestra que si $dim_{VC}(U, \mathcal{F}) = d$ es finita, entonces $\pi_{(U,\mathcal{F})}(n)$ está acotado por un polinomio de grado d

Teorema 1.2.1 ***Lema de Sauer***
Para un espacio de aprendizaje $(U, \mathcal{F})$ con dimensión-VC d finita, se tiene $\pi_{(U,\mathcal{F})}(n) \leq \sum_{i=0}^{d} \binom{n}{i}$

Demostración: La prueba es por inducción en n y d. Para el caso base probamos el resultado para los pares (n, d) con $n \leq d$ o $d = 0$.

Para $n \leq d$, $\sum_{i=0}^{d} \binom{n}{i} = \sum_{i=0}^{n} \binom{n}{i} = 2^n$ y $\pi_{(U,\mathcal{F})}(n) = 2^n$. Para $d = 0$ se tiene que $\forall X \subseteq U$, $|X|_{\mathcal{F}} = 0$ y por lo tanto $\pi_{(U,\mathcal{F})}(n) = 0 \, \forall n \in \mathbb{N}$ y $\sum_{i=0}^{d} \binom{n}{i} = 1$.

Consideremos probados los casos $(n-1, d-1)$ y $(n-1, d)$.

Sea $X = \{x_1, .., x_n\} \subseteq U$ un subconjunto con $|X|_{\mathcal{F}} = \pi_{(U,\mathcal{F})}(n)$ y $\mathcal{G}$ la familia de subconjuntos de X que son aprendidos por $\mathcal{F}$, y sean $X' = \{x_1, .., x_{n-1}\} \subseteq X$ y $\mathcal{G}_1$ la familia de subconjuntos de X' que son aprendidos por $\mathcal{F}$.

Se define $\mathcal{G}_2 = \{g' \subseteq X' \mid (g' \in \mathcal{G}) \quad y \quad (g' \cup \{x_n\} \in \mathcal{G})\}$. Así definidos los conjuntos, $\mathcal{G} - \mathcal{G}_1 = \{g \in \mathcal{G} \mid g - \{x_1\} \notin \mathcal{G}_1\} = \mathcal{G}_2$, y por lo tanto se tiene que $|\mathcal{G}_1| + |\mathcal{G}_2| = |\mathcal{G}|$. Identificando cada conjunto con su función indicatriz, tenemos tres familias de funciones.

Como $dim_{VC}(U, \mathcal{G}_1) \leq dim_{VC}(U, \mathcal{G}) \leq d$, por definición de la función de granularidad máxima y por hipótesis de inducción,

$$|\mathcal{G}_1| \leq \pi_{(U,\mathcal{G}_1)}(n-1) \leq \sum_{i=0}^{d} \binom{n-1}{i}$$

Por otro lado, por definición de $\mathcal{G}_2$, si un conjunto $A \subseteq S'$ es desmenuzado por $\mathcal{G}_2$, entonces también es desmenuzado por $\mathcal{G}$. Por lo tanto, $dim_{VC}(U, \mathcal{G}_2) \leq dim_{VC}(U, \mathcal{G}) - 1 \leq d - 1$, y por definición de la función de granularidad máxima y por hipótesis de inducción,

$$|\mathcal{G}_2| \leq \pi_{(U,\mathcal{G}_2)}(n-1) \leq \sum_{i=0}^{d-1} \binom{n-1}{i}$$

Finalmente,

$$|\mathcal{G}| = |\mathcal{G}_1| + |\mathcal{G}_2| \leq \sum_{i=0}^{d} \binom{n-1}{i} + \sum_{i=0}^{d-1} \binom{n-1}{i} =$$

$$\sum_{i=0}^{d} \binom{n-1}{i} + \binom{n-1}{i-1} = \sum_{i=0}^{d} \binom{n}{i}$$

que completa la demostración. □

Este teorema establece límites a la función de granularidad máxima en los espacios de aprendizaje de dimensión VC finita. La aplicación que Vapnik y Chervonenkis dan a este lema es probar que toda distribución de probabilidad sobre un conjunto de datos puede ser estimada (mediante una familia de funciones con dimensión VC finita) con una muestra finita de los datos con cardinalidad que depende únicamente de la dimensión VC de la familia y de la cota de error exigida. Este resultado se conoce como el Teorema Fundamental del Aprendizaje Estadístico, y se demostrará en este trabajo en el tema cuatro.

Capítulo 2

Ejemplos de Dimensión Vapnik-Chervonenkis

En este capítulo se presentará una recopilación de distintos conjuntos y funciones y el cálculo de su dimensión VC.

2.1. Clases en el plano

Teorema 2.1.1 *Sea $\mathcal{H}$ la clase de las regiones planas determinadas por rectángulos con lados paralelos a los ejes. Entonces $dim_{VC}(\mathbb{R}^2, \mathcal{H}) = 4$*

Demostración: Consideremos cuatro puntos en el plano que forman las esquinas de un diamante. Cualquier subconjunto de estos puntos puede ser desmenuzado por $\mathcal{H}$.

Figura 2.1: Ejemplos de desmenuzados posibles para cuatro puntos dispuestos en forma de diamante.

Sin embargo, no se puede desmenuzar ningún conjunto de cinco puntos. Para comprobarlo consideremos cinco puntos en el plano y busquemos el rectángulo

de mínima área que los contenga. En cada arista hay al menos un punto. Identifiquemos un punto de este tipo por cada arista. Si un punto está en un esquina, puede ser identificado con ambas aristas. Si en una arista hay dos puntos o más, se elige uno cualquiera de ellos.

Hemos definido un conjunto de a lo más cuatro puntos. Cualquier rectángulo que contenga estos puntos contiene también a los demás, por lo tanto estos puntos no pueden ser desmenuzados del resto con rectángulos con aristas paralelas a los ejes, y $\dim_{VC}(\mathbb{R}^2, \mathcal{H}) = 4$. □

Teorema 2.1.2 *Sea $\mathcal{H}$ la clase de las regiones planas determinadas por cuadrados con lados paralelos a los ejes. Entonces* $\dim_{VC}(\mathbb{R}^2, \mathcal{H}) = 3$

Demostración: Existen tres puntos que pueden ser desmenuzados por cuadrados con lados horizontales y verticales. Veamos el ejemplo de puntos formando un triángulo rectángulo. Desmenuzar los subconjuntos de 1 y 3 puntos es trivial. La siguiente imagen muestra como podemos desmenuzar dos puntos.

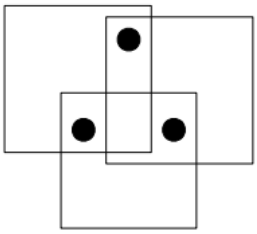


Figura 2.2: Desmenuzado de los tres subconjuntos posibles de dos puntos

Estudiemos ahora el caso de cuatro puntos.

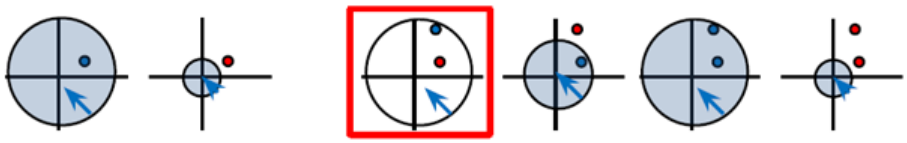
Llamemos A al punto de mayor ordenada, B al de menor ordenada, I al de menor abscisa y D al de mayor abscisa, asumiendo que se pueden definir de forma única. Asumamos sin pérdida de generalidad que la distancia de A a B es mayor que la de I a D. Entonces A y B no pueden ser desmenuzados por $\mathcal{H}$.

Si no se pueden definir estos puntos de forma única, entonces o bien hay un punto contenido en el mínimo rectángulo que encierra a los otros tres de forma que el subconjunto formado por esos tres puntos no puede ser desmenuzado, o los cuatro puntos están alineados y el subconjunto formado por los puntos de los extremos del segmento que une los cuatro puntos no es desmenuzable. Por lo tanto $\dim_{VC}(\mathbb{R}^2, \mathcal{H}) = 3$. □

Teorema 2.1.3 *Sea $\mathcal{H}$ la clase de los círculos centrados en el origen. Entonces* $\dim_{VC}(\mathbb{R}^2, \mathcal{H}) = 1$

Demostración: Consideremos un conjunto formado por un único punto $x \in \mathbb{R}$. Existen dos subconjuntos, el vacío y el propio conjunto. Para desmenuzar el vacío escoger cualquier círculo de radio $r < \|x\|_2$ y para desmenuzar el propio conjunto tomar $r > \|x\|_2$.

Consideremos ahora un conjunto de dos puntos en el plano. Si la distancia de los puntos al origen es la misma, $\mathcal{H}$ no puede desmenuzar ningún subconjunto formado por solo uno de los puntos, y si la distancia al origen es distinta, $\mathcal{H}$ no puede desmenuzar el subconjunto formado por el punto más alejado del origen. Por lo tanto $\mathcal{H}$ no puede desmenuzar ningún conjunto de dos o más puntos y $\dim_{VC}(\mathbb{R}^2, \mathcal{H}) = 1$



□

Teorema 2.1.4 *Sea $\mathcal{H}$ la clase formada por todos los polígonos convexos en el plano. Entonces* $\dim_{VC}(\mathbb{R}^2, \mathcal{H}) = \infty$

Demostración: Dado $n \in \mathbb{N}$ consideremos n puntos en la circunferencia unidad. Tomando cualquier subconjunto de estos puntos se ordenan en sentido horario y se unen con segmentos en ese orden. Esto nos da un polígono convexo que contiene a este subconjunto y a ningún otro punto que no pertenezca al subconjunto. Así esta familia puede desmenuzar subconjuntos arbitrariamente grandes y $\dim_{VC}(\mathbb{R}^2, \mathcal{H}) = \infty$. □

2.2. Clases en la recta real

Teorema 2.2.1 *Sea $\mathcal{H}$ la clase formada por los intervalos en los números reales. Entonces* $\dim_{VC}(\mathbb{R}, \mathcal{H}) = 2$

Demostración: Consideremos un conjunto $A = \{x_1, x_2\} \subset \mathbb{R}$ con $x_1 < x_2$. Hay cuatro subconjuntos de A. Para desmenuzar $\emptyset$ tomar el intervalo $(-\infty, a]$ con $a < x_1$. Para $\{x_1\}$ tomar $(-\infty, b]$ con $x_1 < b < x_2$. Para $\{x_2\}$ tomar $[b, +\infty)$. Para $\{x_1, x_2\}$ tomar $\mathbb{R}$.

Si tenemos un conjunto $B = \{x_1, x_2, x_3\} \subset \mathbb{R}$ con $x_1 < x_2 < x_3$ cualquier intervalo que contenga a x_1 y x_3 también contiene a x_2, luego no existe ningún intervalo que desmenuze $\{x_1, x_3\}$. Por lo tanto $\dim_{VC}(\mathbb{R}, \mathcal{H}) = 2$. □

Teorema 2.2.2 *Sea $\mathcal{H}$ la clase formada por los pares de intervalos en los números reales definidos como la unión de dos intervalos reales. Entonces $dim_{VC}(\mathbb{R}, \mathcal{H}) = 4$*

Demostración: Consideremos un conjunto de cuatro puntos reales. Desmenuzar los subconjuntos de uno, dos y cuatro puntos es trivial. Si tenemos un subconjunto de tres puntos o bien los tres son consecutivos y un intervalo que los contenga y no contenga al punto que no pertenece al subconjunto desmenuza al subconjunto o hay uno aislado de los otros dos por el punto que no pertenece al subconjunto y basta tomar un intervalo que contenga al punto aislado y otro que contenga al par de puntos restante y que no contengan al punto que no pertenece al subconjunto.

Sin embargo, no existe ningún conjunto de cinco puntos que sea desmenuzable, pues no podemos desmenuzar el subconjunto formado por el primer, tercer y quinto puntos con la unión de dos intervalos sin coger los otros dos. Luego $\dim_{VC}(\mathbb{R}, \mathcal{H}) = 5$. □

2.3. Clases en el espacio de d dimensiones

Teorema 2.3.1 *Sea*

$$S(d) := \{(-\infty, t_1] \times (-\infty, t_2] \times \ldots \times (-\infty, t_d] \;\mid\; (t_1, ..., t_d) \in \mathbb{R}^d\}$$

la clase de intervalos medios en $\mathbb{R}^d$. Entonces $\dim_{VC}(\mathbb{R}^d, S(d)) = d$

Demostración: Consideremos los d vectores unitarios. Es posible desmenuzar cualquier subconjunto tomando $t_i = 1$ si el vector i pertenece al subconjunto y $t_i = 0$ si no pertenece.

Sean $d + 1$ puntos arbitrarios $x_1, ..., x_{d+1}$. Tomemos para cada $1 \leq i \leq d$, $t_i = \text{máx}_{1 \leq j \leq d+1}(x_j)_i$. Hay como máximo d puntos distintos en la frontera de $(-\infty, t_1] \times \ldots \times (-\infty, t_d]$. Estos puntos no pueden ser desmenuzados por ningún conjunto de la familia $S(d)$. □

Teorema 2.3.2 *Sea $\mathcal{H}$ la clase formada por los semiespacios en d dimensiones. Entonces* $\dim_{VC}(\mathbb{R}^d, \mathcal{H}) = d + 1$

Se define un semiespacio como el conjunto de todos los puntos que quedan a un lado de un hiperplano, i.e. $\{x \mid a^T x \geq a_0\}$.

DEMOSTRACIÓN: Consideremos el conjunto consistente en el origen y los d vectores con coordenadas unitarias. Supongamos que tenemos un subconjunto A arbitrario. Si el origen pertenece a A, tomemos el vector a que tiene un cero en las coordenadas correspondientes a los vectores unitarios que pertenecen a A y un uno a los que no pertenecen. Entonces el semiespacio $\{x \mid a^T x \leq 0\}$ desmenuza A. Si el origen no pertenece a A se toma el vector a de forma inversa al caso anterior y consideramos el semiespacio $\{x \mid a^T x \geq 1\}$.

Veamos ahora que ningún conjunto de d+2 puntos puede ser desmenuzado por semiespacios. El teorema de Radon afirma que cualquier conjunto de d+2 puntos puede ser separado en dos subconjuntos disjuntos A y B tal que si $convex(A)$ y $convex(B)$ son sus respectivas envolventes convexas, entonces $convex(A) \cap convex(B) \neq \emptyset$.

El teorema de Radon implica que ningún conjunto de d+2 puntos puede ser desmenuzado por semiespacios. Tomando una partición disjunta A y B tal que $convex(A) \cap convex(B) \neq \emptyset$, si un semiespacio desmenuza estos dos conjuntos entonces el hiperplano que lo define separa sus envolventes convexas, por lo que $convex(A) \cap convex(B) = \emptyset$, una contradicción. □

Teorema 2.3.3 (Radon) *Para todo conjunto $S \subset \mathbb{R}^d$ con $|S| \geq d+2$ existe una partición en dos conjuntos disjuntos A y B tal que $convex(A) \cap convex(B) \neq \emptyset$.*

DEMOSTRACIÓN: Sin pérdida de generalidad, asumir que $|S| = d+2$. Sea A la matriz que tiene en cada columna un punto de S. A es una matriz $d \times (d+2)$. Sea B la matriz $(d+1) \times (d+2)$ resultante de añadir un fila de unos a A. Como el rango de B es a lo más $d+1$, sus columnas son linealmente dependientes.

Sea $\mathbf{x} = (x_1, x_2, ..., x_{d+2})$ un vector no nulo con $B\mathbf{x} = 0$. Reordenar los elementos de $\mathbf{x}$ para que $x_1, ..., x_t \geq 0$ y $x_{t+1}, ..., x_{d+2} < 0$. Normalizar $\mathbf{x}$ tal que $\sum_{i=1}^{t} |x_i| = 1$. Sean a_i y b_i las i-ésimas columnas de A y B respectivamente. Entonces $\sum_{i=1}^{t} |x_i| b_i = \sum_{i=t+1}^{d+2} |x_i| b_i = 1$ de donde $\sum_{i=1}^{t} |x_i| a_i = \sum_{i=t+1}^{d+2} |x_i| a_i$ y $\sum_{i=1}^{t} |x_i| = \sum_{i=t+1}^{d+2} |x_i|$. Como $\sum_{i=1}^{t} |x_i| = 1$ y $\sum_{i=t+1}^{d+2} |x_i| = 1$ cada lado de $\sum_{i=1}^{t} |x_i| a_i =$

$\sum_{i=t+1}^{d+2} |x_i| a_i = 1$ es una combinación convexa de columnas de A y por lo tanto de puntos de S. Tomando el conjunto de los primeros t puntos y otro conjunto de los puntos $t+1$ a $d+2$ tras la reordenación tenemos una partición de S que satisface las propiedades del teorema. □

Teorema 2.3.4 *Sea $\mathcal{H}(d)$ la clase formada por las esferas en d dimensiones. Entonces* $\dim_{VC}(\mathbb{R}^d, \mathcal{H}(d)) = d + 1$

Una esfera en d dimensiones es un conjunto

$$\{x \in \mathbb{R}^d \mid |x - x_0| \leq r\} \in \mathcal{H}(d)$$

Demostración: En primer lugar mostremos que ningún conjunto S de $d+2$ puntos puede ser desmenuzado por esferas. Supongamos que cualquier partición A_1 y A_2 de S puede ser desmenuzada por $\mathcal{H}(d)$. Entonces existen esferas $B_1, B_2 \in \mathcal{H}(d)$ tal que $B_1 \cap S = A_1$ y $B_2 \cap S = A_2$.

Si B_1 y B_2 no se intersecan, existe un hiperplano que las separa y entonces separa también a A_1 y A_2. Si se intersecan, no hay puntos de S en la intersección, y hay un hiperplano perpendicular a la recta que une los centros de las dos esferas que deja los puntos de A_1 a un lado y los de A_2 a otro. Por lo tanto S sería desmenuzable por semiespacios, que es falso pues hay $d+2$ puntos en S y los semiepacios tienen dimensión VC $d+1$.

Veamos ahora que el conjunto S de $d+1$ puntos formado por los vectores de coordenada unitaria y el origen es desmenuzable por esferas en d dimensiones. Sea $A \subset S$ un subconjunto y a el número de vectores unitarios en A. Tomemos una esfera con centro a_0 la suma de los vectores de A y radio $r \in (\sqrt{a-1}, \sqrt{a+1})$. La distancia de todo vector unitario perteneciente a A hasta a_0 es $\sqrt{a-1}$, y la de los vectores que no pertenecen a A es $\sqrt{a+1}$. Así que esta esfera desmenuza A.

Entonces las esferas en d dimensiones tienen dimensión VC $d+1$, al igual que los semiespacios en d dimensiones. □

2.4. Funciones reales

Teorema 2.4.1 *La clase de funciones*

$$\mathcal{F} = \{f \mid f(x) = sign(sen(\omega x)), \omega \geq 0\}$$

con dominio $X = [0, 2\pi]$ *tiene* $dim_{VC}(X, \mathcal{F}) = +\infty$

Demostración:

Para demostrar que la dimensión VC de esta familia es infinita buscamos para cualquier $n \in \mathbb{N}$ un conjunto de puntos en X de tamaño n que pueda ser desmenuzado por $\mathcal{F}$. Consideremos $\{(2\pi 10^{-i}, y_i)\}_{i=1}^{n} \subset X^n \times \{-1, 1\}^n$ donde y_i es la etiqueta de cada punto $x_i = 2\pi 10^{-i}$. Para un conjunto de este tipo escogeremos el parámetro

$$\omega = \frac{1}{2}\left(1 + \sum_{i=1}^{n} \frac{1 - y_i}{2} 10^i\right)$$

Primero probamos que se predicen correctamente los x_i con etiquetas negativas.

Se puede reescribir ω como

$$\omega = \frac{1}{2}\left(1 + \sum_{\{i\,:\,y_i=-1\}} 10^i\right).$$

Entonces, para cada j tal que $y_j = -1$

$$\begin{aligned}
\omega x_j &= \pi 10^{-j}\left(1 + \sum_{\{i\,:\,y_i=-1\}} 10^i\right) \\
&= \pi 10^{-j}\left(1 + 10^j + \sum_{\{i\,:\,y_i=-1,\, i\neq j\}} 10^i\right) \\
&= \pi\left(10^{-j} + 1 + \sum_{\{i\,:\,y_i=-1,\, i\neq j\}} 10^{i-j}\right) \\
&= \pi\left(10^{-j} + 1 + \sum_{\{i\,:\,y_i=-1,\, i>j\}} 10^{i-j} + \sum_{\{i\,:\,y_i=-1,\, i<j\}} 10^{i-j}\right).
\end{aligned}$$

Para $i > j$ los términos 10^{i-j} son potencias positivas de 10 y por lo tanto también números pares, así que pueden ser reescritos como $2k_i$ para ciertos $k_i \in \mathbb{N}$. Por lo que se tiene que

$$\sum_{\{i\,:\,y_i=-1,\, i>j\}} 10^{i-j} = \sum_{\{i\,:\,y_i=-1,\, i>j\}} 2k_i = 2k$$

Para cierto $k \in \mathbb{N}$. Para el otro sumando se tiene

$$\sum_{\{i\,:\,y_i=-1,\,i<j\}} 10^{i-j} < \sum_{i=1}^{+\infty} 10^{-i} = \sum_{i=0}^{+\infty} \left(\frac{1}{10}\right)^i - 1 = \frac{1}{1-0{,}1} - 1 = \frac{1}{9}.$$

Ahora, usando esto y que $10^{-j} < 0{,}1$, definimos

$$\epsilon = 10^{-j} + \sum_{\{i\,:\,y_i=-1,\,i<j\}} 10^{i-j} < \frac{1}{10} + \frac{1}{9} = 1$$

se tiene que

$$\omega x_j = \pi(1+\epsilon) + 2k\pi$$

y como $0 < \epsilon < 1$, finalmente $sin(\omega x_j) < 0$ y por lo tanto $f(x_j) = sign(sen(\omega x_j)) = -1 = y_j$ y todas las etiquetas negativas están bien clasificadas.

Ahora probaremos que también se predicen correctamente los x_i con etiquetas positivas.

El procedimiento es similar al anterior, con la diferencia de que al tener $y_j = 1$, el término 10^j se anula ahora en la definición de ω.

$$\begin{aligned}
\omega x_j &= \pi 10^{-j} \left(1 + \sum_{\{i\,:\,y_i=-1,\,i\neq j\}} 10^i\right) \\
&= \pi \left(10^{-j} + \sum_{\{i\,:\,y_i=-1,\,i>j\}} 10^{i-j} + \sum_{\{i\,:\,y_i=-1,\,i<j\}} 10^{i-j}\right) \\
&= \pi\epsilon + 2k\pi
\end{aligned}$$

y como $0 < \epsilon < 1$, y $0 < \pi\epsilon < \pi$ se tiene que $f(x_j) = 1 = y_j$.

Por lo tanto esta familia de funciones clasifica correctamente el conjunto, y lo hace para toda $n \in \mathbb{N}$, por lo que $dim_{VC}(X, \mathcal{F}) = +\infty$

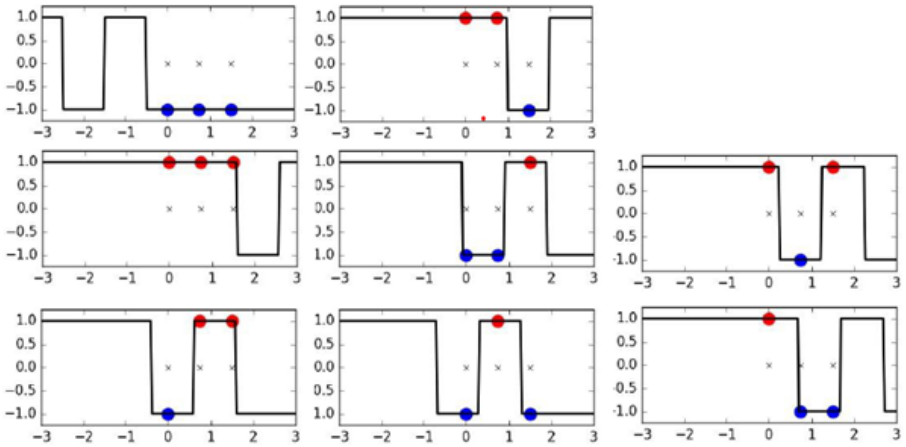
□

Teorema 2.4.2 *Sea* $\mathcal{F} = \{f(\cdot, t) \mid t \in \mathbb{R}\}$ *la clase de funciones con*

$$f(x;t) = \begin{cases} 1 & si\ x \in (-\infty, t] \cup [t+1, +\infty) \\ -1 & si\ x \in (t, t+1) \end{cases}$$

entonces $dim_{VC}(\mathbb{R}, \mathcal{F}) = 3$

DEMOSTRACIÓN: Esta función puede desmenuzar el siguiente conjunto de tres puntos



Sin embargo, no puede desmenuzar conjuntos de cuatro puntos, ya que no puede desmenuzar el subconjunto formado por el primer y tercer punto. Por lo tanto, la dimensión VC de $(\mathbb{R}, \mathcal{F})$ es 3. □

Capítulo 3

Aprendizaje PAC

En este capítulo introducimos el marco de aprendizaje estadístico **aprendizaje correcto probablemente aproximado**, o **aprendizaje PAC** por sus siglas en inglés. Para ello se han utilizado teoremas y demostraciones de [6] y [7].

Cuando se desarrollan y analizan algoritmos que aprenden a partir de ejemplos surgen preguntas fundamentales: ¿Qué puede ser aprendido? ¿Cómo de complejo es el aprendizaje? ¿Cómo influye el tamaño del conjunto de datos en los resultados? ¿Hay un modelo general de aprendizaje? Para responder a estas preguntas se introdujo el marco de aprendizaje PAC.

Valiant introdujo esta teoría en 1984 en [10] para que los científicos de la Computación estudiaran la eficiencia computacional de los algoritmos de aprendizaje. Aunando diferentes aproximaciones que se habían hecho hasta el momento sobre la idea de aprendizaje, consiguió una noción de problemas de aprendizaje que son resolubles, en el sentido de que existe un algoritmo que los resuelve en tiempo polinomial, en analogía a la clase **P** de la Teoría de la Complejidad clásica. Desde 1984 muchos informáticos teóricos e investigadores de aprendizaje automático han obtenido importantes resultados en esta teoría, convirtiendo el aprendizaje PAC en el marco de aprendizaje mas importante dentro de la teoría del aprendizaje computacional.

Como se ha comentado el marco PAC aporta definiciones de la clase de conceptos aprendibles en función del tamaño de la muestra de datos necesario para alcanzar una solución aproximada, y la complejidad del algoritmo en espacio y tiempo. En la siguiente sección se introducen los elementos fundamentales.

3.1. Definiciones básicas

En esta sección introducimos algunas terminologías básicas y notaciones que sirven de sustento en la Teoría Estadística del Aprendizaje.

Definición 3.1.1 *Se denomina* dominio o espacio de instancia *al conjunto X formado por las instancias o ejemplos que queremos etiquetar.*

Podemos pensar en X como el conjunto de codificaciones de los objetos a aprender. En un problema de reconocimiento de caracteres el espacio de instancia puede consistir en los vectores dos-dimensionales de píxeles binarios de una imagen de un tamaño dado. Otro ejemplo sería para el problema de encontrar un intervalo que clasifique correctamente los puntos dentro del intervalo como positivo y los puntos exteriores al rango como negativo, y en este caso $X = \mathbb{R}$.

El conjunto de posibles *etiquetas* o *valores objetivos* se denota por Y. Por ejemplo, en clasificación binaria se tiene que $Y = \{0, 1\}$, y en problemas de regresión $Y = \mathbb{R}$.

Definición 3.1.2 *Se denomina* conjunto de entrenamiento *a una muestra de instancias $x_i \in X$ para las que se conoce los valores correctos de sus etiquetas y_i.*

$$S = \{(x_1, y_1), \ldots, (x_m, y_m)\} \subseteq X \times Y.$$

Definición 3.1.3 *Un* concepto $c : X \longrightarrow Y$ *es una una función de X en Y.*

Si se considera $Y = \{0, 1\}$, se puede identificar un concepto c con el subconjunto $c \subseteq X$ de los elementos en los que toma el valor 1. La función c podría ser una variable aleatoria, lo que permite modelar incertidumbre en las predicciones (e.g. ruido en los datos). En reconocimiento de caracteres c puede ser el conjunto de todos los patrones en $X = \{0, 1\}^n$ que codifican una imagen de la letra "Q". En el problema del intervalo podría ser todos los números comprendidos entre el 0 y π.

Definición 3.1.4 *Una* clase de conceptos $\mathcal{C}$ *es un conjunto de conceptos sobre X.*

Cuando se trabaja en un problema de aprendizaje estadístico se busca un *concepto hipótesis h* entre los elementos de una *clase de conceptos hipótesis $\mathcal{H}$* que etiquete correctamente las instancias de X.

Definición 3.1.5 *Un* algoritmo de aprendizaje *es una función* $L : Z = X \times Y \longrightarrow \mathcal{H}$ *que recibe como entrada un conjunto de entrenamiento* S *y devuelve un concepto hipótesis* $h \in \mathcal{H}$.

Asumimos que los ejemplos están independiente e idénticamente distribuidos (i.i.d.) según una distribución desconocida $\mathcal{D}$ sobre la población. El problema de aprendizaje se formula de la siguiente manera. El algoritmo de aprendizaje considera un *conjunto hipótesis* $\mathcal{H}$ de posibles conceptos, recibe un conjunto de entrenamiento $S = \{(x_1, y_1), \ldots, (x_m, y_m)\}$ de elementos escogidos i.i.d. según $\mathcal{D}$ con probabilidad $P(x, y)$, dónde cada etiqueta $y_i = c(x_i)$ viene dada por un concepto objetivo c. El objetivo del algoritmo es obtener una hipótesis $h_S \in \mathcal{H}$ partiendo de S que se asemeje al concepto objetivo c. Cabe notar que si c es una variable aleatoria, entonces cada y_i no está es funcional (determinista) respecto a x_i, si no que siguen una variable aleatoria con probabilidad $P(y|x)$.

Para medir esa semejanza entre conceptos nos basaremos en el paradigma Minimización del Riesgo Empírico (también conocido por sus siglas en inglés *ERM*)

3.2. El paradigma ERM

La idea fundamental del principio ERM es que no se puede conocer con exactitud cómo de bien trabajará un algoritmo de aprendizaje en la práctica (el **error real**), pero sí se puede medir su desempeño para un conjunto de entrenamiento determinado (**error empírico**). La diferencia entre la función que queremos estimar y la hipótesis que nos aporta un algoritmo de aprendizaje se mide mediante una **función de pérdida**, que se elige dependiendo del problema al que nos enfrentemos. Esta función pérdida nos permitirá definir los errores reales y empíricos del concepto hipótesis.

Definición 3.2.1 *Dado un conjunto* Z *y un conjunto de funciones* $\mathcal{H}$*, se denomina* función de pérdida *o* función coste *a una función* $l : \mathcal{H} \times Z \longrightarrow \mathbb{R}_+$.

La función de pérdida mide como se ajusta un determinado algoritmo a un conjunto de datos, y se debe escoger de forma adecuada para cada problema concreto. Los siguientes dos ejemplo son las funciones más comunes.

En problemas de clasificación binaria $Z = X \times Y$, la salida del supervisor solo puede tomar dos valores, $Y = \{0, 1\}$ y $\mathcal{H}$ es un conjunto de funciones

indicatrices. Consideramos la siguiente función de pérdida:

$$l_{0,1}(h,(x,y)) = \begin{cases} 0, \text{ si } h(x) = y \\ 1, \text{ si } h(x) \neq y \end{cases}$$

En problemas de regresión, la salida del algoritmo tomará valores reales, $Y = \mathbb{R}$. Se puede considerar la siguiente función de pérdida:

$$l_{sq}(h,(x,y)) = (h(x) - y)^2$$

Esta es la denominada *función de pérdida cuadrática*, utilizada por ejemplo para técnicas de mínimos cuadrados.

Una vez definida una función de pérdida, se puede definir el **error real**.

Definición 3.2.2 *Dados una hipótesis $h \in \mathcal{H}$, una función error l, una distribución $\mathcal{D}$ sobre $Z = X \times Y$, el* error real *o* error de generalización *de h se define como*

$$R_{\mathcal{D}}(h) = \mathop{\mathbb{E}}_{z \sim \mathcal{D}} l(h,z) = \int l(h,z) dP(z)$$

Como se mencionó en la sección anterior, un algoritmo de aprendizaje recibe como entrada un conjunto de entrenamiento S, i.i.d. según una distribución de probabilidad $\mathcal{D}$ donde la etiqueta de cada instancia viene dada por un concepto $c : X \to Y$, y debe ofrecer de salida un predictor $h_S : X \to Y$. El objetivo del algoritmo es encontrar h_S que minimice el error respecto a $\mathcal{D}$ y c. Sin embargo, como $\mathcal{D}$ y c son desconocidas, el error real no está disponible para decidir la idoneidad del algoritmo. Por ello para evaluar la bondad del concepto hipótesis h se utiliza el denominado **error empírico** del predictor sobre el conjunto de entrenamiento.

Definición 3.2.3 *Dados una hipótesis $h \in \mathcal{H}$, una función error l y un conjunto de entrenamiento $S = \{z_i\}_i \subseteq Z$ con m puntos, el* error empírico *o* error de entrenamiento *de h sobre S se define como*

$$\widehat{R}_S(h) = \frac{1}{m} \sum_i l(h,z_i)$$

Así pues, el error empírico de $h \in \mathcal{H}$ es el error medio sobre la muestra S, mientras que el error real es el coste esperado de un ejemplo en Z escogido según $\mathcal{D}$.

El paradigma ERM considera al conjunto de entrenamiento S como una muestra representativa del conjunto $Z = X \times Y$ sobre el que está definido el concepto, por lo que es natural buscar una solución que tenga buenos resultados en los datos disponibles, minimizando el error empírico sobre S para tratar así de minimizar el error real que recordemos que no conocemos.

En posteriores secciones se muestran varios resultados que relacionan ambos errores bajo ciertas condiciones generales, y estudiaremos propiedades para discernir la representatividad del conjunto de entrenamiento sobre la población total. Por el momento solo mostramos el siguiente resultado sobre el error empírico esperado.

Proposición 3.2.1 *Sea $h \in \mathcal{H}$ fijo, la esperanza del error empírico en una muestra finita $S \subset Z$ escogida según $\mathcal{D}$ es igual al error real*

$$\mathop{\mathbb{E}}_{S\sim\mathcal{D}^m} \widehat{R}_S(h) = R_{\mathcal{D}}(h).$$

Demostración: Utilizando la linealidad de $\mathbb{E}$ y que los datos son i.i.d.

$$\mathop{\mathbb{E}}_{S\sim\mathcal{D}^m}[\widehat{R}_S(h)] = \mathop{\mathbb{E}}_{S\sim\mathcal{D}^m}\left[\frac{1}{m}\sum_{z\in S} l(h,z)\right]$$

$$= \frac{1}{m}\sum_{i}^{m} \mathop{\mathbb{E}}_{z_i\sim\mathcal{D}}[l(h,z_i)] = \mathop{\mathbb{E}}_{z\sim\mathcal{D}^m}[l(h,z)] = R_{\mathcal{D}}(h)$$

□

3.3. Clases aprendible PAC

Como mencionamos al comienzo del capítulo, el marco de aprendizaje estadístico PAC nos proporciona una forma de discriminar los conceptos que son aprendibles. En las siguientes secciones definiremos las clases de conceptos *aprendibles PAC*, diferenciando dos situaciones: si el concepto objetivo es determinista o no lo es.

- **Aprendizaje PAC**: El concepto objetivo c es una función determinista y el conjunto de entrenamiento i.i.d. S se elige siguiendo una distribución $\mathcal{D}$ sobre X.

- **Aprendizaje PAC agnóstico**: El concepto objetivo c es una función probabilística y el conjunto de entrenamiento i.i.d. S se elige siguiendo una distribución $\mathcal{D}$ sobre $Z = X \times Y$.

El aprendizaje agnóstico es más general y modela numerosos escenarios reales en los que puede haber instancias con más de una única etiqueta correcta. Por ejemplo, si tratamos de predecir el género de una persona en función de su peso y altura, ya que el mismo par de valores de peso y altura puede tener asignado tanto la etiqueta hombre como mujer en una misma población.

La intención de este modelo es que el aprendizaje exitoso de un concepto debería de permitir obtener, con alta probabilidad, una hipótesis que es una buena aproximación.

En concreto, un concepto objetivo $c \in \mathcal{C}$ será *aprendible en el sentido PAC* si existe un algoritmo de aprendizaje L que, dado un conjunto de aprendizaje suficientemente grande y $\epsilon, \delta > 0$, es capaz de producir con probabilidad 1-δ un concepto hipótesis h que sea aproximadamente correcto a c, i.e. error menor que ϵ.

En el caso de existir tal algoritmo L, la hipótesis h resultado de aplicar el algoritmo L es *probablemente aproximada* a c. De ahí el nombre del marco PAC.

Los parámetros ϵ y δ se denominan *parámetro de precisión y de confianza*, respectivamente. El primero, ϵ, estima cómo de lejos está la hipótesis obtenida de la óptima, y es necesario considerarlo para las clases aprendible PAC ya que puede que solo haya una muy pequeña probabilidad de que una muestra aleatoria pequeña vaya a distinguir entre dos hipótesis que difieren en pocos puntos muy improbables del espacio de instancias. El segundo parámetro, δ, acota la probabilidad de obtener una muestra poco representativa. Idealmente los dos parámetros deberían poder tomarse arbitrariamente pequeños sin gran coste.

3.3.1. Aprendizaje PAC para clasificación binaria

Comenzamos definiendo las clases aprendibles PAC en el caso más simple, el de clasificación binaria, y generalizaremos la definición para otros casos.

Para caracterizar el escenario determinista definimos el siguiente supuesto:

Definición 3.3.1 ***(Supuesto de realización)*** *Sea c un concepto objetivo,*

una clase de conceptos $\mathcal{H}$ *cumple el* supuesto de realización *si* $\exists\, h_* \in \mathcal{H}$ *tal que*

$$R_{\mathcal{D}}(h_*) = \mathop{\mathbb{E}}_{x\sim\mathcal{D}} l_{0,1}(h_*, (x, c(x))) = \mathop{\mathbb{P}}_{x\sim\mathcal{D}}[h_*(x) \neq c(x)] = 0$$

El supuesto de realización implica que para cada S tomado según $\mathcal{D}$ i.i.d. se tiene que, con probabilidad 1, $\widehat{R}_S(h_*) = 0$. Por lo tanto un algoritmo ERM devuelve un predictor h_S con $\widehat{R}_S(h_S) = 0$, anulando el error empírico. Sin embargo, lo que interesa realmente es minimizar el error real.

Definición 3.3.2 ***(Aprendizaje PAC para clasificación binaria)*** *Una clase de conceptos deterministas* $\mathcal{H}$ *es* aprendible PAC *si existe una función* $m_{\mathcal{H}} : (0,1)^2 \to \mathbb{N}$ *y un algoritmo de aprendizaje tal que* $\forall\, \epsilon, \delta \in (0,1)$*, para toda distribución de probabilidad* $\mathcal{D}$ *sobre* X*, y para todo concepto* $c : X \to \{0,1\}$*, si se cumple el supuesto de realización para* $\mathcal{H}, \mathcal{D}$ *y* c *entonces si se ejecuta el algoritmo sobre un conjunto de entrenamiento* S *con elementos i.i.d. por* $\mathcal{D}$ *y etiquetados por* c *con* $|S| \geq m_{\mathcal{H}}(\epsilon, \delta)$*, entonces*

$$P[R_{\mathcal{D}}(L(S)) \leq \epsilon] \geq 1 - \delta$$

Por lo tanto si una clase $\mathcal{H}$ es aprendible PAC se puede encontrar una hipótesis que *generaliza bien fuera del conjunto de entrenamiento*. Un algoritmo ERM devolverá una hipótesis que anula el error empírico y con error real acotado por ϵ con probabilidad δ.

La función $m_{\mathcal{H}}$ determina el tamaño necesario del conjunto de entrenamiento para que el algoritmo probablemente devuelva un concepto hipótesis h aproximado. Pueden existir multitud de funciones $m_{\mathcal{H}}$ que sean adecuadas. Por ello se escoge $m_{\mathcal{H}}$ de forma que sea el mínimo entero que satisface el requisito de aprendizaje PAC con precisión ϵ y confianza $1 - \delta$.

Ofrecemos ahora un resultado fundamental que acota el tamaño necesario de la muestra de entrenamiento para conceptos consistentes en el caso en el que la cardinalidad $|\mathcal{H}|$ es finita.

Teorema 3.3.1 ***(Tamaño de muestra - $\mathcal{H}$ finito, caso determinista)*** *Sea* $\mathcal{H}$ *un clase de conceptos consistentes de* X *a* Y*. Sea* L *un algoritmo tal que para todo* $c \in \mathcal{H}$ *y para todo conjunto de entrenamiento* S *devuelve un concepto hipótesis consistente* $h_S : \widehat{R}_S(h_S) = 0$*. Entonces,* $\forall\, \epsilon, \delta \in (0,1)$*, se satisface* $\mathbb{P}_{S\sim\mathcal{D}^m}[R_{\mathcal{D}}(h_S) \leq \epsilon] \geq 1 - \delta$ *si*

$$m \geq \frac{1}{\epsilon} log\left(\frac{|\mathcal{H}|}{\delta}\right)$$

Demostración: Para $\epsilon > 0$ se define

$$\mathcal{H}_\epsilon = \{h \in \mathcal{H} : R_{\mathcal{D}}(h) > \epsilon\}$$

. La probabilidad de que un concepto $h \in \mathcal{H}_\epsilon$ sea consistente en un conjunto de entrenamiento S seleccionado i.i.d. según $\mathcal{D}$ se puede acotar superiormente $\mathbb{P}[\widehat{R}_S(h) = 0] \leq (1-\epsilon)^m$ de lo que se sigue

$$\mathbb{P}\left[\exists h \in \mathcal{H}_\epsilon : \widehat{R}_S(h) = 0\right] = \mathbb{P}\left[\widehat{R}_S(h_1) = 0 \vee \cdots \vee \widehat{R}_S(h_{|\mathcal{H}_\epsilon|}) = 0\right]$$

$$\leq \sum_{h \in \mathcal{H}_\epsilon} \mathbb{P}\left[\widehat{R}_S(h) = 0\right] \leq \sum_{h \in \mathcal{H}_\epsilon} (1-\epsilon)^m \leq |\mathcal{H}|(1-\epsilon)^m \leq |\mathcal{H}|e^{-m\epsilon} \leq \delta$$

Donde en la última desigualdad hemos utilizado que $m \geq \frac{1}{\epsilon} log\left(\frac{|\mathcal{H}|}{\delta}\right)$. Finalmente,

$$\mathbb{P}_{S \sim \mathcal{D}^m}[R_{\mathcal{D}}(h - S) \leq \epsilon] = 1 - \mathbb{P}\left[\exists h \in \mathcal{H}_\epsilon : \widehat{R}_S(h) = 0\right] \geq 1 - \delta$$

□

Una formulación equivalente del teorema establece un límite superior del error real.

Corolario 3.3.2 *Sea $\mathcal{H}$ un clase de conceptos consistentes de X a Y. Sea L un algoritmo tal que para todo $c \in \mathcal{H}$ y para todo conjunto de entrenamiento S devuelve un concepto hipótesis consistente $h_S : \widehat{R}_S(h_S) = 0$. Entonces, $\forall\, \epsilon, \delta \in (0,1)$, se tiene con probabilidad al menos $1 - \delta$*

$$R_{\mathcal{D}}(h_S) \leq \frac{1}{m} log\left(\frac{|\mathcal{H}|}{\delta}\right).$$

En consecuencia a estos resultados, en el caso consistente toda clase de hipótesis finita es aprendible PAC con complejidad de la muestra

$$m_{\mathcal{H}}(\epsilon, \delta) \leq \left\lceil \frac{1}{\epsilon} log\left(\frac{|\mathcal{H}|}{\delta}\right) \right\rceil$$

Hay clases de conceptos infinitas que también son aprendibles PAC. Más adelante mostraremos que lo que determina si una clase es aprendible no es su finitud, sino la dimensión VC.

3.3.2. Aprendizaje PAC agnóstico para clasificación binaria

Abordaremos ahora el caso en el que el concepto objetivo es una variable aleatoria sobre X, este es el caso en el que el conjunto de entrenamiento no es consistente, ya que la misma instancia x puede tener distintas etiquetas escogidas según una probabilidad $\mathbb{P}(y|x)$, y por lo tanto no se cumple el supuesto de realización (no existe una clase hipótesis que anule el error real).

Definición 3.3.3 ***(Aprendizaje PAC agnóstico para clasificación binaria)*** *Una clase de conceptos $\mathcal{H}$ es* aprendible PAC agnóstico *si existe una función $m_{\mathcal{H}} : (0,1)^2 \to \mathbb{N}$ y un algoritmo de aprendizaje tal que $\forall\, \epsilon, \delta \in (0,1)$, para toda distribución de probabilidad $\mathcal{D}$ sobre $Z = X \times \{0,1\}$, si se ejecuta el algoritmo sobre $m \geq m_{\mathcal{H}}(\epsilon, \delta)$ ejemplos i.i.d. por $\mathcal{D}$, entonces el algoritmo devuelve un concepto hipótesis h tal que, con probabilidad al menos $1 - \delta$ se tiene que $R_{\mathcal{D}}(h) \leq min_{h' \in \mathcal{H}} R_{\mathcal{D}}(h') + \epsilon$.*

Que una clase de hipótesis $\mathcal{H}$ sea aprendible PAC agnóstico no garantiza que el error pueda hacerse arbitrariamente pequeño con suficientes ejemplos de entrenamiento como en el caso determinista, pero sí garantiza que el error sea relativamente pequeño respecto al mejor predictor que se puede obtener.

3.3.3. Aprendizaje PAC agnóstico generalizado

Hasta ahora solo hemos considerado problemas de clasificación binaria. Trataremos ahora el caso general que comprende también problemas como clasificación multiclase y regresión. La diferencia fundamental respecto a las dos anteriores definiciones es la consideración de funciones pérdida.

Definición 3.3.4 ***(Aprendizaje PAC agnóstico)*** *Una clase de conceptos $\mathcal{H}$ es* aprendible PAC agnóstico *con respecto a un conjunto Z y una función de pérdida $l : \mathcal{H} \times Z \to \mathbb{R}_+$ si existe una función $m_{\mathcal{H}} : (0,1)^2 \to \mathbb{N}$ y un algoritmo de aprendizaje tal que $\forall\, \epsilon, \delta \in (0,1)$, para toda distribución de probabilidad $\mathcal{D}$ sobre Z, si se ejecuta el algoritmo sobre $m \geq m_{\mathcal{H}}(\epsilon, \delta)$ ejemplos i.i.d. por $\mathcal{D}$, entonces el algoritmo devuelve un concepto hipótesis h tal que, con probabilidad al menos $1 - \delta$ se tiene que*

$$R_{\mathcal{D}}(h) \leq min_{h' \in \mathcal{H}} R_{\mathcal{D}}(h') + \epsilon, \text{ donde } R_{\mathcal{D}}(h) = \mathop{\mathbb{E}}_{z \sim \mathcal{D}} l(h, z)$$

3.4. Convergencia uniforme

En la sección anterior se mostró que, bajo el supuesto de realización, toda clase de concepto finita es aprendible PAC. Ahora desarrollamos una herramienta, la **convergencia uniforme**, y la aplicamos para mostrar que toda clase finita es aprendible PAC agnóstico siempre que el rango de la función de pérdida esté acotado.

Recordemos que, en el marco ERM, se busca el concepto hipótesis $h \in \mathcal{H}$ que minimiza el riego empírico sobre el conjunto de entrenamiento S con la esperanza de que así se minimize el riesgo real sobre la distribución completa $\mathcal{D}$. Esto sucede si el riesgo empírico de los conceptos pertenecientes a la clase de hipótesis es una buena aproximación del riesgo real. Procedamos a formalizar esta idea.

Definición 3.4.1 *Se dice que un conjunto de entrenamiento S es ϵ- representativo respecto a un dominio Z, clase de hipótesis $\mathcal{H}$, función de pérdida l y distribución $\mathcal{D}$ si*

$$\forall h \in \mathcal{H}, \quad |\widehat{R}_S(h) - R_{\mathcal{D}}(h)| \leq \epsilon$$

Los conjuntos de entrenamientos que satisfacen esta condición garantizan que los estimadores obtenidos por un algoritmo ERM son una buena hipótesis, ya que al minimizar el riesgo empírico sobre estos conjuntos también se minimiza el riesgo real.

Lema 3.4.1 *Sea S un conjunto de entrenamiento $\epsilon/2$-representativo. Entonces toda salida de un algoritmo ERM sobre $\mathcal{H}$ que toma como entrada S, i.e. todo $h_S = argmin_{h \in \mathcal{H}} \widehat{R}_S(h)$, verifica:*

$$R_{\mathcal{D}}(h_S) \ \leq \ \min_{h \in \mathcal{H}} R_{\mathcal{D}}(h) + \epsilon$$

Demostración: Para todo $h \in \mathcal{H}$

$$R_{\mathcal{D}}(h_S) \ \leq \ \widehat{R}_S(h_S) + \frac{\epsilon}{2} \ \leq \ \widehat{R}_S(h) + \frac{\epsilon}{2} \ \leq \ R_{\mathcal{D}}(h) + \frac{\epsilon}{2} + \frac{\epsilon}{2} \ \leq \ R_{\mathcal{D}}(h) + \epsilon$$

□

Debido a este lema, para garantizar que $\mathcal{H}$ es aprendible PAC agnóstico, basta con que al escoger S según $\mathcal{D}$, el conjunto de entrenamiento sea ϵ-representativo con probabilidad al menos $1 - \delta$.

Definición 3.4.2 ***(Convergencia Uniforme)*** *Una clase de hipótesis $\mathcal{H}$ tiene la propiedad de* convergencia uniforme *(respecto a un dominio Z y una función pérdida l) si existe una función $m_{\mathcal{H}}^{UC} : (0,1)^2 \to \mathbb{N}$ que para cada par $\epsilon, \delta \in (0,1)$ y para cada distribución de probabilidad $\mathcal{D}$ sobre Z, si $S \subseteq Z$ es un subconjunto de $m \geq m_{\mathcal{H}}^{UC}(\epsilon, \delta)$ ejemplos escogidos i.i.d. según $\mathcal{D}$, entonces S es ϵ-representativo con probabilidad al menos $1 - \delta$.*

Al igual que $m_{\mathcal{H}}$ determinaba el tamaño mínimo de S para que un algoritmo devuelva probablemente un concepto hipótesis aproximado, la función $m_{\mathcal{H}}^{UC}$ mide el tamaño mínimo de S para que sea probablemente ϵ-representativo.

El siguiente corolario se sigue directamente del Lema 3.4.1.

Corolario 3.4.2 *Si una clase $\mathcal{H}$ tiene la propiedad de convergencia uniforme con función $m_{\mathcal{H}}^{UC}$, entonces la clase es aprendible PAC agnóstico con complejidad de muestra $m_{\mathcal{H}}(\epsilon, \delta) \leq m_{\mathcal{H}}^{UC}(\frac{\epsilon}{2}, \delta)$. Por lo tanto, el paradigma $ERM_{\mathcal{H}}$ es un aprendiz PAC agnóstico para $\mathcal{H}$.*

3.5. Las clases finitas son aprendibles PAC agnóstico

En esta sección demostraremos que todas las clases de conceptos finitas tienen la propiedad de convergencia uniforme y por lo tanto son aprendibles PAC agnóstico, siempre que la función de pérdida considerada tenga rango acotado.

Teorema 3.5.1 *Sea $\mathcal{H}$ una clase de hipótesis finita, sea Z el dominio, y $l : \mathcal{H} \times Z \to [0,1]$ una función de pérdida. Entonces, $\mathcal{H}$ tiene la propiedad de convergencia uniforme con*

$$m_{\mathcal{H}}^{UC}(\epsilon, \delta) \leq \left\lceil \frac{log(2|\mathcal{H}|/\delta)}{2\epsilon^2} \right\rceil$$

Además, la clase es aprendible PAC agnóstico con complejidad de muestra

$$m_{\mathcal{H}}(\epsilon, \delta) \leq m_{\mathcal{H}}^{UC}(\frac{\epsilon}{2}, \delta) \leq \left\lceil \frac{2log(|\mathcal{H}|/\delta)}{\epsilon^2} \right\rceil$$

Para la demostración del teorema se utilizará el siguiente lema.

Lema 3.5.2 ***(desigualdad de Hoeffding)****: Sean $\theta_1, \ldots, \theta_m$ una secuencia de variables aleatorias i.i.d tales que $\forall i\ \mathbb{E}(\theta_i) = \mu$ y $\mathbb{P}[a \leq \theta_i \leq b] = 1$. Entonces $\forall \epsilon > 0$ se tiene que*

$$\mathbb{P}\left[\left|\frac{1}{m}\sum_{i=1}^{m}\theta_i - \mu\right| > \epsilon\right] \leq 2exp\left(\frac{-2m\epsilon^2}{(b-a)^2}\right)$$

Demostración: Hay que encontrar m tal que para toda distribución de probabilidad $\mathcal{D}$ sobre Z, al tomar una muestra $S = \{z_1, \ldots, z_m\}$ de m elementos i.i.d. según $\mathcal{D}$ se tenga con probabilidad al menos $1 - \delta$ que

$$\forall h \in \mathcal{H},\ |\widehat{R}_S(h) - R_{\mathcal{D}}(h)| \leq \epsilon.$$

Tenemos

$$\mathcal{D}^m(\{S\ :\ \forall h \in \mathcal{H},\ |\widehat{R}_S(h) - R_{\mathcal{D}}(h)| \leq \epsilon\}) \geq 1 - \delta$$
$$\iff \mathcal{D}^m(\{S\ :\ \exists h \in \mathcal{H},\ |\widehat{R}_S(h) - R_{\mathcal{D}}(h)| > \epsilon\}) \leq \delta$$

Como

$$\{S\ :\ \exists h \in \mathcal{H},\ |\widehat{R}_S(h) - R_{\mathcal{D}}(h)| > \epsilon\} = \bigcup_{h\in\mathcal{H}}\{S\ :\ |\widehat{R}_S(h) - R_{\mathcal{D}}(h)| > \epsilon\}$$

Se tiene entonces que

$$\mathcal{D}^m(\{S\ :\ \exists h \in \mathcal{H}, |\widehat{R}_S(h) - R_{\mathcal{D}}(h)| > \epsilon\}) \leq \sum_{h\in\mathcal{H}}\mathcal{D}^m(\{S\ :\ |\widehat{R}_S(h) - R_{\mathcal{D}}(h)| > \epsilon\})$$

Por la proposición 3.2.1 se tiene que $\mathbb{E}_{S\sim\mathcal{D}^m}\widehat{R}_S(h) = R_{\mathcal{D}}(h)$ y tanto $|\widehat{R}_S(h) - R_{\mathcal{D}}(h)|$ es la desviación del error empírico a su valor esperado.

Sea la variable aleatoria $\theta_i = l(h, z_i)$ para h fijo. Como z_i se han tomado i.i.d., θ_i también son i.i.d. y se tiene que $\widehat{R}_S(h) = \frac{1}{m}\sum_{i=1}^{m} l(h, z_i) = \frac{1}{m}\sum_{i=1}^{m}\theta_i$ y $R_{\mathcal{D}}(h) = \mu$. Además como $rango(l) = [0, 1]$ entonces $\theta_i \in [0, 1]$. Por lo tanto con la desigualad de Hoeffding

$$\{S\ :\ |\widehat{R}_S(h) - R_{\mathcal{D}}(h)| > \epsilon\} = \mathbb{P}\left[\left|\frac{1}{m}\sum_{i=1}^{m}\theta_i - \mu\right| > \epsilon\right] \leq 2exp\left(-2m\epsilon^2\right)$$

Combinado con las inecuaciones anteriores llegamos a

$$\mathcal{D}^m(\{S\ :\ \exists h \in \mathcal{H},\ |\widehat{R}_S(h) - R_{\mathcal{D}}(h)| > \epsilon\})$$
$$\leq \sum_{h\in\mathcal{H}} 2exp\left(-2m\epsilon^2\right) = 2|\mathcal{H}|exp(-2m\epsilon^2).$$

Luego si tomamos

$$m \geq \frac{log\left(\frac{2|\mathcal{H}|}{\delta}\right)}{\epsilon^2}$$

entonces tenemos que

$$\mathcal{D}^m(\{S \,:\, \exists h \in \mathcal{H},\, |\widehat{R}_S(h) - R_{\mathcal{D}}(h)| > \epsilon\}) < \delta$$

y en consecuencia las clases finitas tienen la propiedad de convergencia uniforme.

Por último, utilizando el corolario 3.4.2, tenemos que, además, la clase de hipótesis finita $\mathcal{H}$ es aprendible PAC Agnóstico con complejidad de muestra:

$$m_{\mathcal{H}}(\epsilon, \delta) \leq m_{\mathcal{H}}^{UC}(\frac{\epsilon}{2}, \delta) \leq \left\lceil \frac{2log(|\mathcal{H}|/\delta)}{\epsilon^2} \right\rceil.$$

□

La propiedad de convergencia uniforme de una clase de hipótesis implica que si el conjunto de entrenamiento es suficientemente grande, entonces los errores empíricos de las hipótesis de la clase son similares a los errores reales. Utilizando algoritmos ERM, la convergencia uniforme basta para asegurar el aprendizaje PAC agnóstico en clases de hipótesis finitas. Se puede encontrar información mas detallada en la referencia [5]. Con este resultado finalizamos el capítulo de introducción al Aprendizaje PAC.

Capítulo 4

El Teorema Fundamental del Aprendizaje Estadístico

Hasta ahora hemos definido las ideas básicas del aprendizaje estadístico, el concepto de aprendizaje PAC y la propiedad de convergencia uniforme. Recordemos que una clase de funciones $\mathcal{H}$ es aprendible PAC si dado una muestra de entrenamiento finita con elementos escogidos i.i.d. según una distribución de probabilidad $\mathcal{D}$, se puede encontrar un concepto hipótesis h que minimize el error real respecto al mejor estimador en $\mathcal{H}$.

En el marco de aprendizaje estadístico ERM, los algoritmos tratan de minimizar el error empírico sobre el conjunto de entrenamiento con la intención de minimizar también el error real sobre la distribución.

Una vez presentadas estas ideas surge de manera natural la pregunta de qué clases de conceptos son aprendibles PAC y qué las caracteriza. Ya demostramos que todas las clases de hipótesis finitas son aprendibles, mediante un algoritmo ERM, si tenemos una muestra de $O\left(\frac{2log(|\mathcal{H}|/\delta)}{\epsilon^2}\right)$ ejemplos, pero de momento no mostramos nada sobre las que son infinitas. El resultado más importante sobre esta cuestión nos indica que las clases aprendibles no se caracterizan por su tamaño, sino por su dimensión Vapnik-Chervonenkis.

En este capítulo estableceremos en primer lugar el **teorema No-Free-Lunch (NFL)** [4], que establece que no existe un algoritmo de aprendizaje universal, es decir, un algoritmo que pueda aprender bien cualquier clase de problema, sea cual sea la distribución de probabilidad de los datos. En tareas de clasificación, para cada algoritmo de aprendizaje, siempre existirá una distribución de probabilidad $\mathcal{D}$ (que genera los datos entrada-salida) en la que *falle*.

Existen numerosas versiones de este teorema, la que se presenta en este trabajo proviene de [9].

A continuación trataremos la **dimensión VC** en el contexto de aprendizaje PAC y su relación con algoritmos de aprendizaje, finalizando con el resultado conocido como **Teorema Fundamental del Aprendizaje Estadístico (FTSL)** que relaciona la aprendibilidad de una clase de hipótesis con su dimensión VC y en un sentido muy fuerte: una clase es aprendible PAC si y solo si su dimensión VC es finita.

4.1. Teorema No-Free-Lunch

Teorema 4.1.1 ***(No-Free-Lunch)*** *Sea L un algoritmo de aprendizaje ERM de clasificación binaria para la función pérdida $l_{0,1}(h,(x,y)) = I(h(x) = y)$, con I función indicatriz, sobre un dominio X. Sea $m \in \mathbb{N}$, $m \leq |X|/2$. Entonces existe una distribución de probabilidad $\mathcal{D}$ sobre $X \times \{0,1\}$ tal que:*

- *Existe un concepto $f : X \to \{0,1\}$ con $R_{\mathcal{D}}(f) = 0$.*
- *Con probabilidad al menos $1/7$ sobre un conjunto de entrenamiento $S \sim \mathcal{D}^m$ se tiene que $R_{\mathcal{D}}(L(S)) \geq 1/8$.*

El teorema NFL indica que ningún algoritmo de aprendizaje puede ser exitoso para todos los problemas. La formalización de esta idea es que, para todo algoritmo de aprendizaje, existe una distribución de probabilidad para la cual el algoritmo falla. Es decir que existen ϵ, δ para los cuales el algoritmo no puede devolver con probabilidad $1 - \delta$ una hipótesis que acote el error según ϵ, y esto se cumple aun cuando otro algoritmo sí es exitoso. Un ejemplo trivial es el de un algoritmo ERM sobre la clase de hipótesis $\mathcal{H} = \{f\}$, y mas generalmente, uno con una clase de hipótesis que contenga a f y con un conjunto de entrenamiento de tamaño $m \geq 8log(7|\mathcal{H}|/6)$ (ver teorema 3.3.1).

Antes de demostrar el teorema NFL consideremos el siguiente lema.

Lema 4.1.2 *Sea Z una variable aleatoria que toma valores en $[0,1]$ tal que $\mathbb{E}[Z] = \mu$. Entonces, para $a \in (0,1)$ se tiene que*

$$\mathbb{P}[Z > 1 - a] \geq \frac{\mu - (1 - a)}{a}$$

Esto también implica que para todo $a \in (0,1)$

$$\mathbb{P}[Z > a] \geq \frac{\mu - a}{1 - a} \geq \mu - a$$

DEMOSTRACIÓN: Sea $Y = 1 - Z$, que es una variable aleatoria no negativa con $\mathbb{E}[Y] = 1 - \mu$. Aplicando la desigualdad de Markov se tiene:

$$\mathbb{P}[Z \leq 1 - a] \geq \mathbb{P}[1 - Z \geq 1 - a] = \mathbb{P}[Y \geq a] \leq \frac{\mathbb{E}[Y]}{a} = \frac{1 - \mu}{a}$$

Y por lo tanto

$$\mathbb{P}[Z > 1 - a] \geq 1 - \frac{1 - \mu}{a} = \frac{a + \mu - 1}{a}.$$

Para la segunda inecuación, basta considerar $1 - a$ en lugar de a. □

Demostremos ahora el teorema NFL 4.1.1:

DEMOSTRACIÓN: Sea $C \subseteq X$ un subconjunto del dominio con tamaño $2m$ y sean $f_1, \ldots, f_T$ con $T = 2^{2m}$ todas las posibles funciones de C a $\{0, 1\}$. Para cada $i = 1, \ldots, T$ sea $\mathcal{D}_i$ la distribución sobre X definida como

$$\mathcal{D}_i(\{(x, y)\}) = \begin{cases} 1/|C| & \text{si } y = f_i(x) \\ 0 & \text{en otro caso} \end{cases}$$

que es una distribución uniforme sobre los pares (x, y) tal que $y = f_i(x)$ y se cumple $R_{\mathcal{D}_i}(f_i) = \underset{(x,y)\sim\mathcal{D}_i}{\mathbb{E}} l_{0,1}(f_i, (x, y)) = 0$.

Sea L un algoritmo de aprendizaje que recibe de entrada un conjunto S de m ejemplos de $C \times \{0, 1\}$ y devuelve como salida un concepto $L(S) : C \to \{0, 1\}$. Denotamos por $S_1, \ldots, S_k$ las $k = (2m)^m$ posibles secuencias de m ejemplos de C.

Para $S_j = (x_1, \ldots, x_m)$ denotamos por $S_j^i = ((x_1, f_i(x_1)), \ldots, (x_m, f_i(x_m)))$ la secuencia de los pares de instancias en S_j con su etiqueta según f_i. Considerando la distribución $\mathcal{D}_i$ los posibles conjuntos de entrenamiento que L puede recibir son $S_1^i, \ldots, S_k^i$ con la misma probabilidad de ser escogidos. Por lo tanto

$$\underset{S\sim\mathcal{D}_i^m}{\mathbb{E}}[R_{\mathcal{D}_i}(L(S))] = \frac{1}{k}\sum_{j=1}^{k} R_{\mathcal{D}_i}(L(S_j^i)) \tag{4.1}$$

Utilizando que el máximo de una muestra es mayor que su media, y que la media es, a su vez, mayor que su mínimo, tenemos que

$$\begin{aligned}\max_{1\leq i\leq T}\frac{1}{k}\sum_{j=1}^{k}R_{\mathcal{D}_i}(L(S_j^i)) &\geq \frac{1}{T}\sum_{i=1}^{T}\frac{1}{k}\sum_{j=1}^{k}R_{\mathcal{D}_i}(L(S_j^i))\\ &= \frac{1}{k}\sum_{j=1}^{k}\frac{1}{T}\sum_{i=1}^{T}R_{\mathcal{D}_i}(L(S_j^i))\\ &\geq \min_{1\leq j\leq k}\frac{1}{T}\sum_{i=1}^{T}R_{\mathcal{D}_i}(L(S_j^i)) \qquad (4.2)\end{aligned}$$

Dada $j \in \{1,\ldots,k\}$, sean $v_i,\ldots,v_p$ las instancias en C que no aparecen en S_j. Se tiene que $p \geq m$. Por lo tanto, para toda función $h : C \to \{0,1\}$ y para todo i se tiene:

$$\begin{aligned}R_{\mathcal{D}_i}(h) &= \frac{1}{2m}\sum_{x\in C}I_{[h(x)\neq f_i(x)]}\\ &\geq \frac{1}{2m}\sum_{r=1}^{p}I_{[h(v_r)\neq f_i(v_r)]}\\ &\geq \frac{1}{2p}\sum_{r=1}^{p}I_{[h(v_r)\neq f_i(v_r)]} \qquad (4.3)\end{aligned}$$

Y por lo tanto,

$$\begin{aligned}\frac{1}{T}\sum_{i=1}^{T}R_{\mathcal{D}_i}(L(S_j^i)) &\geq \frac{1}{T}\sum_{i=1}^{T}\frac{1}{2p}\sum_{r=1}^{p}I_{[A(S_j^i)(v_r)\neq f_i(v_r)]}\\ &= \frac{1}{2p}\sum_{r=1}^{p}\frac{1}{T}\sum_{i=1}^{T}I_{[A(S_j^i)(v_r)\neq f_i(v_r)]}\\ &\geq \frac{1}{2}\min_{1\leq r\leq p}\frac{1}{T}\sum_{i=1}^{T}I_{[A(S_j^i)(v_r)\neq f_i(v_r)]} \qquad (4.4)\end{aligned}$$

Fijado $r \in \{1,\ldots,p\}$, podemos particionar las funciones f_i para $i \leq T$ en $T/2$ pares disjuntos $(f_i, f_{i'})$ tales que para cada $c \in C$, $f_i(c) \neq f_{i'}(c)$ si y solo si $c = v_r$. Como para estos pares se debe tener que $S_j^i = S_j^{i'}$, se tiene que

$$I_{[A(S_j^i)(v_r)\neq f_i(v_r)]} + I_{[A(S_j^{i'})(v_r)\neq f_i(v_r)]} = 1.$$

por lo que

$$\frac{1}{T}\sum_{i=1}^{T} I_{[A(S_j^i)(v_r)\neq f_i(v_r)]} = \frac{1}{2}.$$

Combinando esto con las ecuaciones (4.1), (4.2), (4.3) y (4.4), hemos demostrado que para L se cumple que

$$\underset{1\leq i\leq T}{max} \underset{S\sim\mathcal{D}_i^m}{\mathbb{E}}[R_{\mathcal{D}_i}(L(S))] \geq 1/4.$$

Por lo tanto, para todo algoritmo de aprendizaje L' que recibe un conjunto de entrenamiento de tamaño m de $X \times \{0,1\}$ existe una función $f : X \to \{0,1\}$ y una distribución $\mathcal{D}$ sobre $X \times \{0,1\}$, tal que $R_{\mathcal{D}}(f) = 0$ y

$$\underset{S\sim\mathcal{D}^m}{\mathbb{E}}[R_{\mathcal{D}}(L'(S))] \geq 1/4.$$

Se tiene que $R_{\mathcal{D}}(L(S)) = \underset{z\sim\mathcal{D}}{\mathbb{E}}[l_{0,1}(L(S), z)] \in [0,1]$. Aplicamos ahora el lema 4.1.2, obteniendo que

$$\begin{aligned}\mathbb{P}[R_{\mathcal{D}}(L(S)) \geq 1/8] &\geq \frac{\underset{S\sim\mathcal{D}^m}{\mathbb{E}}[R_{\mathcal{D}}(L(S))] - 1/8}{7/8} \\ &\geq \frac{1/4 - 1/8}{7/8} = \frac{1}{7}\end{aligned}$$

Lo que termina la prueba. □

Desde que definimos las clases de conceptos siempre hemos supuesto que un algoritmo de aprendizaje devuelve una hipótesis que pertenece a un conjunto hipótesis $\mathcal{H}$. Esto se puede interpretar como que se supone que el algoritmo posee conocimiento a priori - el conocimiento de que uno de los miembros de $\mathcal{H}$ es una buena aproximación del concepto objetivo.

En el caso de no hacer ninguna suposición sobre como es la solución, el conjunto $\mathcal{H}$ considerado es la clase de todas funciones que tienen como dominio X infinito y rango $\{0,1\}$. El siguiente corolario expresa esta necesidad de restringir el conjunto hipótesis para que pueda ser aprendible.

Corolario 4.1.3 *Sea X un dominio infinito y $\mathcal{H}$ el conjunto de todas funciones de X en $\{0,1\}$. Entonces $\mathcal{H}$ no es aprendible PAC.*

DEMOSTRACIÓN: Asumamos, por reducción al absurdo, que $\mathcal{H}$ es aprendible. Sean $\epsilon < 1/8$ y $\delta < 1/7$. Que $\mathcal{H}$ sea PAC aprendible implica que debe existir un algoritmo L y un entero $m = m(\epsilon, \delta)$, tal que para toda distribución $\mathcal{D}$ sobre $X \times \{0,1\}$, si existe una función $f : X \to \{0,1\}$ con $R_{\mathcal{D}}(f) = 0$, entonces al darle a L un conjunto S de tamaño m escogido según $\mathcal{D}$, se verifica que $R_{\mathcal{D}}(L(S)) \leq \epsilon$ con probabilidad al menos $1 - \delta$.

Sin embargo, como $|X| > 2m$, aplicando el teorema NFL para todo algoritmo de aprendizaje existe una distribución $\mathcal{D}$ tal que $R_{\mathcal{D}}(L(S)) > 1/8 > \epsilon$ con probabilidad mayor que $1/7 = \delta$. □

Hemos visto en esta sección que no existe un algoritmo universal de aprendizaje que sea exitoso en todo problema que se le presente. Para que un algoritmo de clasificación sea aprendible PAC hay que establecer unas presunciones mínimas sobre la solución.

4.2. Las clases de hipótesis infinitas pueden ser aprendibles

En el estudio de las clases de hipótesis que son aprendibles PAC hemos demostrado que todas las clases finitas son aprendibles si se aporta un conjunto de entrenamiento de tamaño suficiente, mientras que la clase de todas las funciones de X infinito en $\{0,1\}$ no lo es. Podríamos pensar que la condición de ser aprendible PAC está caracterizada por la finitud o el tamaño de la clase, pero no es así, existen clases infinitas que pueden ser aprendibles.

Lema 4.2.1 *Sea $\mathcal{H}$ la clase de funciones umbral sobre los números reales, $\mathcal{H} = \{h_a : a \in \mathbb{R}\}$, con $h_a : \mathbb{R} \to \{0,1\}$ definida como $I_{[x<a]}$. Entonces $\mathcal{H}$ es PAC aprendible, con $m_{\mathcal{H}}(\epsilon, \delta) \leq \lceil log(2/\delta)/\epsilon \rceil$.*

DEMOSTRACIÓN: Sea $a^\star \in \mathbb{R}$ tal que el concepto objetivo $h^\star = I_{[x<a\star]}$. Sea $\mathcal{D}$ la distribución de probabilidad sobre $\mathbb{R}$ y sean $a_0 < a^\star < a_1$ tales que

$$\mathbb{P}_{x\sim\mathcal{D}}[x \in (a_0, a^\star)] = \mathbb{P}_{x\sim\mathcal{D}}[x \in (a^\star, a_1)] = \epsilon$$

(Si $\mathcal{D}(-\infty, a^\star) \leq \epsilon$ tomamos $a_0 = \infty$ y similarmente para a_1). Dado un conjunto de entrenamiento $S \subset \mathbb{R} \times \{0,1\}$, definimos

$$b_0 = max\{x : (x,1) \in S\}, b_1 = min\{x : (x,0) \in S\}$$

(Si no hay ejemplos positivos en S tomamos $b_0 = -\infty$ y si no hay ejemplos negativos $b_1 = \infty$). Sea h_S la hipótesis del algoritmo ERM asociada a S y sea b_S su umbral, por lo que b_S $in(b_0, b_1)$. Se cumple

$$\begin{aligned} \underset{S\sim\mathcal{D}}{P}[R_{\mathcal{D}}(h_S) > \epsilon] &\leq \underset{S\sim\mathcal{D}}{P}[b_0 < a_0 \vee b_1 > a_1] \\ &\leq \underset{S\sim\mathcal{D}}{P}[b_0 < a_0] + \underset{S\sim\mathcal{D}}{P}[b_1 > a_1] \end{aligned} \tag{4.5}$$

El caso $b_0 < a_0$ sucede si y solo si no hay ningún caso en S en el intervalo $(a_0, a^\star)$, que tiene probabilidad definida como ϵ

$$\underset{S\sim\mathcal{D}}{P}[b_0 < a_0] = \underset{S\sim\mathcal{D}}{P}[\forall(x,y) \in S, x \notin (a_0, a^\star)] = (1-\epsilon)^m \leq e^{-\epsilon m} \leq \delta/2$$

En la última desigualdad se usa que $m > log(2/\delta)/\epsilon$. De la misma manera, se puede probar que $\underset{S\sim\mathcal{D}}{P}[b_1 > a_1] \leq \delta/2$. Combinando esto con la ecuación (4.5), terminamos la demostración. □

4.3. Dimensión VC y aprendizaje PAC

Hemos visto que la finitud de una clase de conceptos no determina si es aprendible PAC. Veremos ahora que la característica que lo determina es su dimensión Vapnik-Chervonenkis.

Recordemos que en el Capítulo 1 definimos la dimensión VC de un espacio de aprendizaje $(U, \mathcal{F})$, con $U = \mathbb{R}^d$ y y $\mathcal{F}$ una familia de funciones de $\mathbb{R}^d$ en $\mathbb{R}$, como la cardinalidad del mayor conjunto de puntos en U que puede ser desmenuzado por alguna función de $\mathcal{F}$, i.e. la mayor nube de puntos $X \subset U$ tal que $\forall A \subseteq X \quad \exists f \in \mathcal{F}$ que satisface $\forall x \in X$ $(I_A(x) = 1 \Leftrightarrow f(x) > 0)$.

Tratando el problema de clasificación binaria la familia de funciones de interés será $\mathcal{H}$ de funciones con dominio X e imagen $\{0, 1\}$. Todos los conceptos y definiciones del primer capítulo son fácilmente aplicables a este caso.

Se estableció anteriormente (en el teorema NFL) que cuando no se imponen restricciones sobre la clase de hipótesis $\mathcal{H}$, entonces existe una distribución de probabilidad para la cuál una función tiene error nulo mientras que el algoritmo

falla para esa distribución. En la prueba del teorema se tomaban todos las posibles funciones binarias sobre un subconjunto finito $C \subset X$ y para cada una de ellas se consideraba una distribución con probabilidad concentrada en C. Para mostrar que el algoritmo fallaba se utilizaba la capacidad de poder escoger una función objetivo entre la clase de todas las posibles funciones de C a $0, 1$.

Cuando una clase de hipótesis $\mathcal{H}$ desmenuza algún conjunto $C \subset X$ lo que tenemos es que, efectivamente, $\mathcal{H}$ no tiene ninguna restricción cuando la consideramos sobre C. Esto nos lleva al siguiente corolario del teorema NFL.

Corolario 4.3.1 *Sea $\mathcal{H}$ una clase de hipótesis de X a $\{0, 1\}$. Sea m el tamaño del conjunto de entrenamiento. Si existe un conjunto $C \subset X$ de tamaño $2m$ que es desmenuzado por $\mathcal{H}$, entonces para todo algoritmo de aprendizaje L existen una distribución $\mathcal{D}$ sobre $X \times \{0, 1\}$ y una hipótesis $h \in \mathcal{H}$ tal que $R_{\mathcal{D}}(h) = 0$ pero para la que se tiene con probabilidad al menos $1/7$ sobre $S \sim \mathcal{D}^m$ que $R_{\mathcal{D}}(L(S)) \geq 1/8$.*

Este corolario nos relaciona intuitivamente el aprendizaje PAC con la definición de dimensión VC. Si $\mathcal{H}$ desmenuza algún conjunto C de tamaño $2m$ entonces no se puede aprender $\mathcal{H}$ con m ejemplos, mientras que $dim_{VC}(X, \mathcal{H})$ es el máximo de los tamaños de los conjuntos que desmenuza $\mathcal{H}$. Una consecuencia directa del corolario es el siguiente teorema:

Teorema 4.3.2 *Sea $\mathcal{H}$ una clase de funciones con dimensión VC infinita. Entonces $\mathcal{H}$ no es aprendible PAC.*

Demostración: Como $\mathcal{H}$ tiene dimensión VC infinita, para cualquier conjunto de entrenamiento de tamaño m existe un subconjunto $C \subset X$ se tamaño $2m$ desmenuzado por $\mathcal{H}$. La conclusión sigue de aplicar el corolario 4.3.1. □

4.4. El Teorema Fundamental del Aprendizaje Estadístico

Ya hemos mostrado que las clases de hipótesis con dimensión VC infinita no son aprendible PAC. El teorema fundamental del aprendizaje estadístico afirma que esa relación se cumple en ambos sentidos: si la clase tiene dimensión VC finita entonces será aprendible PAC.

Teorema 4.4.1 ***Teorema Fundamental del Aprendizaje Estadístico*** *Sea $\mathcal{H}$ una clase de hipótesis de un dominio X al conjunto $\{0,1\}$ y consideremos la función pérdida $l_{0,1}$. Las siguientes afirmaciones son equivalentes:*

1. *$\mathcal{H}$ tiene la propiedad de convergencia uniforme.*
2. *$\mathcal{H}$ es aprendible PAC agnóstico.*
3. *$\mathcal{H}$ es aprendible PAC.*
4. *$(X, \mathcal{H})$ tiene dimensión VC finita.*

Dos versiones ligeramente distintas del teorema, que aporta límites numéricos al tamaño necesario de la muestra, pueden encontrarse en [9] y [2].

Demostración: En el tema 3 se demostró (1) $\Rightarrow$ (2) con el Corolario 3.4.2. La implicación (2) $\Rightarrow$ (3) es trivial. La implicación (3) $\Rightarrow$ (4) es consecuencia del teorema NFL como se especificó en al apartado anterior concluyendo con el Teorema 4.3.2. La parte complicada es (4) $\Rightarrow$ (1). En el resto de este apartado se presentará una serie de resultados teóricos que, junto con los resultados numéricos del anexo, concluirán con la demostración. □

Recordemos que en el capítulo 1 se definió la *función de granularidad máxima* de tamaño n para un espacio de aprendizaje $(X, \mathcal{H})$ como el mayor número de subconjuntos de algún $A \subseteq X$ de tamaño n que pueden ser aprendidos por $\mathcal{H}$

$$\pi_{(X,\mathcal{H})}(n) = max\{|A|_{\mathcal{H}} \mid A \subseteq X \; y \; |A| = n\}.$$

El lema 1.2.1 establece una cota para la granularidad máxima. Al combinarlo con el Apéndice .0.7, se puede reformular de la siguiente forma:

Lema 4.4.2 *Sea $\mathcal{H}$ una clase de hipótesis con dominio X y $dim_{VC}(X, \mathcal{H}) = d < \infty$, entonces para todo n, $\pi_{(X,\mathcal{H})}(n) \leq \sum_{i=0}^{d} \binom{n}{i}$. En particular, si $n > d+1$ entonces $\pi_{(X,\mathcal{H})}(n) \leq (e \cdot n/d)^d$.*

El siguiente teorema muestra que si $\pi_{(X,\mathcal{H})}(n)$ crece *lentamente* entonces $\mathcal{H}$ tiene convergencia uniforme.

Teorema 4.4.3 *Sea $\mathcal{H}$ una clase de hipótesis con dominio X, y $\pi_{(X,\mathcal{H})}(n)$ su función de granularidad máxima. Entonces para toda distribución $\mathcal{D}$ y para todo $\delta \in (0,1)$ se tiene con probabilidad al menos $1-\delta$ sobre un conjunto de entrenamiento $S \sim \mathcal{D}^m$ que*

$$|R_{\mathcal{D}}(h) - \widehat{R}_S(h)| \leq \frac{4 + \sqrt{\log(\pi_{(X,\mathcal{H})}(2m))}}{\delta\sqrt{2m}}$$

Demostración: Sea

$$a(m) = \frac{4 + \sqrt{\log(\pi_{(X,\mathcal{H})}(2m))}}{\delta\sqrt{2m}}.$$

Demostraremos que

$$\mathop{\mathbb{E}}_{S\sim\mathcal{D}^m}\left[\sup_{h\in\mathcal{H}} |R_{\mathcal{D}}(h) - \widehat{R}_S(h)|\right] \leq a(m)\cdot\delta \tag{4.6}$$

La desigualdad de Markov establece que si X es una variable aleatoria y $a > 0$, entonces

$$\mathbb{P}[|X| \geq a] \leq \frac{\mathbb{E}[|X|]}{a}.$$

Como $sup_{h\in\mathcal{H}} |R_{\mathcal{D}}(h) - \widehat{R}_S(h)| \geq 0$ es una variable aleatoria sobre $\mathcal{D}^m$, utilizando la desigualdad de Markov y la inecuación (4.6), se llega a

$$\mathbb{P}\left[sup_{h\in\mathcal{H}} |R_{\mathcal{D}}(h) - \widehat{R}_S(h)| \geq a(m)\right] \leq \frac{\mathbb{E}_{S\sim\mathcal{D}^m}\left[\sup_{h\in\mathcal{H}} |R_{\mathcal{D}}(h) - \widehat{R}_S(h)|\right]}{a(m)} \leq \delta$$

y por lo tanto

$$\mathbb{P}\left[|R_{\mathcal{D}}(h) - \widehat{R}_S(h)| \leq a(m)\right] \leq 1 - \delta$$

lo que finaliza la demostración. Probemos entonces la ecuación (4.6).

$$\mathop{\mathbb{E}}_{S\sim\mathcal{D}^m}\left[\sup_{h\in\mathcal{H}} |R_{\mathcal{D}}(h) - \widehat{R}_S(h)|\right] = \mathop{\mathbb{E}}_{S\sim\mathcal{D}^m}\left[\sup_{h\in\mathcal{H}} |\mathop{\mathbb{E}}_{S'\sim\mathcal{D}^m}[\widehat{R}_{S'}(h)] - \widehat{R}_S(h)|\right]$$

Por la desigualdad $abs(\mathbb{E}[X]) \leq \mathbb{E}[abs(X)]$ se tiene

$$\left|\mathop{\mathbb{E}}_{S'\sim\mathcal{D}^m}[\widehat{R}_{S'}(h) - \widehat{R}_S(h)]\right| \leq \mathop{\mathbb{E}}_{S'\sim\mathcal{D}^m}\left[|\widehat{R}_{S'}(h) - \widehat{R}_S(h)|\right]$$

y como el supremo del valor esperado es menor que el valor esperado del supremo,

$$\sup_{h\in\mathcal{H}} \mathop{\mathbb{E}}_{S'\sim\mathcal{D}^m}\left[|\widehat{R}_{S'}(h) - \widehat{R}_S(h)|\right] \leq \mathop{\mathbb{E}}_{S'\sim\mathcal{D}^m}\left[\sup_{h\in\mathcal{H}} |\widehat{R}_{S'}(h) - \widehat{R}_S(h)|\right]$$

Por lo tanto

$$\begin{aligned}\mathop{\mathbb{E}}_{S\sim\mathcal{D}^m}\left[\sup_{h\in\mathcal{H}}|R_{\mathcal{D}}(h)-\widehat{R}_S(h)|\right] &\leq \mathop{\mathbb{E}}_{S',S\sim\mathcal{D}^m}\left[\sup_{h\in\mathcal{H}}|\widehat{R}_{S'}(h)-\widehat{R}_S(h)|\right] \\ &= \mathop{\mathbb{E}}_{S',S\sim\mathcal{D}^m}\left[\sup_{h\in\mathcal{H}}\frac{1}{m}|\sum_{i=1}^{m}(l(h,z_i')-l(h,z_i))|\right].\end{aligned} \tag{4.7}$$

En el último valor esperado los vectores aleatorios z' y z son los pertenecientes a los conjuntos S' y S, respectivamente, que se escogen i.i.d. según $\mathcal{D}$. Como se escogen i.i.d., si se sustituye z_i por z_i' no cambiaría nada, por lo que $(l(h,z_i')-l(h,z_i))$ y $-(l(h,z_i')-l(h,z_i))$ son intercambiables para cada i. Así que para cada $\sigma\in\{\pm1\}^m$ la ecuación (4.7) equivale a

$$\mathop{\mathbb{E}}_{S',S\sim\mathcal{D}^m}[\sup_{h\in\mathcal{H}}\frac{1}{m}|\sum_{i=1}^{m}\sigma_i(l(h,z_i')-l(h,z_i))|]$$

Esto también sucede si cada σ_i se escoge según una distribución $U_\pm$ uniforme sobre $\{\pm1\}$. Luego (4.7) es igual a

$$\mathop{\mathbb{E}}_{\sigma\sim U_\pm^m}\mathop{\mathbb{E}}_{S',S\sim\mathcal{D}^m}[\sup_{h\in\mathcal{H}}\frac{1}{m}|\sum_{i=1}^{m}\sigma_i(l(h,z_i')-l(h,z_i))|]$$

Fijados S y S' consideremos $C=S\cup S'$. En este caso podemos considerar el supremo sobre la restricción de $\mathcal{H}$ a C

$$\mathop{\mathbb{E}}_{\sigma\sim U_\pm^m}[\sup_{h\in\mathcal{H}}\frac{1}{m}|\sum_{i=1}^{m}\sigma_i(l(h,z_i')-l(h,z_i))|]=\mathop{\mathbb{E}}_{\sigma\sim U_\pm^m}[\sup_{h\in\mathcal{H}_C}\frac{1}{m}|\sum_{i=1}^{m}\sigma_i(l(h,z_i')-l(h,z_i))|]$$

Fijado $h\in\mathcal{H}_C$, denotemos $\theta_h=\frac{1}{m}\sum_{i=1}^{m}\sigma_i(l(h,z_i')-l(h,z_i))$. Como z y z' provienen i.i.d. de la misma distribución, sus respectivos valores esperados son iguales, por lo que se tiene que $\mathbb{E}[\theta_h]=0$. Como θ_h es la media de variables independientes que toman valores en $[-1,1]$, aplicando la desigualdad de Hoeffding del Lema 3.5.2 se tiene que para cada $\rho>0$,

$$\mathbb{P}[|\theta_h|>\rho]\leq 2\exp(-2m\rho^2).$$

y tomando el máximo en $\mathcal{H}_C$

$$\mathbb{P}[\max_{\mathcal{H}_C} |\theta_h| > \rho] \leq 2|\mathcal{H}_C| \exp(-2m\rho^2).$$

Aplicando el lema .0.6 del apéndice se obtiene

$$\mathbb{E}[\max_{\mathcal{H}_C} |\theta_h|] \leq \frac{4 + \sqrt{log(|\mathcal{H}_C|)}}{\sqrt{2m}}.$$

Estos resultados combinados con la definición de $\pi_{(X,\mathcal{H})}(n)$ muestran que (4.6) es cierto. □

Tras probar el teorema 4.4.3 ya estamos en condiciones de demostrar la última parte del Teorema Fundamental del Aprendizaje Estadístico. Para ello hay que probar que si la dimensión VC de un espacio de aprendizaje es finita, entonces la clase de hipótesis tiene la propiedad de convergencia uniforme.

Demostración: (4) ⇒ (1). Para probar la convergencia uniforme daremos una cota de la función $m_{\mathcal{H}}^{UC}(\epsilon, \delta)$ que depende de $d = dim_{VC}(X, \mathcal{H})$.

Por el lema de Sauer 4.4.2 se tiene que para $m > d$, $\pi_{(x,\mathcal{H})}(2m) \leq (2em/d)^d$. Esto, junto con el teorema 4.4.3 anterior muestra que, con probabilidad al menos $1 - \delta$, se tiene

$$|R_{\mathcal{D}}(h) - \widehat{R}_S(h)| \leq \frac{4 + \sqrt{d \log(2em/d)}}{\delta\sqrt{2m}}$$

Por simplicidad, podemos asumir que $\sqrt{d \log(2em/d)} \geq 4$, ya que si es menor tomando un tamaño $m \geq 32/(\delta\epsilon)^2$ tenemos $|R_{\mathcal{D}}(h) - \widehat{R}_S(h)|$ está acotado por ϵ. Por lo tanto,

$$|R_{\mathcal{D}}(h) - \widehat{R}_S(h)| \leq \frac{1}{\delta}\sqrt{\frac{2d \log(2em/d)}{m}}$$

y para asegurar que esto sea menor a ϵ tomamos

$$m \geq \frac{2d \log(m)}{(\delta\epsilon)^2} + \frac{2d \log(2e/d)}{(\delta\epsilon)^2}$$

Mediante operaciones algebraicas expuestas en el Apéndice .0.5 podemos demostrar que una condición suficiente para que se cumpla la anterior desigualdad es

$$m \geq 4\frac{2d}{(\epsilon\delta)^2}\log\left(\frac{2d}{(\epsilon\delta)^2}\right) + \frac{4d\log(2e/d)}{(\epsilon\delta)^2}$$

Luego para satisfacer la propiedad de convergencia uniforme se puede tomar

$$m_{\mathcal{H}}^{UC}(\epsilon,\delta) \leq 4\frac{16d}{(\epsilon\delta)^2}\log\left(\frac{16d}{(\epsilon\delta)^2}\right) + \frac{16d\log(2e/d)}{(\epsilon\delta)^2}$$

Y con esto finaliza la demostración del teorema fundamental del aprendizaje estadístico.

□

Apéndice

Lema .0.4 ***(Apéndice A.1)*** *Sea $a > 0$. Entonces $x \geq 2a\log(a) \Rightarrow x \geq a\log(x)$.*

Demostración:

Para $a \in (0, \sqrt{e}]$ la desigualdad siempre se cumple y la demostración es trivial. Supongamos que $a > \sqrt{e}$. Consideremos la función $f(x) = x - a\log(x)$, que tiene derivada $f'(x) = 1 - a/x$, que es positivo si $x > a$ y por lo tanto la función es creciente. Se tiene que $2a\log(a) > a$ si $a > \sqrt{e}$, y evaluando la función

$$\begin{aligned} f(2a\log(a)) &= 2a\log(a) - a\log(2a\log(a)) \\ &= 2a\log(a) - a\log(a) - a\log(2\log(a)) \\ &= a\log(a) - a\log(2\log(a)) \\ &= a\log\left(\frac{a}{2\log(a)}\right) \end{aligned} \tag{8}$$

Cómo $a > 2\log(a) \forall a > 0$ se tiene el resultado.

□

Lema .0.5 ***(Apéndice A.2)*** *Sea $a \geq 1$ y $b > 0$. Entonces $x \geq 4a\log(2a) + 2b \Rightarrow x \geq a\log(x) + b$*

Demostración: Como $a \geq 1$ entonces claramente $x \geq 2b$. Por otro lado, como $b > 0$, se tiene $x \geq 4a\log(2a)$, y aplicando el Lema .0.4 tenemos $x \geq 2a\log(x)$. Sumando las dos desigualdades $2x \geq 2a\log(x) + 2b$. □

Lema .0.6 ***(Apéndice A.3)*** *Sea X una variable aleatoria y $x' \in \mathbb{R}$ un escalar. Si existen $a > 0$ y $b \geq e$ tales que $\forall t \geq 0$ se cumple $\mathbb{P}[|X - x'| > t] \leq 2be^{-t^2/a^2}$, entonces $\mathbb{E}[|X - x'|] \leq a(2 + \sqrt{\log(b)})$.*

DEMOSTRACIÓN: Para $i = 1, 2, \ldots$ denotamos $t_i = a(i+\sqrt{log(b)})$. t_i es monótonamente creciente, por lo que

$$\mathbb{E}[|X - x'|] \leq a\sqrt{\log(b)} + \sum_{i=1}^{\infty} t_i \mathbb{P}[|X - x'| > t_{i-1}]$$

Desarrollando el sumatorio y usando las condiciones del lema se tiene

$$\begin{aligned}
\sum_{i=1}^{\infty} t_i \mathbb{P}[|X - x'| > t_{i-1}] &\leq 2ab \sum_{i=1}^{\infty} (i + \sqrt{\log(b)}) e^{-(i-1+\sqrt{\log(b)})^2} \\
&\leq 2ab \int_{1+\sqrt{\log(b)}}^{\infty} x e^{-(x-1)^2} dx \\
&= 2ab \int_{\sqrt{\log(b)}}^{\infty} (y+1) e^{-y^2} dy \\
&\leq 4ab \int_{\sqrt{\log(b)}}^{\infty} y e^{-y^2} dy \\
&= 2ab \left[-e^{-y^2} \right]_{\sqrt{\log(b)}}^{\infty} \\
&= 2ab/b = 2a
\end{aligned}$$

Combinando este resultado con la primera desigualdad completamos la prueba. □

Lema .0.7 ***(Apéndice A.4)*** *Sean $m, d \in \mathbb{N}$ tales que $d \leq m - 2$. Entonces.*

$$\sum_{k=0}^{d} \binom{m}{k} \leq \left(\frac{em}{d}\right)^d.$$

DEMOSTRACIÓN: Por inducción. Para $d = 1$ se tiene $1 + m \leq em$ que es cierto. Supongamos ahora que es cierto para d y probémoslo para $d + 1$. Por hipótesis de inducción se tiene

$$\begin{aligned}
\sum_{k=0}^{d+1} \binom{m}{k} &\leq \left(\frac{em}{d}\right)^d + \binom{m}{d+1} \\
&= \left(\frac{em}{d}\right)^d \cdot \left(1 + \left(\frac{d}{em}\right)^d \frac{m(m-1)(m-2)\cdots(m-d)}{(d+1)d!}\right) \\
&\leq \left(\frac{em}{d}\right)^d \cdot \left(1 + \left(\frac{d}{e}\right)^d \frac{m-d}{(d+1)d!}\right)
\end{aligned}$$

La fórmula de Stirling está dada por

$$d! = \sqrt{2\pi d}\left(\frac{d}{e}\right)^d e^{\frac{1}{12d}-\frac{1}{360d^3}+\frac{1}{1260d^5}-\cdots} \geq \sqrt{2\pi d}\left(\frac{d}{e}\right)^d$$

y aplicándola a la desigualdad anterior

$$\begin{aligned}
&\leq \left(\frac{em}{d}\right)^d \cdot \left(1 + \left(\frac{d}{e}\right)^d \frac{m-d}{(d+1)\sqrt{2\pi d}(d/e)^d}\right) \\
&= \left(\frac{em}{d}\right)^d \cdot \left(1 + \frac{m-d}{(d+1)\sqrt{2\pi d}}\right) \\
&= \left(\frac{em}{d}\right)^d \cdot \frac{d+1+(m-d)/\sqrt{2\pi d}}{d+1} \\
&\leq \left(\frac{em}{d}\right)^d \cdot \frac{d+1+(m-d)/2}{d+1} \\
&= \left(\frac{em}{d}\right)^d \cdot \frac{d/2+1+m/2}{d+1} \\
&\leq \left(\frac{em}{d}\right)^d \cdot \frac{m}{d+1}
\end{aligned}$$

Se usado que $d \leq m-2$ en la última desigualdad. Continuamos por el otro lado

$$\begin{aligned}
\left(\frac{em}{d+1}\right)^{d+1} &= \left(\frac{em}{d}\right)^d \cdot \frac{em}{d+1} \cdot \left(\frac{d}{d+1}\right)^d \\
&= \left(\frac{em}{d}\right)^d \cdot \frac{em}{d+1} \cdot \frac{1}{(1+1/d)^d} \\
&\geq \left(\frac{em}{d}\right)^d \cdot \frac{em}{d+1} \cdot \frac{1}{e} \\
&= \left(\frac{em}{d}\right)^d \cdot \frac{m}{d+1}.
\end{aligned}$$

Lo que finaliza la prueba □

Conclusiones

Para cumplir el objetivo de presentar y comprender el concepto de la dimensión de Vapnik-Chervonenkis, a lo largo de esta memoria se han desarrollado las bases de la teoría del aprendizaje estadístico, fundamentado en el principio de minimización del riesgo empírico.

Cabe destacar la importancia de esto en la ciencia de los datos, ya que trata el problema de inferencia estadística de encontrar una función predictiva basada en un conjunto de datos, y se ha aplicado con éxito en problemas tecnológicos actuales como la visión artificial, el reconocimiento del habla o el diagnóstico de enfermedades.

El resultado más relevante expuesto en este trabajo es el del Teorema Fundamental del Aprendizaje Estadístico, que muestra claramente la importancia de la dimensión VC en los modelos de clasificación binaria. La colección de ejemplos de su cálculo aporta una mayor comprensión sobre como actúa la dimensión VC sobre distintas familias de funciones.

En conclusión, considero que este trabajo aporta una base de conocimiento fundamental de la teoría de aprendizaje estadístico y explica el papel que juega la dimensión VC en la ciencia de datos.

Trabajo Futuro

Existen numerosos aspectos de la dimensión VC que se podrían desarrollar más allá de lo expuesto en esta memoria:

- Cálculo de la dimensión VC de las clases de funciones devueltas por algoritmos comúnmente utilizados la ciencia de los datos como redes neuronales y árboles de decisión.

- Desarrollo desde el punto de vista geométrico, y concretamente en el campo de geometría computacional.
- Aportar cotas numéricas al Teorema Fundamental del Aprendizaje Estadístico.
- Desarrollo de distintas generalizaciones propuestas para la dimensión VC como la dimensión de Natarajan.
- La complejidad de Redemacher es un concepto similar que también aporta cotas al error empírico.

Bibliografía

[1] Blum, A., Hopcroft, J., and Kannan, R. *Foundations of Data Science*. Cambridge University Press, 2020.

[2] Blumer, A., Ehrenfeucht, A., Haussler, D., and Warmuth, M. K. Learnability and the vapnik-chervonenkis dimension. *J. ACM 36*, 4 (oct 1989), 929–965.

[3] Díaz, J. B. Las matemáticas en el país de los datos (i): De puntos a mónadas. *La Gaceta de la RSME 20* (2017), 113–142.

[4] H. Wolpert, D. The supervised learning no-free-lunch theorems. *Roy R., Köppen M., Ovaska S., Furuhashi T., Hoffmann F. (eds) Soft Computing and Industry* (2002), 25–42.

[5] Haussler, D. Probably approximately correct learning. In *AAAI-90 Proceedings* (1990), AAAI'90, AAAI Press, p. 1101–1108.

[6] Kearns, M. J., and Vazirani, U. V. *An Introduction to Computational Learning Theory*. MIT Press, 1994.

[7] Mitchel, T. *Machine Learning*. McGraw Hill, 1997.

[8] Rostamizadeh, A., Talwalkar, A., and Mohri, M. *Foundations of Machine Learning*. The MIT Press, 2018.

[9] Shalev-Shwartz, S., and Ben-David, S. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, 2014.

[10] Valiant, L. G. A theory of the learnable. *Commun. ACM 27*, 11 (nov 1984), 1134–1142.

[11] Vapnik V.N., C. A. *On the Uniform Convergence of Relative Frequencies of Events to Their Probabilities*. Springer, 2015.