UNIVERSITAT DE BARCELONA
Facultat de Matemàtiques i Informàtica

GRAU DE MATEMÀTIQUES

Treball final de grau

APRENSIBILITAT I COTES DE GENERALITZACIÓ EN APRENENTATGE AUTOMÀTIC

Autor: Joaquim Guiot Cusidó

Director: Dr. Oriol Pujol Vila i Dr. Carles Rovira Escofet
Realitzat a: Departament de matemàtiques i informàtica

Barcelona, 15 de gener de 2025

Abstract

The aim of this project is to study the mathematical foundations of machine learning, focusing on the domain of learnability and generalization bounds, leading to its most important results. Additionally, a recent research finding is analyzed from the perspective of the results introduced throughout the work. First, the desired and required theoretical concepts and results are presented. Finally, the mentioned analysis is carried out.

Resum

L'objectiu d'aquesta memòria és estudiar els fonaments matemàtics de l'aprenentatge automàtic, restringint-se a l'àmbit de l'aprensibilitat i les cotes de generalització, fins arribar als resultats més importants del mateix, així com analitzar un resultat recent de recerca des de la perspectiva dels resultats introduits al llarg del treball. En primer lloc es presenten els conceptes i resultats teòrics desitjats i requerits. Finalment es realitza l'anàlisi mencionat.

2020 Mathematics Subject Classification. 68, 60, 62, 28

Agraïments

En primer lloc, vull mostrar el meu agraïment als dos tutors que han dirigit aquest treball. Vull agrair al Dr. Oriol Pujol i Vila no només el seu acompanyament i guiatge al llarg de tota la realització del treball, des de la proposta del tema fins a les últimes revisions i correccions de la memòria, sinó també haver-me introduit al fascinant món de l'aprenentatge automàtic, del qual no tenia cap coneixement previ a emprendre l'aventura d'aquest treball final de grau. Vull agrair al Dr. Carles Rovira i Escofet no només el seu acompanyament i guiatge al llarg del treball, sinó també el seu esforç al dirigir aquest treball malgrat que el tema no fos del seu àmbit habitual d'estudi. En segon lloc, vull agraïr a tots aquells agents que, d'alguna manera o altra, m'han ajudat al llarg d'aquests 5 anys i mig d'experiència universitària; companys, familiars i professors, però especialment als meus amics dins del grau qui, dia rere dia, han fet el procés més fàcil i agradable. Finalment, vull dedicar un especial agraïment als meus familiars més propers. Als meus pares, Miquel i Teresa, per sempre confiar en mi i donar-me suport, així com per brindar-me l'oportunitat de poder cursar un grau universitari i d'assegurar-se que la meva vida fos el més fàcil i còmode possible per tal de poder dedicar-m'hi a temps complert. També al meu germà gran Miquel, qui no només sempre s'ha ofert per ajudar-me en tot allò que pogués necessitar sinó també contribuir a fer de mi la persona que soc avui, gràcies de tot cor.

Índex

Capítol 1

Introducció

Des de l'origen de la humanitat fins a l'actualitat, allò que ha caracteritzat i distingit als éssers humans de la resta d'espècies animals que han habitat la Terra és la seva capacitat per aprendre i comunicar-se; és a dir, la seva habilitat i facilitat per aprendre i transmetre a altres éssers humans allò que han aprés. Si bé la gran majoria d'espècies animals també són capaces d'aprendre i comunicar-se, podria argumentar-se que la magnitut de l'habilitat i la facilitat per fer-ho és el que ha marcat la diferència entre elles i l'ésser humà. Així doncs, resulta evident que la noció d'aprenentatge és i ha sigut de vital importància al llarg de la història de la humanitat. Fent una breu aturada per descriure què significa aquest concepte, l'aprenentatge es pot definir com el procés mitjançant el qual un individu, a través de l'experiència, l'observació i/o l'estudi, és capaç d'adquirir coneixements per després poder aplicar-los en futures situacions idèntiques, similars o completament diferents. D'ençà que l'ésser humà és metacognitiu, és a dir té la capacitat de pensar i avaluar els seus propis pensaments i habilitats, molts són els individus i moltes han sigut les societats que han buscat desxifrar quina és la clau que permet que aquest procés tingui lloc. No obstant, com que la comprensió d'aquesta causa no ha sigut mai necessària en el procés d'aprenetatge, mai ha esdevingut una necessitat determinar i quantificar aquells paràmetres (si existeixen) que determinen la capcitat de l'ésser humà per aprendre o la pròpia aprensibilitat d'una tasca. Aquesta situació ha canviat radicalment amb l'arribada dels ordinadors, ja que aquest procés que pels éssers humans resulta tan natural i està tan present en la vida quotidiana ha intentat ser extrapolat a les màquines; és a dir, s'ha plantejat la següent pregunta: poden aprendre les màquines?

Aquest concepte s'ha formalitzat en el que avui es coneix com *machine learning*, és a dir aprenentatge automàtic. L'aprenentatge automàtic es defineix com l'ús i desenvolupament de sistemes de computació capaços d'aprendre i adaptar-se sense instruccions específiques, fent servir algoritmes i models estadístics per analitzar i extreure conclusions de patrons significatius en les dades.

Així doncs, no n'hi ha prou que el programa emprat sigui capaç de memoritzar informació, sinò que ha de ser capaç de generalitzar i realitzar un raonament inductiu a partir d'aquesta. Ser capaç de predir, en base a dades vistes prèviament, el comportament que s'està estudiant per a dades noves. Aquesta és la gran diferència entre la memorització i l'aprenentatge.

Un cop plantejada la qüestió sobre si els ordinadors són capaços d'aprendre, sorgeixen noves incògnites com a conseqüències lògiques d'aquesta. En primer lloc, què es pot aprendre? Existeixen situacions impossibles d'aprendre i, per tant, de predir? En segon lloc, com es diferencia l'aprenentatge útil i/o correcte de l'aprenentatge erroni?

La primera pregunta es respon més endavant en aquesta memòria, veient per a diferents casos de problema i de noció d'aprenentatge condicions que garanteixen la capacitat d'aprenentatge o la impossibilitat de la mateixa. Pel que fa a la segona qüestió, una primera resposta a aquesta pot trobar-se en l'existència d'un coneixement previ que permet a l'aprenent intuir quina llei segueix el patró que està buscant. Aquest concepte pròximament queda formalitzat pel cas que s'aborda en aquest treball.
Cal aclarir abans de començar que, malgrat que hi ha diversos tipus d'aprenentatge, en aquest treball tan sols s'estudia el cas d'aprenentatge estadístic, supervisat per lots amb un aprenent passiu i un professor neutral; és a dir, l'aprenent tindrà a la seva disposició una mostra finita de dades prèviament classificades, sense cap biaix en l'elecció/generació d'aquestes, abans d'haver de començar a realitzar prediccions o proposar un model que resolgui la problemàtica. A més a més, l'aprenent no podrà interactuar ni amb les dades ni amb la realitat de cap manera més enllà d'observar-les; és a dir, no podrà realitzar experiments pel seu compte per tal de comprovar les seves hipòtesis o extreure nova informació.
Posant per un moment l'atenció en la pròpia raó de ser darrere de l'aprenetatge automàtic, hom podria qüestionar la utilitat del mateix; és a dir, és realment necessari l'aprenentatge automàtic? Existeixen tasques que requereixin estrictament de l'aplicació d'alguna mena d'aprenentatge automàtic? No n'hi ha prou que els éssers humans ens encarreguem de les tasques creatives i que involucrin aprenenetatge i pensament i deixem a les màquines tan sols aquelles tasques farragoses i consistents en seguir instruccions? Per sort o per desgràcia, la realitat és que, efectivament, existeixen tasques que son inconcebibles d'abordar sense la implementació de l'aprenentatge automàtic. Principalment, destaquen dues característiques que determinen la necessitat d'implementació d'aprenentatge autòmatic en la resolució d'un problema; la complexitat del mateix i la necessitat d'adaptabilitat. En primer lloc, posant l'èmfasi en el cas de problemes amb un grau massa elevat de complexitat, s'hi troben totes aquelles tasques que són massa complicades de programar, generalment degut que no es té un coneixment suficientment detallat de quines són les cadenes de raonament i passos lògics que se segueixen al realitzar-les. El principal conjunt de problemes d'aquesta índole del qual s'acostuma a parlar són totes aquelles accions que els éssers humans (i/o fins i tot altres espècies animals) realitzem però no en tenim un coneixement introspectiu suficientment detallat com per traduir-les en un programa informàtic; alguns exemples serien cuinar, conduir, reconeixer imatges, etc. També dins del conjunt de problemes amb un grau massa elevat de complexitat s'hi troben aquelles tasques que sobrepassen les capacitats humanes, com podria ser el tractament de conjunts de dades massa grans i complexos. En totes aquestes qüestions, la implementació de programes informàtics que "aprenen"de la pròpia experiència ha demostrat ser una eina útil oferint uns resultats satisfactoris. Per altra banda, pel que fa als problemes que requereixen d'adaptabilitat, un cop redactat el codi d'un programa informàtic, aquest romandrà immutable fins que algú el modifiqui manualment. Així doncs, si el problema que es preté resoldre és susceptible de mutar, el programa és susceptible de quedar obsolet i, per tant, requerir ser modificat. Estar revisant que el prgrama funcioni correctament per cada cas concret del problema seria extremadament ineficient, per tant la implementació de l'aprenentatge automàtic resulta imprescindible per tal de poder abordar la tasca de forma viable. Alguns exemples de problemes d'aquesta índole resolts mitjançant aprenentatge automàtic són els programes de reconeixement de veu o els filtres de detecció de correus brossa.

Pel que fa al tipus de problemes que s'estudien en aquesta memòria, si bé l'aprenentatge automàtic serveix i es fa servir per multitut de problemes diferents (tal i com s'acaba d'exposar), en aquest treball tan sols s'estudien problemes de classificació, restringint

l'estudi d'aquests al cas particular de classificació binària, amb l'excepció puntual del resultat que es presenta a la secció 5.2. No obstant, aquest resultat també és cert per al cas de classificació binària, que és l'òptica des de la qual es discuteix a la secció 5.3, en consonància amb la resta de la memòria.

Finalment, convé destacar que la majoria dels continguts i demostracions d'aquest treball, així com l'estructura i l'ordre que s'ha seguit al llarg del mateix, s'ha extret del llibre *Understanding Machine Learning: From Theory to Algorithms* dels professors universitaris Shai Shalev-Shwartz de la Universitat Hebrea de Jerusalem i Shai Ben-David de la Universitat de Waterloo [1]. Addicionalment, l'article al qual es fa esment a l'últim capítol i del qual s'ha extret el resultat discutit s'estitula *On Generalitzation Bounds for Neural Networks with Low Rank Layers* dels doctors Andrea Pinto, Akshay Rangamani i Tomaso Poggio; publicat a la revista *Proceedings of Machine Learning Research, 30* [2].

Objectius del treball

- Entendre el problema de generalització de l'aprenentatge automàtic, els elements que el conformen i les tècniques que s'empren per abordar-lo.
- Introduir i desenvolupar els fonaments matemàtics de l'aprenentatge automàtic; més concretament, de l'aprenentatge automàtic probablement aproximadament correcte, l'aprensibilitat d'espais d'hipòtesis i les cotes de generalització.
- Discutir i comparar un resultat recent de recerca en cotes de generalització en aprenentatge profund des de la perspectiva del teorema fonamental de l'aprnentatge estadístic.

Estructura de la memòria

En primer lloc, convé destacar que l'objectiu d'aquesta memòria és introduir i desenvolupar els fonaments matemàtics de l'aprenentatge automàtic; més concretament, de l'aprenentatge automàtic probablement aproximadament correcte, l'aprensibilitat i les cotes de generalització. Al llarg de l'escrit, s'aniran presentant i desenvolupant/construint des de la perspectiva matemàtica, fent servir fonamentalment eines de teoria de la probabilitat, teoria de la mesura i desigualtats, els conceptes i resultats més rellevants dins d'aquest àmbit. No obstant, aquesta memòria no es limitarà tan sols a tractar el tema des d'una vesant merament teòrica, sinò que conclourà presentant un tema punter de recerca sobre un àmbit lleugerament diferent com és l'aprenentatge profund (*deep learning*) en xarxes neuronals i comparant els resultats obtinguts a l'article amb els resultats teòrics que s'hauran vist prèviament sobre l'aprenentatge automàtic.

Normalment els documents sobre aquesta temàtica tenen tendència a enfocar-ho des d'una perspectiva purament informàtica, centrant-se en exposar les aplicacions pràctiques i limitant-se merament a comentar una idea intuitiva d'allò que està passant sense entrar gaire en detall a les matemàtiques que hi ha al darrere. Per altra banda, tampoc convé perdre la perspectiva sobre quina és la problemàtica que es preté resoldre i per tant no tenir una intuició sobre quin és el significat dels conceptes que van apareixent com a conseqüència d'estudiar el tema des d'una perspectiva merament matemàtica. Així doncs, aquest treball combina un estil divulgatiu on, a mesura que es van introduint nous conceptes, es proporciona una idea intuitiva sobre el seu significat, amb el formalisme i rigor matemàtic pertinent en les definicions d'aquests, així com a l'hora d'enunciar i

demostrar tan els resultats que se'n deriven com aquells resultats auxiliars que són requerits per tal d'obtenir-los.

Tot seguit, a mode de síntesi, es llisten els capítols dels quals consta aquesta estudi, així com els temes que s'hi tracten.

En primer lloc, el segon capítol està integrament dedicat a presentar el problema que s'aborda al treball, així com els elements que el conformen.

En segon lloc, al capítol 3 s'introdueix el primer model formal d'aprenentatge automàtic, l'aprenentatge probablement aproximadament correcte (PAC). A més a més, també es presenten altres nocions d'aprenentatge que se'n deriven com l'aprenentatge PAC agnòstic o la propietat de convergència uniforme. El capítol conclou amb l'enunciat i demostració del primer gran resultat del treball, el teorema NFL.

En tercer lloc, al capítol 4 es defineix la que serà la noció que caracteritza l'aprensibilitat o no d'una classe d'hipòtesis, la dimensió VC. A més a més, en aquest capítol és on es presenta i demostra el resultat teòric més important de tota la memòria, el teorema fonamental de l'aprenentatge estadístic.

Tot seguit, al penúltim capítol es presenta un resultat recent d'un àmbit de recerca punter i, a mode de contribució pròpia de l'autor, s'efectua una comparativa respecte la versió quantitativa del teorema fonamental de l'aprenentatge estadístic i una discussió dels resultats obtinguts. Prèviament al mateix capítol s'introdueixen tots els conceptes necessaris per tal de poder comprendre aquest resultat.

Finalment, el darrer capítol són les conclusions generals extretes en aquesta memòria.

Guia de lectura

A fi de proveir al lector una guia de lectura dels capitols dos, tres i quatre, s'inclou a continuació un breu resum que emfatitza la línia argumental i deductiva dels conceptes i resultats introduits.

En primer lloc es presenta el problema de classificació binària que consta d'un conjunt de domini $\mathcal{X}$, un conjunt de classificacions $\mathcal{Y} = \{0, 1\}$ i una mostra d'entrenament S. L'objectiu és que l'aprenent (algoritme d'aprenentatge) sigui capaç de retornar una funció de predicció $h : \mathcal{X} \to \{0, 1\}$ amb el mínim error possible. D'entrada, s'assumeix que existeix una distribució de probabilitat $\mathcal{D}$ sobre $\mathcal{X}$ i una funció $f : \mathcal{X} \to \{0, 1\}$ de classificació que és la correcta, les quals són desconegudes per l'aprenent. Tot seguit es defineixen l'error empíric i l'error real, així com el paradigma d'aprenentatge que es considera al llarg de la memòra; l'ERM. Per prevenir el sobreajustament, es delimita l'espai de les hipòtesis h a una classe $\mathcal{H}$ i s'imposa la restricció que tingui un nombre finit d'elements; d'ara en endavant, es preté determinar què caracteritza l'aprensibilitat de $\mathcal{H}$. Fins indicació contrària, s'assumeixen les assumpcions de realitzabilitat i l'assumpció i.i.d. Vist que no es pot garantir ni un error nul ni una probabilitat 1 d'assolir-lo, s'introdueixen els paràmetres de confiança i precisió (δ i ε) per tal de paliar-ho. Així doncs, el nou objectiu és garantir amb probabilitat major o igual a $1 - \delta$ que l'error real sigui menor que ε.

Al següent capítol, es formalitza la noció anterior en el primer model d'aprenetatge d'aquesta memòria; l'aprenentatge PAC. A partir d'ara, es relaxa l'assumpció de realitzabilitat i $\mathcal{D}$ deixa de ser una distribució sobre $\mathcal{X}$ per esdevenir una distribució sobre $\mathcal{X} \times \{0, 1\}$. Ara ja no es pot garantir un error arbitràriament petit, sinò tan sols un error arbitràriament pròxim al mínim possible (al de l'element $h \in \mathcal{H}$ amb l'error mínim). Aquesta noció es formalitza en el segon model d'aprenetatge, l'aprenentatge PAC agnòstic.

Per tal de generalitzar aquesta noció d'aprenentatge a altres problemes de classificació, s'introdueix la noció de funció de pèrdua i es defineixen la funció de risc i el risc empíric com a alternatives a l'error real i a l'error empíric. A continuació, es detecta un problema intrínsec al propi funcionament de l'ERM, aquest selecciona el predictor que minimitza l'error empíric, esperant que aquest predictor tingui un error real reduit; malgrat no hi hagi cap garantia al respecte. Per tal d'establir aquesta garantia, s'introdueix la noció de mostra ε-representativa; la qual, si es dona amb probabilitat major o igual a $1 - \delta$ per una mostra S, garanteix que l'ERM sigui un aprenent PAC agnòstic. Aquesta noció es generalitza en la condició de convergència uniforme. Tot seguit, s'introduexi el teorema "No-Free-Lunch"(NFL), el qual nega l'existència d'un aprenenent universal i, per tant, posa de manifest la necessitat de restringir $\mathcal{H}$.

A fi de caracteritzar l'aprensibilitat d'una classe d'hipòtesis $\mathcal{H}$ infinita, es defineix el concepte de dimensió VC; posteriorment, es presenta el teorema fonamental de l'aprenentatge estadístic, la versió qualitativa del qual finalment caracteritza l'aprensibilitat d'una classe d'hipòtesis per a problemes de classificació binària, recoltzant-se en la noció de dimensió VC. També s'inclou la versió quantitativa, la qual proporciona cotes a la complexitat de la mostra i serà emprada a l'últim capítol per tal d'analitzar un resultat recent de recerca en l'àmbit de les cotes de generalització en aprenentatge profund.

Capítol 2

Primer model formal del problema

Es considera el següent problema: Un aprenent A vol trobar una norma general per resoldre un problema de classificació binària en base a uns paràmetres que coneix (Ex: Un zoòleg vol trobar un sistema per classificar una nova espècie de cavall, descoberta recentment, segons si els exemplars tenen o no sobrepés en base a la longitud de la columna vertebral i el pes de l'animal).
En primer lloc, l'aprenent té accés a la següent informació:

- **Conjunt de domini:** Un conjunt arbitrari $\mathcal{X}$, el qual és el conjunt d'objectes que l'aprenent vol classificar. Cada punt d'aquest conjunt és una instància i està representat per un vector de característiques. A l'exemple dels cavalls, el domini seria el conjunt de tots els cavalls d'aquesta espècie; cada punt estaria representat per un vector de dues component, la llargada de la columna vertebral i el pes de l'animal.
- **Conjunt de classificacions/etiquetes:** Un conjunt $\mathcal{Y}$ d'etiquetes que poden prendre els diferents elements de $\mathcal{X}$. Al llarg d'aquest treball només considerarem problemes de classificació binària; és a dir $\mathcal{Y} = \{0, 1\}$ o $\mathcal{Y} = \{-1, 1\}$. A l'exemple anterior, 0 representaria la classificació "no té sobrepés" i 1 la classificació "té sobrepés".
- **Dades d'entrenament:** Una seqüencia $S = ((x_1,y_1), ..., (x_m,y_m))$ finita de parelles de $\mathcal{X} \times \mathcal{Y}$; és a dir, una seqüencia de punts del domini etiquetats. Aquesta és l'entrada ("input") que rep l'aprenent. Sovint s'anomena a S conjunt d'entrenament i als seus elements exemples d'entrenament. A l'exemple, seria una colecció d'exemplars d'aquesta espècie (les seves longituts de columna i pesos) marcats prèviament com a exemplars amb o sense sobrepés.

En segon lloc, a l'aprenent se li requereix que retorni una funció h: $\mathcal{X} \to \mathcal{Y}$ que predigui la classificació d'un punt x qualsevol del domini. A aquesta funció se l'anomena regla de predicció, predictor, hipòtesi o classificador. Així doncs, la informació de sortida del programa que actuï com a aprenent serà una funció d'aquesta forma.
Dit això, per tal d'explicar com es generen els elements x de la mostra d'entrenament S, en tot moment es farà l'assumpció (que en pròximes seccions serà relaxada) que existeix una distribució de probabilitat $\mathcal{D}$ sobre $\mathcal{X}$, d'acord a la qual es generen els elements de les dades d'entrenament. És important recalcar que en cap moment s'assumeix que l'aprenent

sàpiga res d'aquesta distribució de probabilitat. A més a més, pel que fa a les etiquetes dels elements del domini, s'assumeix que existeix una funció de classificació correcta f: $\mathcal{X} \to \mathcal{Y}$, la qual és desconeguda per l'aprenent, ja que és justament allò que està intentant determinar.
Així doncs, a mode de resum i per il·lustrar millor el problema, hom pot pensar en el següent exemple: Un zoòleg descobreix una nova espècie de mamífer similar al cavall i vol trobar una funció que el permeti classificar els membres d'aquesta espècie segons si tenen sobrepés o no basant-se en el pes i la longitud de la columna vertebral. Com a dades de mostra en les que basar-se, disposa d'un conjunt d'exemplars els quals ha pogut classificar prèviament i de manera correcta segons si tenen o no sobrepés.
Òbviament, l'objectiu de l'aprenent no és tan sols retornar una funció qualsevol, sinó que vol trobar el millor predictor possible (a poder ser la f correcta). No obstant, per ser capaç de diferenciar la qualitat de les diferents hipòtesis candidates, és imprescindible disposar d'un mecanisme per mesurar la precisió (encert), o equivalentment la mancança d'aquesta (error), per a una hipòtesi qualsevol. No obstant, primer cal definir què és l'error d'un classificador.
Es defineix l'error d'un classificador h com la probabilitat que h predigui malament (és a dir, de manera diferent a la funció verdadera f) la classificació d'un punt x qualsevol de $\mathcal{X}$ generat d'acord a $\mathcal{D}$. S'introdueix la següent notació que es farà servir d'ara en endavant: $L_{\mathcal{D},f}$; on L representa la pèrdua[1], $\mathcal{D}$ la distribució de probabilitat sobre $\mathcal{X}$ i f la funció que prediu correctament en base als paràmetres de $\mathcal{X}$ la classificació corresponent de $\mathcal{Y}$.

Definició 1. *Sigui $\mathcal{X}$ un conjunt de domini, sigui $\mathcal{Y} = \{0, 1\}$ un conjunt de classificacions per als punts de $\mathcal{X}$, sigui $\mathcal{D}$ una distribució de probabilitat segons la qual es generen aleatòriament els punts x de $\mathcal{X}$, sigui f: $\mathcal{X} \to \mathcal{Y}$ la funció correcta de classificació (per a tot $x \in \mathcal{X}$, $f(x) = y$), sigui h: $\mathcal{X} \to \mathcal{Y}$ un classificador qualsevol. L'error real ("true error") de h és:*

$$L_{\mathcal{D},f}(h) := \mathop{\mathbb{P}}_{x \sim \mathcal{D}}[h(x) \neq f(x)].$$

Sovint s'emprarà la notació $\mathcal{D}(B)$ per indicar la mesura del conjunt B segons la distribució $\mathcal{D}$.

Hom pot observar que aquesta noció d'error depén de la distribució de probabilitat $\mathcal{D}$ sobre $\mathcal{X}$, així com de la funció f; no obstant, prèviament s'ha posat de manifest que l'aprenent desconeix completament aquests dos elements, aleshores no és una noció que l'aprenent pugui emprar per determinar la qualitat de les diferents hipòtesis. Per altra banda, la informació que sí que està disponible i és coneguda per l'aprenent és la mostra S que rep, i en base a la qual ha de retornar una hipòtesi h_S: $\mathcal{X} \to \mathcal{Y}$ amb l'objectiu que minimitzi l'error respecte $\mathcal{D}$ i f. Així doncs, sembla lògic buscar una hipòtesi que funcioni bé a la mostra S amb l'esperança de minimitzar també $L_{\mathcal{D},f}(h)$. Per tal de mesurar la qualitat d'un predictor respecte una mostra S es defineix la següent noció d'error:

Definició 2. *Sigui $\mathcal{X}$ un domini, sigui $\mathcal{Y} = \{0, 1\}$, sigui S una mostra de $m \in \mathbb{N}$ elements de $\mathcal{X} \times \mathcal{Y}$ generada d'acord a una distribució $\mathcal{D}$ sobre $\mathcal{X}$ desconeguda i classificada d'acord a una funció correcta f: $\mathcal{X} \to \mathcal{Y}$. Sigui $h : \mathcal{X} \to \mathcal{Y}$ un predictor. Es defineix l'error d'entrenament ("training error") de h com:*

[1] Generalització del concepte d'error per a altres problemes de classificació, no necessariament binaris. S'introdueix a la secció 3.2.

$$L_S(h) = \frac{|\{i \in [m] : h(x_i) \neq y_i\}|}{m} \quad , \quad \text{on } [m] = \{1, ..., m\} \text{ i } y_i = f(x_i).$$

I així s'arriba a la definició del primer paradigma d'aprenentatge que es proposarà:

Definició 3. *Sigui $\mathcal{X}$ un domini, sigui $\mathcal{Y} = \{0, 1\}$, sigui S una mostra de $m \in \mathbb{N}$ elements de $\mathcal{X} \times \mathcal{Y}$. La minimització de l'error empíric*[2] *("Empirical Risk Minimitzation") és un paradigma d'aprenentatge en el qual l'aprenent busca seleccionar un predictor h_S: $\mathcal{X} \to \mathcal{Y}$ tal que minimitzi $L_S(h_S)$.*

El motiu de la notació h_S és emfatitzar que el predictor seleccionat per l'aprenent depén de la mostra S.

L'ERM, en base al plantejament realitzat sobre la manca d'informació sobre $\mathcal{D}$ i f, d'entrada pot semblar una regla general per escollir predictors lògica i efectiva; no obstant, si no es fa servir amb prudència pot abocar a l'aprenent al fracàs. El següent exemple ho posa de manifest:

Exemple: Siguin $\mathcal{X}$ un domini, $\mathcal{Y} = \{0, 1\}$, $S = ((x_1, y_1), ..., (x_m, y_m))$ una col·lecció d'elements de $\mathcal{X} \times \mathcal{Y}$ generada aleatòriament segons una distribució $\mathcal{D}$ sobre $\mathcal{X}$ i sigui f: $\mathcal{X} \to \mathcal{Y}$ la funció de classificació correcta. Consideri's el següent predictor h_S: $\mathcal{X} \to \mathcal{Y}$:

$$h_{\mathcal{S}}(x) = \begin{cases} y_i & \text{si } \exists i \in [m] \text{ tal que } x_i = x, \\ 0 & \text{altrament.} \end{cases}$$

Resulta evident que aquest predictor no tindrà un bon error real, ja que fallarà en tots els punts $x \in \mathcal{X} \setminus S$ tals que $f(x) = 1$.

Aquest fenomen pel qual una hipòtesi fruit d'aplicar l'ERM fracassa respecte $\mathcal{D}$ i f malgrat resultar exitosa a la mostra S rep el nom de sobreajustament ("overfitting").

Definició 4. *Sigui $\mathcal{X}$ un domini, $\mathcal{Y} = \{0, 1\}$, $S = ((x_1, y_1), ..., (x_m, y_m))$ una col·lecció d'elements de $\mathcal{X} \times \mathcal{Y}$ generada aleatòriament segons una distribució $\mathcal{D}$ sobre $\mathcal{X}$, sigui f: $\mathcal{X} \to \mathcal{Y}$ la funció de classificació correcta. S'anomena sobreajustament al fenomen en el qual un model d'aprenentatge retorna una funció de predicció h: $\mathcal{X} \to \mathcal{Y}$ la qual no s'ajusta bé a la distribució real per ajustar-la massa a les dades de la mostra S.*

Un cop vista aquesta problemàtica, resulta imperativa la necessitat de solventar-la. Fonamentalment, aquesta resolució pot arribar per dues vies diferents; o bé modificar el paradigma d'aprenentatge ja que l'ERM és inconsistent, o bé realitzar-hi els ajustos pertinents per garantir-ne la fiabilitat. En tant que el raonament pel qual s'ha obtat per implementar l'ERM és correcte, s'optarà per reparar-lo enlloc de reemplaçar-lo per un paradigma d'aprenentatge completament nou. El mecanisme que s'emprarà consisteix en cercar condicions sota les quals està garantit que l'ERM no caurà mai en sobreajustament; qualitativament, significa imposar condicions sota les quals un baix error empíric impliqui amb alta probabilitat un baix error real.
La solució més freqüent a aquest problema consisteix en restringir l'espai de búsqueda sobre el qual s'aplica l'ERM, és a dir, fixar un conjunt $\mathcal{H}$ de funcions predictores, anomenat classe d'hipòtesis, sobre el qual s'aplicarà l'ERM. Formalment:

[2] El qual es denotarà com ERM, d'acord a la nomenclatura que reb a la literatura anglesa.

Definició 5. *Sigui $\mathcal{X}$ un domini, $\mathcal{Y} = \{0, 1\}$. Una classe d'hipòtesis és un conjunt $\mathcal{H}$ de predictors h: $\mathcal{X} \to \mathcal{Y}$.*

És interessant recalcar que la definició de classe d'hipòtesi no depén ni de la mostra S, ni de la distribució $\mathcal{D}$ ni tampoc de la funció d'etiquetatge f. Això es deu que l'aprenent ha d'escollir la classe d'hipòtesis abans de veure les dades; és a dir, al moment de restringir el conjunt de funcions sobre el qual s'aplicarà l'ERM, l'aprenent no coneix la mostra S. Un cop seleccionada la classe d'hipòtesis $\mathcal{H}$, l'aprenent té accés a la mostra S i aplica l'ERM sobre $\mathcal{H}$ per tal de trobar el predictor $h \in \mathcal{H}$ amb el menor error possible sobre S. A aquest procediment se l'anomena aprenent $\text{ERM}_{\mathcal{H}}$. Formalment:

Definició 6. *$ERM_{\mathcal{H}}(S) = h' \in \mathcal{H}$ tal que $h' \in \underset{h \in \mathcal{H}}{argmin}\ L_S(h)$.*
On $\underset{h \in \mathcal{H}}{argmin}\ L_S(h) := \{\ h' \in \mathcal{H} : L_S(h') \leq L_S(h)\ \forall\ h \in \mathcal{H}\ \}$; és a dir, el conjunt d'hipòtesis h de $\mathcal{H}$ tals que $L_S(h)$ assoleix el seu valor mínim sobre $\mathcal{H}$.

Un cop vista la imposició de restriccions sobre el conjunt de funcions que es consideren candidates per tal de garantir que l'ERM no caigui en sobreajustament, al ser realitzades abans de veure la mostra S, se'n deriva una nova qüestió ineludiblement. Com s'escolleixen aquestes restriccions? Sota quin criteri s'escull una família de funcions enlloc d'una altra? Idealment, aquesta elecció s'hauria de basar en coneixement previ sobre el problema que es busca resoldre.

Observació: La restricció de $\mathcal{H}$, així com la forma i la raó d'aquesta, formalitzen la noció de coneixement previ que s'havia deixat entreveure a la introducció del treball.

Reprenent momentàniament l'exemple del zoòleg i la nova espècie de cavall; en aquest cas l'aprenent (el zoòleg) podria escollir com a classe d'hipòtesis la família de les funcions polinòmiques de grau n, amb n fixat, en el pla de $\mathbb{R}^2$ per separar l'espai en dos subespais; els exemplars amb sobrepés i els exemplars sense sobrepés. A l'hora d'escollir aquesta familia de funcions, ho faria basant-se en el seu coneixement previ sobre el comportament del sobrepés en funció de la llargada de la columna i del pes en altres mamífers similars (com per exemple una espècie ja coneguda de cavall).

Deixant de banda aquest incís, una pregunta més interessant que quins són els criteris sota els quals s'escull una classe d'hipòtesis enlloc d'una altra, és a dir quines restriccions s'imposen sobre $\mathcal{H}$, és quines condicions garanteixen que l'$\text{ERM}_{\mathcal{H}}$ no caurà en sobreajustament. Aquesta incògnita és equivalent a plantejar-se quines classes d'hipòtesis poden ser apreses mitjançant l'ERM. Aquesta és una qüestió fonamental en la teoria de l'aprenentatge i s'estudiarà diverses vegades, en funció de les premises sota les quals s'estigui treballant en cada moment, al llarg d'aquesta memòria.

La primera i més simple restricció que es pot imposar sobre una classe d'hipòtesis és acotar superiorment la seva mida, és a dir el nombre d'elements (predictors) que en formen part. Aquesta condició és suficient per tal de garantir que, sobre una classe d'hipòtesis $\mathcal{H}$, a partir d'una certa mida de la mostra S (la qual depén de la mida de $\mathcal{H}$) l'$\text{ERM}_{\mathcal{H}}$ no sobreajustarà. Aquest resultat està demostrat posteriorment en aquest mateix capítol del treball. No obstant, abans de passar a la demostració d'aquest resultat, cal deixar constància de les assumpcions sota les quals es treballarà.

En primer lloc, es considerarà certa l'assumpció de realitzabilitat:

Definició 7. *L'assumpció de realitzabilitat ("Realizability Assumption") enuncia que existeix $h^* \in \mathcal{H}$ tal que $L_{(\mathcal{D},f)}(h^*) = 0$.*

És rellevant notar que aquesta asumpció estableix que, amb probabilitat 1 sobre les mostres aleatòries S, es complirà $L_{\mathcal{S}}(h*) = 0$, on les instàncies de les mostres són generades d'acord a una distribució $\mathcal{D}$ i classificades segons una funció f. De fet, és fàcil deduir que això significa que $f \in \mathcal{H}$. No obstant, és molt més interessant i rellevant l'error vertader, $L_{(\mathcal{D},f)}(h_S)$, que no pas l'error empíric.

Per altra banda, com que l'aprenent només té accés a la mostra S i aquesta està generada per una distribució $\mathcal{D}$ de la qual no en té cap informació, qualsevol garantia sobre l'error respecte la distribució subjacent dependrà ineludibement de la relació entre la mostra $\mathcal{S}$ i la distribució $\mathcal{D}$. Aquesta relació es formalitza en la següent definició:

Definició 8. *L'assumpció i.i.d. estableix que les instàncies d'una mostra S són independentment idènticament distribuides (i.i.d.) segons la distribució $\mathcal{D}$.*

Aquesta assumpció es denota generalment mitjançant la notació $S \sim \mathcal{D}^m$.

A mode de breu recordatori, el resultat que es vol demostrar és que sobre una classe d'hipòtesis $\mathcal{H}$ finita, a partir d'una mida m donada de la mostra S, l'ERM$_{\mathcal{H}}$ no sobreajustarà; és a dir, es vol garantir que $L_{(\mathcal{D},f)}(h_S)$ prendrà un valor raonable.

Trivialment $L_{(\mathcal{D},f)}(h_S)$ depén de l'elecció del predictor h_S que realitzi l'algoritme d'aprenentatge, la qual depén alhora de la mostra S que rep aquest. Aquesta situació és potencialment problemàtica, ja que S està generada de manera completament aleatòria segons la distribució $\mathcal{D}$. Però, per què pot resultar problemàtic aquest fet? Aquesta situació implica l'existència d'un factor d'atzar en la generació de S i, per tant, en l'elecció de h_S; aleshores, inevitablement l'error $L_{(\mathcal{D},f)}(h_S)$ tindrà certa aleatorietat conseqüència de la seva pròpia definició. Vist això, es parla de $L_{(\mathcal{D},f)}(h_S)$ com una variable aleatòria (v.a.). Intuitivament, la probabilitat que la mostra S obtinguda sigui molt poc representativa de $\mathcal{D}$ no serà mai zero, per molt gran que sigui la mostra S. Com a conseqüència, no es pot garantir amb probabilitat 1 que l'error $L_{(\mathcal{D},f)}(h_S)$ no sigui gaire gran, sinò que simplement s'haurà de treballar en base a la probabilitat que $L_{(\mathcal{D},f)}(h_S)$ tingui un valor raonable. Per tal de realitzar aquest ajust, es denota la probabilitat d'obtenir una mostra S que sigui no representativa de $\mathcal{D}$ com un paràmetre δ, denotant llavors $1-\delta$ la probabilitat d'obtenir una mostra representativa; aquest paràmetre $1-\delta$ rep el nom de paràmetre de confiança.

Com probablement el lector haurà notat, fins ara s'ha parlat de la mida dels errors (generalment de $L_{(\mathcal{D},f)}(h_S)$) de manera merament arbitrària i imprecisa mitjançant expressions com "no massa gran"o "raonable", entre d'altres, enlloc de parlar d'error nul o de donar-li un valor o cota concreta. Si bé resultaria ideal treballar en base a garantir (amb probabilitat igual o major a $1-\delta$) error nul, d'acord a la definició de $L_{(\mathcal{D},f)}(h_S)$, $L_{(\mathcal{D},f)}(h_S) = 0$ si i només si, $h_S = f$. Per tant, exceptuant aquest cas concret, la majoria d'ocasions $h_S \neq f$ i per tant $L_{(\mathcal{D},f)}(h_S) > 0$. Així doncs, tan sols es pot aspirar a acotar superiorment $L_{(\mathcal{D},f)}(h_S)$ per un valor per sota del qual l'error es consieri acceptable. Per tal d'assolir-ho, s'introdueix un nou paràmetre ε, anomenat paràmetre de qualitat, el qual té l'objectiu de mesurar la qualitat d'un predictor. Un cop fixat aquest valor, s'interpreta l'esdeveniment $L_{(\mathcal{D},f)}(h_S) > \varepsilon$ com un fracàs de l'aprenent, mentre que si $L_{(\mathcal{D},f)}(h_S) \leq \varepsilon$ es valora h_S com un predictor aproximadament correcte.

Un cop definits aquests conceptes, fixada una certa funció $f\colon \mathcal{X} \to \mathcal{Y}$, es pot formalitzar la noció qualitativa de garantir que $L_{(\mathcal{D},f)}(h_S)$ sigui no gaire gran reescrivint-ho com garantir $\underset{S\sim\mathcal{D}^m}{\mathbb{P}}(L_{(\mathcal{D},f)}(h_S) \leq \varepsilon) \geq 1-\delta$ o, equivalentment, $\underset{S\sim\mathcal{D}^m}{\mathbb{P}}(L_{(\mathcal{D},f)}(h_S) > \varepsilon) < \delta$. Aquesta noció és equivalent a acotar superiorment la probabilitat d'obtenir una mostra S de m elements tal que condueixi a l'aprenent al fracàs (és a dir, que sigui no representativa); per

tant, l'objectiu no és altre que, essent $S = ((x_1, y_1), ..., (x_m, y_m))$ una mostra d'elements de $\mathcal{X} \times \mathcal{Y}$ on $S|_x = (x_1, ..., x_m)$ i $S \sim \mathcal{D}^m$, acotar superiorment $\mathcal{D}^m(\{\mathcal{S}|_x : L_{(\mathcal{D},f)}(h_\mathcal{S}) > \varepsilon\})$.

Proposició 9. *Sigui $\mathcal{X}$ un conjunt de domini, $\mathcal{Y} = \{0, 1\}$, sigui $\mathcal{H}$ una classe d'hipòtesis $h: \mathcal{X} \to \mathcal{Y}$ de mida finita $|\mathcal{H}|$, sigui $S = ((x_1, y_1), ..., (x_m, y_m))$ una mostra d'elements de $\mathcal{X} \times \mathcal{Y}$ on $S|_x = (x_1, ..., x_m)$ i $S \sim \mathcal{D}^m$. Si es compleix l'assumpció de realitzabilitat, per a tot $\varepsilon \in (0, 1)$, sigui h_S el predictor seleccionat per l'algoritme $ERM_\mathcal{H}$, aleshores es compleix:*

$$\mathcal{D}^m(\{S|_x : L_{(\mathcal{D},f)}(h_S) > \varepsilon\}) \leq |\mathcal{H}|e^{-\varepsilon m}.$$

Demostració:
Sigui $\mathcal{H}_B$ el següent conjunt:

$$\mathcal{H}_B = \{h \in \mathcal{H} : L_{(\mathcal{D},f)}(h) > \varepsilon\}.$$

És a dir, el conjunt de les hipòtesis amb un error superior a l'acceptable. Per simplicitat, se l'anomena conjunt de "males"hipòtesis.

Per altra banda, sigui M el següent conjunt:

$$M = \{S|_x : \exists h \in \mathcal{H}_B, L_S(h) = 0\}.$$

És a dir, el conjunt de les mostres tals que existeix almenys una hipòtesis de $\mathcal{H}$ amb error real inacceptable però error empíric zero. Intuitivament, és el conjunt de les mostres que poden enganyar a l'aprenent i fer-lo fracassar, ja que hi ha almenys una mala hipòtesis que funciona molt bé a la mostra.

Com que en tot moment s'està acceptant com a certa l'assumpció de realitzabilitat, es complirà sempre $L_S(h_S) = 0$. Així doncs, per tal que tingui lloc l'esdeveniment que s'està estudiant, $L_{(\mathcal{D},f)}(h_S) > \varepsilon$, han de succeir simultàniament $L_{(\mathcal{D},f)}(h_S) > \varepsilon$ i $L_S(h_S) = 0$.
És a dir, $L_{(\mathcal{D},f)}(h_S) > \varepsilon \Leftrightarrow \exists\ h \in \mathcal{H}_B$ tal que $L_S(h) = 0$. És a dir, aquest esdeveniment tan sols pot succeir si la mostra que rep l'aprenent pertany al conjunt M.

Així doncs:

$$\{\mathcal{S}|_x : L_{(\mathcal{D},f)}(h_\mathcal{S}) > \varepsilon\} \subseteq M.$$

Per tant, per demostrar el resultat n'hi ha prou amb acotar la probabilitat d'obtenir una mostra de M.
A més a més, noti's que, per la seva pròpia definició, M es pot reescriure com:

$$M = \bigcup_{h \in \mathcal{H}_B} \{S|_x : L_S(h) = 0\}.$$

Així doncs, unint-ho tot s'obté:

$$\mathcal{D}^m(\{\mathcal{S}|_x : L_{(\mathcal{D},f)}(h_S) > \varepsilon\}) \leq \mathcal{D}^m(M) = \mathcal{D}^m\left(\bigcup_{h \in \mathcal{H}_B} \{S|_x : L_S(h) = 0\}\right).$$

El següent pas de la demostració consisteix en acotar aquesta darrera expressió. Com que no és altra cosa que la probabilitat d'una unió, aplicant la propietat de la sub-additivitat es pot acotar fàcilment de la següent forma:

$$\mathcal{D}^m(\{S|_x : L_{(\mathcal{D},f)}(h_S) > \varepsilon\}) \leq \sum_{h \in \mathcal{H}_B} \mathcal{D}^m\left(\{S|_x : L_S(h) = 0\}\right).$$

Tot seguit, s'acotarà cadascun dels sumants d'aquest sumatori. Fixi's un predictor $h \in \mathcal{H}_B$; l'esdeveniment $L_S(h) = 0$ és equivalent a l'esdeveniment $h(x_i) = f(x_i)$ per a qualsevol i. Per tant, com que en tot moment s'està considerant certa l'assumpció i.i.d., aplicant que totes les instàncies estan generades i.i.d. s'obté:

$$\mathcal{D}^m\left(\{S|_x : L_S(h) = 0\}\right) = \mathcal{D}^m\left(\{S|_x : \forall i, h(x_i) = f(x_i)\}\right) = \prod_{i=i}^{m} \mathcal{D}\left(\{x_i : h(x_i) = f(x_i)\}\right).$$

Per tant, per cada extracció individual de la mostra es té:

$$\mathcal{D}\left(\{h(x_i) = y_i\}\right) = 1 - L_{(\mathcal{D},f)}(h) \leq 1 - \varepsilon.$$

On s'ha fet servir a l'última desigualtat que, com $h \in \mathcal{H}_B$, aleshores $L_{\mathcal{D},f}(h) > \varepsilon$. Així doncs, aplicant-ho a tota la mostra i no només a una instància s'obté:

$$\mathcal{D}^m\left(\{S|_x : L_S(h) = 0\}\right) \leq (1 - \varepsilon)^m \leq e^{-\varepsilon m}.$$

On a l'última desigualtat s'ha emprat que $1 - \varepsilon \leq e^{-\varepsilon}$. Finalment, combinant els darrers resultats i generalitzant per a tot el conjunt $\mathcal{H}_B$ (enlloc de considerar un sol predictor h) s'obté:

$$\mathcal{D}^m(\{S|_x : L_{(\mathcal{D},f)}(h_S) > \varepsilon\}) \leq |\mathcal{H}_B| e^{-\varepsilon m} \leq |\mathcal{H}| e^{-\varepsilon m} < \infty.$$

Tal i com es volia demostrar. □

Destaca que en cap moment de la proposició ni de la seva demostració ha aparegut el paràmetre de confiança $1 - \delta$. Això es deu que l'objectiu era justament acotar aquesta probabilitat a la que fa referència el paràmetre de confiança; de fet, imposant que aquesta cota obtinguda a la proposició referent a la probabilitat que falli l'aprenent sigui menor que δ, s'obté el següent corol·lari:

Corol·lari 10. *Sigui $\mathcal{H}$ una classe d'hipòtesis finita, siguin $\delta \in (0, 1)$ i $\epsilon > 0$, sigui m un enter tal que satisfà*

$$m \geq \frac{\log(|\mathcal{H}|/\delta)}{\varepsilon}.$$

Aleshores, donat un conjunt de domini $\mathcal{X}$, un conjunt d'etiquetes $\mathcal{Y} = \{0, 1\}$, per qualsevol funció de classificació f i per qualsevol distribució $\mathcal{D}$ tal que es compleixi l'assumpció de realitzabilitat, amb probabilitat major o igual a $1 - \delta$ sobre l'elecció d'una mostra S i.i.d. de m elements de $\mathcal{X} \times \mathcal{Y}$, es té que per qualsevol hipòtesi ERM, h_S, es compleix:

$$L_{(\mathcal{D},f)}(h_S) \leq \varepsilon. \tag{2.0.1}$$

Demostració.
Es vol demostrar que per a qualsevol $m \geq \frac{\log(|\mathcal{H}|/\delta)}{\varepsilon}$ es compleix

$$\mathcal{D}^m(\{S|_x : L_{(\mathcal{D},f)}(h_S) > \varepsilon\}) < \delta.$$

O, equivalentment,

$$\mathcal{D}^m(\{S|_x : L_{(\mathcal{D},f)}(h_S) \leq \varepsilon\}) \geq 1 - \delta.$$

Tal i com s'ha vist a la proposició 10, es té:

$$\mathcal{D}^m(\{S|_x : L_{(\mathcal{D},f)}(h_S) > \varepsilon\}) \leq |\mathcal{H}|e^{-\varepsilon m}.$$

L'expressió $|\mathcal{H}|e^{-\varepsilon m}$ decreix a mesura que m augmenta, per tant, com que δ és un valor constant, existeix $m \in \mathbb{N}$ tal que

$$\mathcal{D}^m(\{S|_x : L_{(\mathcal{D},f)}(h_S) > \varepsilon\}) \leq |\mathcal{H}|e^{-\varepsilon m} < \delta.$$

Així doncs, aïllant la m per tal de trobar el valor concret:

$$|\mathcal{H}|e^{-\varepsilon m} \leq \delta \Leftrightarrow \frac{|\mathcal{H}|}{\delta} \leq e^{\varepsilon m} \Leftrightarrow \frac{\log(|\mathcal{H}|/\delta)}{\varepsilon} \leq m.$$

Que és el que es volia provar. □

Capítol 3

Primer model formal d'aprenentatge

Al capítol anterior s'ha presentat una definició formal del problema que es tracta en aquesta memòria, així com un seguit d'assumpcions inicials que s'han realitzat sobre el mateix. Finalment, s'ha provat que per qualsevol classe d'hipòtesis de mida finita, si l'algoritme ERM relatiu a aquesta classe s'aplica a una mostra d'entrenament suficientment gran, es pot garantir amb certa probabilitat sobre l'elecció d'aquesta mostra que el predictor retornat per l'algoritme tindrà un error raonable. És a dir, es pot garantir que el predictor retornat per l'algoritme serà probablement aproximadament correcte. Aquesta noció es formalitzarà en el primer model formal d'aprenentatge que es tractarà, l'aprenentatge probablement aproximadament correcte (PAC), originalment "Probably Aproximately Correct (PAC) learning". Tot seguit, es proporciona una definició formal de la noció que una classe d'hipòtesis sigui aprensible mitjançant aprenentatge PAC, és a dir, que sigui aprensible PAC ("PAC learnable").

Definició 11. *Sigui $\mathcal{H}$ una classe d'hipòtesis. Diem que $\mathcal{H}$ és aprensible PAC si existeix una funció $m_{\mathcal{H}} : (0, 1)^2 \to \mathbb{N}$ i un algoritme d'aprenetatge tal que: Siguin ε , $\delta \in (0, 1)$ qualsevols, sigui $\mathcal{D}$ una distribució qualsevol sobre $\mathcal{X}$ i sigui $f: \mathcal{X} \to \{0, 1\}$ una funció d'etiquetatge qualsevol, si l'assumpció de realitzabilitat es compleix respecte $\mathcal{D}$, $\mathcal{H}$ i f, aleshores quan s'executi l'algoritme amb $m \geq m_{\mathcal{H}}(\varepsilon, \delta)$ i.i.d exemples generats d'acord a $\mathcal{D}$ i etiquetats d'acord a f, l'algoritme retornarà una hipòtesi h tal que, amb probabilitat igual o major a $1 - \delta$ sobre l'elecció dels exemples, $L_{(\mathcal{D},f)}(h) \leq \varepsilon$.*

Hom es pot fixar que en aquesta definició apareixen els paràmetres de precisió i confiança ε i δ que han aparegut al capítol anterior, els quals són respectivament els responsables dels adverbis *aproximadament* i *probablement* que acompanyen a la paraula *correcte*. Aquestes imprecisions, degut al model d'accés a les dades i la naturalesa de les mateixes, són inevitables tal i com s'ha posat de manifest a l'última secció del capítol anterior.
Parant atenció a la definició d'aprensibilitat PAC, és fàcil notar que tota la definició gira al voltant de l'existència de la funció $m_{\mathcal{H}} : (0, 1)^2 \to \mathbb{N}$, si bé també es requereix de l'existència d'un algoritme d'aprenentatge competent. En altres paraules, el que realment caracteritza l'aprensibilitat o la no aprensibilitat PAC d'una classe d'hipòtesis $\mathcal{H}$ donada és l'existència d'aquesta funció $m_{\mathcal{H}}$ (ε, δ), la qual determina la *complexitat de la mostra* ("*sample complexity*"); és a dir, la mida mínima requerida a tenir per la mostra de dades

d'entrenament per tal de garantir amb probabilitat major o igual a $1 - \delta$ l'aprenentatge amb error menor o igual a ε de la classe d'hipòtesis $\mathcal{H}$. És fàcil notar que, per una mateixa classe $\mathcal{H}$ que sigui aprensible PAC, existeixen infinites funcions $m_{\mathcal{H}}$ que satisfan els requeriments de la definció d'aprensibilitat PAC; per tant, noti's que la complexitat de la mostra es defineix com la mida mínima (equivalentment funció mínima), és a dir l'enter més petit que satisfà els requeriments de la definició (per a uns certs valors de ε i δ donats).

3.1 Aprenentatge PAC agnòstic

La noció d'aprensibilitat PAC, malgrat útil i consistent com a primer model formal d'aprenentatge, és certament limitada, ja que no només té diversos requeriments, sinò que a més a més alguns d'aquests són bastant restrictius. En concret, per tal que una classe d'hipòtesis $\mathcal{H}$ sigui aprensible, necessàriament han d'existir una funció $m_{\mathcal{H}}$ com la requerida a la definició, un algoritme d'aprenentatge suficientment capaç i, a més a més, cal que es compleixi l'assumpció de realitzabilitat. Aquesta última és una assumpció molt forta; fins ara en tot moment s'ha assumit l'existència d'una distribució $\mathcal{D}$ sobre el conjunt de domini $\mathcal{X}$ segons la qual es generen les dades i l'existència d'una funció f la qual determina completament per a cada instància la classificació que li pertoca. No obstant, a l'hora d'abordar un problema real, el més probable és que aquesta condició no es compleixi; ja sigui perquè les magnituds considerades a les instàncies no són suficients per determinar unívocament la classificació, o perquè la pròpia naturalesa del problema provoca que no existeixi cap funció correcta. Recuperant l'exemple dels cavalls, podria ser que la longitut de la columna i el pes no determinin unívocament la condició (o no) de sobrepés; podria ser necessari considerar també la densitat òssea, l'amplitud de la caixa toràssica, etc. Per tant, per tal d'ampliar aquesta idea de model d'aprenentatge i fer que sigui aplicable a més situacions, cal relaxar aquesta assumpció.

A partir d'ara, ja no es considerarà una distribució $\mathcal{D}$ sobre $\mathcal{X}$, sinò que es considerarà una distribució $\mathcal{D}$ sobre el conjunt $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$, on $\mathcal{X}$ és el conjunt de domini i $\mathcal{Y} = \{0, 1\}$ el conjunt de classificacions; és a dir, $\mathcal{D}$ és una distribució conjunta sobre punts del domini i del conjunt de classificació. Una perspectiva interessant i més intuitiva des de la qual pensar aquesta distribució és separant-la en una distribució marginal $\mathcal{D}_x$ sobre el conjunt de domini $\mathcal{X}$ i en una probabilitat condicionada $\mathcal{D}((x, y)|x)$ sobre les classificacions per a cada punt del domini.

Un cop modificada la pròpia definició de $\mathcal{D}$, com a conseqüència, també caldrà ajustar (o com a mínim revisar) la definició de totes les nocions que en depenguin. En primer lloc, l'error real es redefinirà de la següent manera:

Definició 12. *Sigui $\mathcal{X}$ un conjunt de domini, sigui $\mathcal{Y} = \{0, 1\}$ un conjunt de classificacions, sigui $\mathcal{D}$ una distribució qualsevol sobre $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$. Aleshores per a un predictor h: $\mathcal{X} \to \mathcal{Y}$ qualsevol, es pot mesurar la probabilitat que h cometi un error per al cas on els punts classificats (x, y) són aleatòriament generats en base a $\mathcal{D}$ (és a dir, que donat $(x,y) \in \mathcal{X} \times \mathcal{Y}$, es tingui $h(x) \neq y$) mitjançant el seu error real, el qual es defineix com:*

$$L_{\mathcal{D}}(h) := \mathop{\mathbb{P}}_{(x,y)\sim\mathcal{D}}[h(x) \neq y] := \mathcal{D}(\{(x, y) : h(x) \neq y\}).$$

És a dir, la probabilitat de seleccionar d'acord a $\mathcal{D}$ un element $(x, y) \in \mathcal{X} \times \mathcal{Y}$ tal que

$y \neq h(x)$.

Pel que fa a l'error empíric, com que aquest no depén en cap moment de la distribució $\mathcal{D}$, ja que tan sols es mesura en base a la mostra S, se segueix definint de la mateixa forma que abans.

Convé recordar que l'aprenent desconeix completament la distribució $\mathcal{D}$, l'única informació a la qual té accés són les dades de la mostra. Aquest fet cobra especial rellevància ara que la distribució $\mathcal{D}$ és sobre $\mathcal{X} \times \{0, 1\}$ enlloc de sobre $\mathcal{X}$, ja que aleshores, tal i com es demostrarà posteriorment, cap predictor h: $\mathcal{X} \to \mathcal{Y}$ té un error real menor que l'anomenat predictor òptim de Bayes, el qual es defineix de la següent manera:

Definició 13. *Sigui $\mathcal{X}$ un conjunt de domini, sigui $\mathcal{Y} = \{0, 1\}$ un conjunt de classificacions, sigui $\mathcal{D}$ una distribució de probabilitat qualsevol sobre $\mathcal{X} \times \{0, 1\}$. S'anomena predictor òptim de Bayes a la funció $f_{\mathcal{D}}$:$\mathcal{X} \to \{0, 1\}$, la qual és de la forma:*

$$f_{\mathcal{D}}(x) = \begin{cases} 1 & \text{si } \mathbb{P}[y{=}\,1|x] \geq 1/2\ , \\ 0 & \text{altrament.} \end{cases}$$

Proposició 14. *Sigui $\mathcal{X}$ un conjunt de domini, sigui $\mathcal{Y} = \{0, 1\}$ un conjunt de classificacions, sigui $\mathcal{D}$ una distribució de probabilitat qualsevol sobre $\mathcal{X} \times \{0, 1\}$. Aleshores per a qualsevol h: $\mathcal{X} \to \{0, 1\}$, $L_{\mathcal{D}}(f_{\mathcal{D}}) \leq L_{\mathcal{D}}(h)$.*

Aquesta proposició afirma que el predictor òptim de Bayes és, efectivament, òptim; és a dir, no n'hi ha cap altre amb un error real menor.

Demostració:

Sigui $x \in \mathcal{X}$, siguin X i Y variables aleatòries (v.a.) que denoten punts de $\mathcal{X}$ i classificacions de $\mathcal{Y}$ respectivament. Es denota λ_x la probabilitat condicionada que el punt x donat sigui classificat com a 1. Consideri's ara la probabilitat que un punt x sigui mal classificat:

$$\begin{aligned}
&\mathop{\mathbb{P}}_{(x,y)\sim\mathcal{D}}[f_{\mathcal{D}}(X) \neq y|X = x] \\
&= \mathbb{1}_{[\lambda_x \geq 1/2]} \times \mathop{\mathbb{P}}_{(x,y)\sim\mathcal{D}}[Y = 0|X = x] + \mathbb{1}_{[\lambda_x < 1/2]} \times \mathop{\mathbb{P}}_{(x,y)\sim\mathcal{D}}[Y = 1|X = x] \\
&= \mathbb{1}_{[\lambda_x \geq 1/2]} \times (1 - \lambda_x) + \mathbb{1}_{[\lambda_x < 1/2]} \times \lambda_x = \min\{\lambda_x, 1 - \lambda_x\}.
\end{aligned}$$

Sigui h: $\mathcal{X} \to \{0, 1\}$ un classificador qualsevol, es té:

$$\begin{aligned}
&\mathop{\mathbb{P}}_{(x,y)\sim\mathcal{D}}[h(X) \neq Y|X = x] = \\
&= \mathop{\mathbb{P}}_{x\sim\mathcal{D}|_x}[h(X) = 0|X = x] \times \mathop{\mathbb{P}}_{(x,y)\sim\mathcal{D}}[Y = 1|X = x] \\
&\qquad\qquad + \mathop{\mathbb{P}}_{x\sim\mathcal{D}|_x}[h(X) = 1|X = x] \times \mathop{\mathbb{P}}_{(x,y)\sim\mathcal{D}}[Y = 0|X = x] \\
&= \mathop{\mathbb{P}}_{x\sim\mathcal{D}|_x}[h(X) = 0|X = x] \times \lambda_x + \mathop{\mathbb{P}}_{x\sim\mathcal{D}|_x}[h(X) = 1|X = x] \times (1 - \lambda_x) \\
&\geq \mathop{\mathbb{P}}_{x\sim\mathcal{D}|_x}[h(X) = 0|X = x] \times \min\{\lambda_x, 1 - \lambda_x\} \\
&\qquad\qquad + \mathop{\mathbb{P}}_{x\sim\mathcal{D}|_x}[h(X) = 1|X = x] \times \min\{\lambda_x, 1 - \lambda_x\} \\
&= \min\{\lambda_x, 1 - \lambda_x\}.
\end{aligned}$$

A la primera igualtat s'ha fet servir que $h(X)$ i Y són variables independents, ja que si bé ambdues depenen del valor de X, aquesta dependència desapareix al condicionar la probabilitat. Com que aquest resultat és cert per qualsevol $x \in \mathcal{X}$, aplicant la llei de l'esperança total al predictor òptim de Bayes es té:

$$\begin{aligned} L_{\mathcal{D}}(f_{\mathcal{D}}) &= \mathbb{E}_{(x,y)\sim\mathcal{D}}[\mathbb{1}_{[f_{\mathcal{D}}(x)\neq y]}] \\ &= \mathbb{E}_{x\sim\mathcal{D}_X}[\mathbb{E}_{y\sim\mathcal{D}_{Y|x}}[\mathbb{1}_{[f_{\mathcal{D}(x)\neq y]}}|X=x]] = \mathbb{E}_{x\sim\mathcal{D}_X}[\lambda_x] \\ &\leq \mathbb{E}_{x\sim\mathcal{D}_X}[\mathbb{E}_{y\sim\mathcal{D}_{Y|x}}[\mathbb{1}_{[h(x)\neq y]}]|X=x]] = L_{\mathcal{D}}(h). \end{aligned}$$

□

Fruit d'aquesta propietat, es diu que (donats $\mathcal{X}$, $\mathcal{Y} = \{0, 1\}$, $\mathcal{D}$ sobre $\mathcal{X} \times \{0, 1\}$) el predictor òptim de Bayes és la millor funció de classificació possible de $\mathcal{X}$ a $\{0, 1\}$. Malgrat conèixer quin és el millor predictor possible, el problema de trobar-ne un amb el mínim error real possible en base a la mostra segueix vigent, ja que aquest predictor només té sentit emprar-lo quan la distribució $\mathcal{D}$ és coneguda (tal i com s'observa del fet que la pròpia definció de $f_{\mathcal{D}}$ depén de la probabilitat d'un cert valor de $y \in \mathcal{Y} = \{0, 1\}$ condicionada a un valor fixat de $x \in \mathcal{X}$, la qual depén directament de la distribució $\mathcal{D}$).

Convé aturar-se un instant abans de proseguir per destacar una observació de gran rellevança, la qual es deriva de la definició del predictor òptim de Bayes, així com del seu estatus de millor predictor possible. De la definició del predictor òptim de Bayes es dedueix que el seu error real no serà mai 0 (tan sols si la probabilitat condicionada de $y = 1$ a un valor fixat de x fos sempre 1 o 0 respectivament; fet que implicaria que x determina completament y i que, per tant, existeix una funció determinista f: $\mathcal{X} \to \{0, 1\}$, arribant així altre cop a l'assumpció de realitzabilitat), ja que pels valors de x tals que $\mathbb{P}[y = 1|x] \neq 0$ o $\mathbb{P}[y = 1|x] \neq 1$ la probabilitat que el predictor òptim de Bayes classifiqui malament serà superior a 0 (convé recordar i tenir present la nova definició d'error real amb la que s'està treballant). Així doncs, l'error real del predictor òptim de Bayes serà sempre més gran que 0, de fet serà algun valor λ constant. Aquest valor λ pot valer fins a $1/2$ (Exemple: llançar una moneda no esbiaixada), per tant no es pot requerir que sigui arbitràriament petit, sobretot tenint en compte que, al desconèixer $\mathcal{D}$, no es pot calcular. Com a conseqüència, com que qualsevol predictor de $\mathcal{X}$ a $\{0, 1\}$ té error igual o major a λ, no es pot requerir a l'aprenent una cota absoluta sobre l'error real del predictor que retorni, ja que aquest serà sempre major o igual a λ. Per tant, la mida absoluta de l'error ja no pot ser un requeriment, com sí que ho era en el model d'aprenentatge PAC; així doncs, cal adaptar aquesta noció. Per fer-ho, ja no es requerirà a l'algoritme d'aprenentatge que retorni un predictor que tingui un error absolutament petit, sinò que es requerirà que retorni un error que sigui comparativament petit; és a dir, que no sigui gaire major al menor error possible. Aquest nou model d'aprenentatge s'anomena aprenentatge PAC agnòstic ("agnostic PAC learning") i es formalitza en la següent definció:

Definció 15. *Sigui $\mathcal{D}$ una distribució qualsevol sobre $\mathcal{X} \times \mathcal{Y}$, on $\mathcal{X}$ és un conjunt de domini i $\mathcal{Y} = \{0, 1\}$ és un conjunt de classificacions. Es diu que una classe d'hipòtesis $\mathcal{H}$ és aprensible PAC agnòsticament si existeixen un algoritme d'aprenentatge i una funció $m_{\mathcal{H}} : (0, 1)^2 \to \mathbb{N}$ tals que per a qualssevol ε, $\delta \in (0, 1)$, en implementar l'algoritme amb una quantitat $m \geq m_{\mathcal{H}}(\varepsilon, \delta)$ d'exemples i.i.d. generats d'acord a $\mathcal{D}$, aleshores l'agoritme retornarà una hipòtesi $h: \mathcal{X} \to \{0, 1\}$ tal que, amb una probabilitat igual o major a $1 - \delta$, compleix:*

$$L_{\mathcal{D}}(h) \leq \min_{h'\in\mathcal{H}} L_{\mathcal{D}}(h') + \varepsilon.$$

Així doncs, quan no s'està sota el paraigües de l'assumpció de realitzabilitat, la noció d'aprenentatge PAC agnòstic ens garanteix que, donada una classe d'hipòtesis $\mathcal{H}$ amb les condicions adecuades, un algoritme d'aprenentatge adient és capaç de seleccionar un predictor $h \in \mathcal{H}$ l'error del qual no sigui gaire superior al de la millor funció de la classe.

A mode d'observació, tal i com s'ha donat a entendre de forma intuitiva al paràgraf previ a la definició d'aprensibilitat PAC agnòstica, si es pren com a certa l'assumpció de realitzabilitat aleshores es recupera la definició d'aprenentatge PAC (ja que $\min(L_{\mathcal{D}}(h')) = L_{\mathcal{D}}(h*) = 0$).

3.2 Funció de pèrdua

D'aquest nou model d'aprenentatge, com ja ha passat amb les nocions considerades previament, se'n deriven un seguit de problemàtiques i qüestions que cal abordar per tal de dotar de consistència el model a l'hora d'implementar-lo. En primer lloc, quines condicions ha de satisfer una classe d'hipòtesis $\mathcal{H}$ per tal que sigui aprensible PAC agnòsticament? Quines són necessàries? Quines són suficients? Aquestes preguntes no s'abordaran encara, ja que seran adecuadament resposes a la secció 4.2. En segon lloc, com es mesura l'error? Aquesta qüestió pot resultar desconcertant, ja que fins ara s'ha estat emprant l'error real, el qual està ben definit i, de moment, funciona correctament. No obstant, depenent de la naturalesa del problema que s'estigui abordant, l'error real pot resultar inconvenient. Per exemple, si hom considerés un problema de classificació en múltiples classes (Ex: classificar articles d'un diari segons la temàtica) o volgués traçar un patró a les dades, és a dir una relació funcional entre $\mathcal{X}$ i $\mathcal{Y}$, l'error real és una eina que pot quedar-se-li curta. Si bé en aquest treball no s'aborden situacions fora de la classificació binària fins a la secció 5.2, per solventar problemàtiques associades a aquesta qüestió s'introdueix el concepte de funció de pèrdua.

Definició 16. *Sigui $\mathcal{H}$ un conjunt qualsevol i sigui $\mathcal{Z}$ un conjunt de domini qualsevol. S'anomena funció de pèrdua a una funció qualsevol de la forma:*

$$\ell : \mathcal{H} \times \mathcal{Z} \to \mathbb{R}^+.$$

On pels problemes de predicció que es tractaran en aquesta publicació, $\mathcal{H}$ sempre serà una classe d'hipòtesis o de models i $\mathcal{Z}$ serà $\mathcal{Z} = \mathcal{X} \times \mathcal{Y} = \mathcal{X} \times \{0, 1\}$.

La funció de pèrdua ℓ pot prendre moltes formes, però aquest estudi es limitarà a considerar la funció de pèrdua $0 - 1$.

Definició 17. *Sigui $\mathcal{Z} = \mathcal{X} \times \{0, 1\}$ un conjunt de domini, sigui h: $\mathcal{X} \to \{0, 1\}$. Es defineix la funció de pèrdua $0 - 1$ com la funció ℓ_{0-1}: $\mathcal{H} \times \mathcal{Z} \to \mathbb{R}^+$ de la forma:*

$$\ell_{0-1}(h, (x, y)) := \begin{cases} 0 & \text{si } h(x) = y \ , \\ 1 & \text{si } h(x) \neq y. \end{cases}$$

Ara que ha sigut introduït el concepte de funció de pèrdua, ja es poden introduir els conceptes que, d'ara en endavant, substituiran l'error real i l'error empíric. Començant pel substitut de l'error real, es defineix la funció de risc com:

Definició 18. *Sigui $\mathcal{Z} = \mathcal{X} \times \{0, 1\}$ un domini, sigui $\mathcal{D}$ una distribució sobre $\mathcal{Z}$, sigui $\mathcal{H}$ una classe d'hipòtesis, sigui h: $\mathcal{X} \to \{0, 1\}$, $h \in \mathcal{H}$, un predictor, sigui ℓ: $\mathcal{H} \times \mathcal{Z}$ $\to \mathbb{R}^+$ una funció de pèrdua. Es defineix la funció de risc de h com:*

$$L_{\mathcal{D}}(h) := \underset{z\sim\mathcal{D}}{\mathbb{E}}[\ell(h,z)].$$

De manera similar es defineix el substituit de l'error empíric, el risc empíric, com:

Definició 19. *Sigui $\mathcal{Z} = \mathcal{X} \times \{0, 1\}$ un domini, sigui $S = (z_1, ..., z_m) \in \mathcal{Z}^m$, sigui $\mathcal{H}$ una classe d'hipòtesis, sigui h: $\mathcal{X} \to \{0, 1\}$, $h \in \mathcal{H}$, un predictor, sigui ℓ: $\mathcal{H} \times \mathcal{Z} \to \mathbb{R}^+$ una funció de pèrdua. Es defineix el risc empíric de h respecte S com:*

$$L_S(h) := \frac{1}{m}\sum_{i=1}^{m} \ell(h, z_i).$$

És interessant notar que, pel cas de classificació binària, sigui α una variable aleatòria que pot prendre els valors 0 o 1, es té $\mathbb{E}_{\alpha\sim\mathcal{D}}[\alpha] = \mathbb{P}_{\alpha\sim\mathcal{D}}[\alpha = 1]$. Per tant, pel cas de classificació binària (l'unic que es tractarà en aquest treball) l'error real i la funció de risc coincideixen, per tant és equivalent treballar amb una definició o l'altra.

Recordant la definció d'aprensibilitat PAC agnòstica, aquesta depén de l'error real i per tant, en funció de la naturalesa del problema, pot resultar inconvenient. Es pot generalitzar la noció d'aprensibilitat PAC agnòstica per a una funció de pèrdua qualsevol. Malgrat que pel cas que es tractarà no cal aquesta nova definició (ja que s'ha vist que la funció de risc i l'error real coincideixen), resulta interessant incloure-la.

Definició 20. *Sigui $\mathcal{D}$ una distribució qualsevol sobre un conjunt $\mathcal{Z}$, sigui $\mathcal{H}$ una classe d'hipòtesis, sigui $\ell : \mathcal{H} \times \mathcal{Z} \to \mathbb{R}_+$ una funció de pèrdua. Es diu que $\mathcal{H}$ és aprensible PAC agnòsticament respecte $\mathcal{Z}$ i ℓ si existeixen un algoritme d'aprenentatge i una funció $m_{\mathcal{H}} : (0, 1)^2 \to \mathbb{N}$ tals que per a qualssevol ε, $\delta \in (0, 1)$, en implementar l'algoritme amb una quantitat $m \geq m_{\mathcal{H}}(\varepsilon, \delta)$ d'exemples i.i.d. generats d'acord a $\mathcal{D}$, aleshores l'algoritme retornarà una hipòtesi h tal que, amb una probabilitat igual o major a $1-\delta$, compleix:*

$$L_{\mathcal{D}}(h) \leq \min_{h'\in\mathcal{H}} L_{\mathcal{D}}(h') + \varepsilon,$$

on $L_{\mathcal{D}}(h) := \underset{z\sim\mathcal{D}}{\mathbb{E}}[\ell(h,z)]$.

A mode d'observació, en aquesta última definició es tracta, per cada $h \in \mathcal{H}$, la funció $\ell(h, \cdot)$: $\mathcal{Z} \to \mathbb{R}^+$ com una variable aleatòria i es defineix la seva funció de risc $L_{\mathcal{D}}(h)$ com el valor esperat d'aquesta variable aleatòria. Per tant, cal que la funció $\ell(h, \cdot)$ sigui mesurable. Així doncs, s'assumeix l'existència d'una certa σ-àlgebra de subconjunts de $\mathcal{Z}$, sobre els quals la probabilitat $\mathcal{D}$ està definida i que conté la preimatge de qualsevol segment inicial de $\mathbb{R}^+$ (és a dir, segments de la forma $[0, a]$ o $[0, a)$, on $a \in \mathbb{R}^+$)). Més concretament, pel cas de classificació binària amb la funció de pèrdua $0 - 1$, la σ-àlgebra és sobre $\mathcal{X} \times \{0, 1\}$ i l'assumpció sobre ℓ és equivalent a considerar que, per a qualsevol $h \in \mathcal{H}$, el conjunt $\{(x, h(x)) : x \in \mathcal{X}\}$ està inclós a la σ-àlgebra.

3.3 Convergència uniforme

Malgrat que les nocions d'aprenetatge PAC i d'aprenentatge PAC agnòstic són útils i consistents, si hom repassa la seva definició pot notar que presenten un problema fruit de la contradicció entre l'objectiu de l'algoritme d'aprenentatge i el funcionament del mateix. En aquests models d'aprenentatge, l'aprenent selecciona una classe d'hipòtesis $\mathcal{H}$ abans de veure la mostra d'entrenament S, tot seguit l'algoritme $\text{ERM}_{\mathcal{H}}$ selecciona un predictor h

$\in \mathcal{H}$ tal que minimitzi el risc empíric respecte S amb l'esperança que també es minimitzi alhora l'error real. Per tant, l'èxit d'aquests models es basa en confiar que minimitzar l'error empíric minimitzi de rebot l'error real també, si bé no n'hi ha cap certesa. Així doncs, cal establir una garantia per tal que en minimitzar el risc empíric es minimitzi (o es redueixi fins a un valor proper al mínim) també la funció de risc. Per fer-ho, n'hi ha prou amb garantir que el risc empíric de tots els membres de $\mathcal{H}$ siguin bones aproximacions del seu error real; és a dir, que uniformement sobre totes les hipòtesis $h \in \mathcal{H}$, el risc empíric és proper a l'error real. Aquesta noció es formalitza en la següent definició.

Definició 21. *Sigui $\mathcal{Z}$ un domini, sigui $\mathcal{D}$ una distribució sobre $\mathcal{Z}$, sigui S una mostra d'entrenament, sigui $\mathcal{H}$ una classe d'hipòtesis, sigui ℓ una funció de pèrdua. Aleshores S és ε-representativa si:*

$$\forall h \in \mathcal{H}, \quad |L_S(h) - L_{\mathcal{D}}(h)| \leq \varepsilon.$$

Un resultat interessant fruit d'aquesta definició és que qualsevol mostra que sigui $(\varepsilon/2)$-representativa garanteix que l'algoritme ERM seleccionarà una bona hipòtesi. Formalment:

Lema 22. Sigui $\mathcal{Z}$ un domini, sigui $\mathcal{H}$ una classe d'hipòtesis, sigui ℓ: $\mathcal{H} \times \mathcal{Z} \to \mathbb{R}^+$ una funció de pèrdua, sigui $\mathcal{D}$ una distribució sobre $\mathcal{Z}$, sigui S una mostra $(\varepsilon/2)$-representativa d'elements de $\mathcal{Z}$ respecte $\mathcal{H}$, $\mathcal{D}$ i ℓ. Aleshores, per a tota $h_S \in \operatorname{argmin}_{h \in \mathcal{H}} L_S(h)$ se satisfà:

$$L_{\mathcal{D}}(h_S) \leq \min_{h \in \mathcal{H}} L_{\mathcal{D}}(h) + \varepsilon.$$

Demostració:
Per a tota $h \in \mathcal{H}$, com que S és $(\varepsilon/2)$-representativa i h_S és un predictor ERM es compleix:

$$L_{\mathcal{D}}(h_S) \leq L_S(h_S) + \frac{\varepsilon}{2} \leq L_S(h) + \frac{\varepsilon}{2} \leq L_{\mathcal{D}}(h) + \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = L_{\mathcal{D}}(h) + \varepsilon.$$

Aquest lema té una implicació molt interessant, per garantir que l'ERM és un aprenent PAC agnòstic, n'hi ha prou que amb probabilitat major o igualtat a $1 - \delta$ sobre l'elecció de la mostra d'entrenament, serà una mostra ε-representativa. Aquesta noció es formalitza en la definició de propietat de convergència uniforme.

Definició 23. *Sigui $\mathcal{Z} = \mathcal{X} \times \{0, 1\}$ un domini, sigui $\mathcal{H}$ una classe d'hipòtesis h: $\mathcal{X} \to \{0, 1\}$, sigui ℓ: $\mathcal{H} \times \mathcal{Z} \to \mathbb{R}^+$. Es diu que $\mathcal{H}$ té la propietat de convergència uniforme si existeix una funció $m_{\mathcal{H}}^{UC}$: $(0, 1)^2 \to \mathbb{N}$ tal que per a qualssevol ε, $\delta \in (0, 1)$ i qualsevol distribució $\mathcal{D}$ sobre $\mathcal{Z}$, si S és una mostra de $m \in \mathbb{N}$ exemples, on $m \geq m_{\mathcal{H}}^{UC}(\varepsilon, \delta)$, extrets i.i.d. segons $\mathcal{D}$, aleshores, amb probabilitat major o igual a $1 - \delta$, S és ε-representativa.*

De manera equivalent a $m_{\mathcal{H}}(\varepsilon, \delta)$ a la definició d'aprenentatge PAC i a la definició d'aprenentatge PAC agnòstic, $m_{\mathcal{H}}^{UC}(\varepsilon, \delta)$ indica la complexitat de la mostra per tal d'obtenir la propietat de convergència uniforme; és a dir, fixats ε i δ, quin és el nombre mínim d'exemples necessaris a la mostra per tal de garantir que amb probabilitat major o igual a $1 - \delta$ la mostra serà ε-representativa.

Combinant el lema 22 i la definició 23 s'obté el següent resultat rellevant:

Corol·lari 24. *Sigui $\mathcal{H}$ una classe d'hipòtesis amb la propietat de convergència uniforme per una funció $m_{\mathcal{H}}^{UC}$ donada. Aleshores $\mathcal{H}$ és aprensible PAC agnòsticament amb una complexitat de la mostra $m_{\mathcal{H}}(\varepsilon, \delta) \leq m_{\mathcal{H}}^{UC}(\varepsilon/2, \delta)$. De fet, en aquest cas el paradigma $ERM_{\mathcal{H}}$ és un aprenent exitós de $\mathcal{H}$.*

Demostració:
Siguin $\mathcal{Z}$, $\mathcal{H}$, $\mathcal{D}$, i ℓ el domini, la classe d'hipòtesis, la distribució i la funció de pèrdua respecte les quals la mostra S té la propietat de convergència uniforme. Com que $\mathcal{H}$ té la propietat de convergència uniforme, aleshores donats $\varepsilon/2$ i $\delta \in (0, 1)$, existeix $m_{\mathcal{H}}^{UC}(\varepsilon/2, \delta) \in \mathbb{N}$ tal que

$$\begin{aligned}&\mathop{\mathbb{P}}_{S\sim\mathcal{D}^m}(\{S \text{ és } \varepsilon/2\text{-representativa}\}) \geq 1-\delta \\ &\equiv \mathop{\mathbb{P}}_{S\sim\mathcal{D}^m}(\{\forall h \in \mathcal{H}, |L_S(h) - L_{\mathcal{D}}(h)| \leq \varepsilon/2\}) \geq 1-\delta.\end{aligned}$$

Pel lema 22, l'esdeveniment $\{\forall h \in \mathcal{H}, |L_S(h) - L_{\mathcal{D}}(h)| \leq \varepsilon/2\}$ és equivalent a l'esdeveniment $\{\ L_{\mathcal{D}}(h_S) \leq \min_{h\in\mathcal{H}} L_{\mathcal{D}}(h) + \varepsilon\ \}$. Així doncs, per a qualsevol $m \geq m_{\mathcal{H}}^{UC}(\varepsilon/2, \delta)$ es té :

$$\mathop{\mathbb{P}}_{S\sim\mathcal{D}^m}(\{L_{\mathcal{D}}(h_S) \leq \min_{h\in\mathcal{H}} L_{\mathcal{D}}(h) + \varepsilon\}) \geq 1-\delta.$$

Aquesta desigualtat és molt similar a la definició d'aprensibilitat PAC agnòstica, la qual diu que, donats ε, $\delta \in (0, 1)$, si la mostra té $m \geq m_{\mathcal{H}}(\varepsilon, \delta)$ exemples, aleshores l'ERM_H retornarà una $h \in \mathcal{H}$ tal que

$$\mathop{\mathbb{P}}_{S\sim\mathcal{D}^m}(\{L_{\mathcal{D}}(h) \leq \min_{h\in\mathcal{H}} L_{\mathcal{D}}(h) + \varepsilon\}) \geq 1-\delta.$$

No obstant, aquesta condició és més feble que l'anterior, ja que $h_S \in \text{argmin}_{h\in\mathcal{H}}\ L_S(h)$, però h no està garantit que també hi pertanyi. Per tant, es tindrà:

$$\begin{aligned}&\mathop{\mathbb{P}}_{S\sim\mathcal{D}^m}(\{L_{\mathcal{D}}(h) \leq \min_{h\in\mathcal{H}} L_{\mathcal{D}}(h) + \varepsilon\}) \\ &\geq \mathop{\mathbb{P}}_{S\sim\mathcal{D}^m}(\{L_{\mathcal{D}}(h_S) \leq \min_{h\in\mathcal{H}} L_{\mathcal{D}}(h) + \varepsilon\}) \geq 1-\delta.\end{aligned}$$

I, com a conseqüència, $m_{\mathcal{H}}(\varepsilon, \delta) \leq m_{\mathcal{H}}^{UC}(\varepsilon/2, \delta)$. De fet, pel lema 22 això se satisfà per l'$\text{ERM}_{\mathcal{H}}$, tal i com es volia provar. □

Aquest resultat més endavant servirà per demostrar que qualsevol classe d'hipòtesis finita és aprensible PAC agnòsticament. No obstant, per poder assolir-ho, és indispensable provar primer que tota classe d'hipòtesis finita té la propietat de convergència uniforme.

Proposició 25. *Tota classe d'hipòtesis $\mathcal{H}$ finita té la propietat de convergència uniforme.*

Per tal de realitzar aquesta demostració calen tot un seguit de resultats matemàtics que s'enunciaran i demostraran a continuació; posteriorment es demostrarà la proposició 25.

Lema 26 (Desigualtat de Jensen): *Sigui φ: $A \to \mathbb{R}$ una funció real convexa, on A és el seu domini, sigui x_1, ..., $x_n \in A$ nombres, on $n \in \mathbb{N}$, i siguin a_1, ..., a_n els seus pesos. Aleshores es compleix:*

$$\varphi\left(\frac{\sum_{i=1}^n a_i x_i}{\sum_{i=1}^n a_i}\right) \leq \frac{\sum_{i=1}^n a_i \varphi(x_i)}{\sum_{i=1}^n a_i}.$$

Per altra banda, si la funció φ és còncava, aleshores es té:

$$\varphi\left(\frac{\sum_{i=1}^n a_i x_i}{\sum_{i=1}^n a_i}\right) \geq \frac{\sum_{i=1}^n a_i \varphi(x_i)}{\sum_{i=1}^n a_i}.$$

Demostració:
Es provarà per inducció:
Es defineix $\gamma_i = \frac{a_i}{\sum_{j=1}^{n} a_j}$.
Cas inicial $n = 2$:
Siguin γ_1, γ_2 dos nombres reals no negatius qualssevol tals que $\gamma_1 + \gamma_2 = 1$, aleshores la convexitat de φ implica que, per a qualssevol x_1, x_2:

$$\varphi(\gamma_1 x_1 + \gamma_2 x_2) \leq \gamma_1 \varphi(x_1) + \gamma_2 \varphi(x_2).$$

Hipòtesi d'inducció:
Se suposarà cert per a $n \in \mathbb{N}$ elements; siguin γ_1, ..., γ_n n nombres reals no negatius qualssevol tals que $\gamma_1 + ... + \gamma_n = 1$, aleshores per a qualssevol x_1, ..., x_n es compleix:

$$\varphi\left(\sum_{i=1}^{n} \gamma_i x_i\right) \leq \sum_{i=1}^{n} \gamma_i \varphi(x_i).$$

Tan sols resta provar que el cas n implica el cas $n+1$. Com que, com a mínim algun dels γ_i és estrictament més petit que 1, prenent sense pèrdua de gerneralitat que és γ_{n+1}, per la convexitat de φ es té:

$$\begin{aligned}\varphi\left(\sum_{i=1}^{n+1} \gamma_i x_i\right) &= \varphi\left((1-\gamma_{n+1}) \sum_{i=1}^{n} \frac{\gamma_i x_i}{1-\gamma_{n+1}} + \gamma_{n+1} x_{n+1}\right) \\ &\leq (1-\gamma_{n+1}) \varphi\left(\sum_{i=1}^{n} \frac{\gamma_i x_i}{1-\gamma_{n+1}}\right) + \gamma_{n+1} \varphi(x_{n+1}).\end{aligned}$$

Per altra banda, com que $\sum_{i=1}^{n+1} \gamma_i = 1$, aleshores es compleix:

$$\sum_{i=1}^{n} \frac{\gamma_i}{1-\gamma_{n+1}} = 1.$$

Aplicant l'hipòtesi d'inducció s'obté

$$\varphi\left(\sum_{i=1}^{n} \frac{\gamma_i x_i}{1-\gamma_{n+1}}\right) \leq \sum_{i=1}^{n} \frac{\gamma_i \varphi(x_i)}{1-\gamma_{n+1}}.$$

Així doncs:

$$\varphi\left(\sum_{i=1}^{n+1} \gamma_i x_i\right) \leq (1-\gamma_{n+1}) \sum_{i=1}^{n} \frac{\gamma_i \varphi(x_i)}{1-\gamma_{n+1}} + \gamma_{n+1} \varphi(x_{n+1}).$$

Per tant, ha quedat provada la implicació $n \to n+1$ i, per tant, la desigualtat de Jensen. El cas de la concavitat és anàleg. □

Lema 27: (Desigualtat de la mitjana aritmètica i la mitjana geomètrica) *Siguin x_1, x_2, ..., x_n un conjunt de nombres reals. Aleshores se satisfà:*

$$\sqrt[n]{x_1 \cdot ... \cdot x_n} \leq \frac{x_1 + ... + x_n}{n}.$$

Demostració:
Com que la funció logaritme és còncava, per la desigualtat de Jensen es té:

$$\log\left(\frac{\sum_{i=1}^n x_i}{n}\right) \geq \frac{1}{n}\sum_{i=1}^n \log(x_i) = \frac{1}{n}\log\left(\prod_{i=1}^n x_i\right) = \log\left(\sqrt[n]{\prod_{i=1}^n x_i}\right)$$

I treient els logaritmes a ambdós costats s'obté:

$$\frac{\sum_{i=1}^n x_i}{n} \geq \sqrt[n]{\prod_{i=1}^n x_i}.$$

Tal i com es volia provar. □

Lema 28. (Lema de Hoeffding) *Sigui X una variable aleatòria tal que pren valors en $[a, b]$ i tal que $\mathbb{E}[X] = 0$. Aleshores, per a qualsevol $\lambda > 0$, és té*

$$\mathbb{E}[e^{\lambda X}] \leq e^{\frac{\lambda^2(b-a)^2}{8}}.$$

Demostració:
Com que $f(x) = e^{\lambda x}$ és una funció convexa, aleshores per a qualsevol $x \in [a, b]$, prenent $\gamma = \frac{b-x}{b-a} \in [0, 1]$ es compleix

$$f(x) \leq \gamma f(a) + (1-\gamma)f(b).$$

La desigualtat pren la forma

$$f(x) = e^{\lambda x} \leq \frac{b-x}{b-a}e^{\lambda a} + \frac{x-a}{b-a}e^{\lambda b}.$$

Prenent esperances

$$\mathbb{E}[e^{\lambda X}] \leq \frac{b-\mathbb{E}[X]}{b-a}e^{\lambda a} + \frac{\mathbb{E}[X]-a}{b-a}e^{\lambda b}.$$

I, com que $\mathbb{E}[X] = 0$, aleshores

$$\mathbb{E}[e^{\lambda X}] \leq \frac{b}{b-a}e^{\lambda a} - \frac{a}{b-a}e^{\lambda b}.$$

Denotant $h = \lambda(b-a)$, $p = \frac{-a}{b-a}$ i $L(h) = -hp + \log(1-p+pe^h)$, es pot comprovar que l'expressió de la dreta de la igualtat es pot reescriure de la forma $e^{L(h)}$:

$$\frac{b}{b-a}e^{\lambda a} - \frac{a}{b-a}e^{\lambda b} = \frac{b-a+a}{b-a}e^{\lambda a} - \frac{a}{b-a}e^{\lambda b} = e^{\lambda a}\left(1 + \frac{a}{b-a}\left(1 - e^{\lambda(b-a)}\right)\right) =$$

$$= e^{\lambda a} \cdot e^{\log\left(1+\frac{a}{b-a}-\frac{a}{b-a}e^{\lambda(b-a)}\right)} = e^{-hp+\log(1-p+pe^h)} = e^{L(h)}.$$

Per tant, n'hi ha prou amb provar que $L(h) \leq \frac{h^2}{8}$. Per fer-ho, s'empra el polinomi de Taylor:

$$L(h) = L(0) + L'(0)\cdot(h-0) + \frac{L^{(2)}(0)}{2!}\cdot(h-0)^2 + \mathcal{O}(h^3).$$

On s'aproximarà fins a termes de segon ordre. Trivialment, $L(0) = 0$. Per altra banda, com $L'(h) = -p + \frac{pe^h}{1-p+pe^h}$ i $L^{(2)}(h) = \frac{p(1-p)e^h}{(1-p+pe^h)^2}$, és directe veure que $L'(0) = 0$ i, mitjançant

la desigualtat de les mitjanes geomètrica i aritmètica, es pot provar que $L^{(2)}(h) \leq 1/4$ per a qualsevol h:

$$\frac{p(1-p)e^h}{(1-p+pe^h)^2} \leq \frac{1}{(1-p+pe^h)^2} \cdot \left(\frac{(1-p)+pe^h}{2}\right)^2 = \frac{1}{4}.$$

Substituïnt al polinomi de Taylor s'obté:

$$L(h) \leq 0 + 0 \cdot h + \frac{1}{4} \cdot \frac{h^2}{2} = \frac{h^2}{8}.$$

I per tant:

$$\mathbb{E}[e^{\lambda x}] \leq e^{L(h)} \leq e^{\frac{h^2}{8}} = e^{\frac{\lambda^2(b-a)^2}{8}}.$$

Tal i com es volia provar. □

Lema 29. (Desigualtat de Markov) *Sigui Z una variable aleatòria no negativa. Aleshores, per a qualsevol $a \geq 0$, es compleix:*

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

Demostració:
L'esperança de Z es pot escriure de la forma

$$\mathbb{E}[Z] = \int_{x=0}^{\infty} \mathbb{P}[Z \geq x]dx.$$

Com que $\mathbb{P}[Z \geq x]$ és monòtonament no creixent, s'obté per a qualsevol $a \geq 0$,

$$\mathbb{E}[Z] \geq \int_{x=0}^{a} \mathbb{P}[Z \geq x]dx \geq \int_{x=0}^{a} \mathbb{P}[Z \geq a]dx = a\mathbb{P}[Z \geq a]$$

Així doncs, per a qualsevol $a \geq 0$ es té:

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

Que és el que es volia provar. □

Lema 30. (Desigualtat de Hoeffding) *Sigui θ_1, ..., θ_m una seqüència de variables aleatòries i.i.d. i consideri's que per a qualsevol i, $\mathbb{E}[\theta_i] = \mu$ i $\mathbb{P}[a \leq \theta_i \leq b] = 1$. Aleshores, per a qualsevol $\varepsilon > 0$ es compleix:*

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

Demostració:
Siguin $\alpha_i = \theta_i$ - $\mathbb{E}[\theta_i]$, $\bar{\alpha} = \frac{1}{m}\sum_{i=1}^{m}\alpha_i$. Utilitzant la monotonia de la funció exponencial i la desigualtat de Markov, s'obté que per a qualssevol λ, $\varepsilon > 0$:

$$\mathbb{P}[\bar{\alpha_i} \geq \varepsilon] = \mathbb{P}[e^{\lambda\bar{\alpha}} \geq e^{\lambda\varepsilon}] \leq e^{-\lambda\varepsilon}\mathbb{E}[e^{\lambda\bar{\alpha}}].$$

Fent servir l'assumpció d'independència es té

$$\mathbb{E}[e^{\lambda\bar{\alpha}}] = \mathbb{E}\left[\prod_i e^{\lambda\alpha_i/m}\right] = \prod_i \mathbb{E}[e^{\lambda\alpha_i/m}].$$

Fent servir el lema de Hoeffding, per cada i es té

$$\mathbb{E}[e^{\lambda\alpha_i/m}] \leq e^{\frac{\lambda^2(b-a)^2}{8m^2}}.$$

Així doncs,

$$\mathbb{P}[\bar{\alpha} \geq \varepsilon] \leq e^{-\lambda\varepsilon} \prod_i e^{\frac{\lambda^2(b-a)^2}{8m^2}} = e^{-\lambda\varepsilon + \frac{\lambda^2(b-a)^2}{8m}}.$$

Prenent $\lambda = 4m\varepsilon/(b-a)^2$ s'obté

$$\mathbb{P}[\bar{\alpha} \geq \varepsilon] \leq e^{-\frac{2m\varepsilon^2}{(b-a)^2}}.$$

Anàlogament, repetint el mateix argument per la variable $-\bar{\alpha}$ s'obté $\mathbb{P}[\bar{\alpha} \leq -\varepsilon] \leq e^{-\frac{2m\varepsilon^2}{(b-a)^2}}$. Així doncs:

$$\mathbb{P}[-\varepsilon \leq \bar{\alpha} \leq \varepsilon] = 1 - \mathbb{P}[\bar{\alpha} \leq -\varepsilon] - \mathbb{P}[\varepsilon \leq \bar{\alpha}] = 1 - e^{-\frac{2m\varepsilon^2}{(b-a)^2}} - e^{-\frac{2m\varepsilon^2}{(b-a)^2}} = 1 - 2e^{-\frac{2m\varepsilon^2}{(b-a)^2}}.$$

Per tant

$$\mathbb{P}[|\bar{\alpha}| > \varepsilon] = 1 - \mathbb{P}[|\bar{\alpha}| \leq \varepsilon] = 1 - \mathbb{P}[-\varepsilon \leq \bar{\alpha} \leq \varepsilon] = 2e^{-\frac{2m\varepsilon^2}{(b-a)^2}}.$$

Finalment, desfent el canvi de variable α_i i $\bar{\alpha}$ s'obté el resultat desitjat. □

Finalment, un cop enunciats i demostrats tots els resultats matemàtics necessaris, ja es pot procedir a demostrar la proposició 25.

Demostració de la proposició 25:
Siguin ε, $\delta \in (0, 1)$ qualsevols fixats. Es vol trobar una mida $m \in \mathbb{N}$ tal que garanteixi que per qualsevol $\mathcal{D}$

$$\mathop{\mathbb{P}}_{S\sim\mathcal{D}^m}(\{\forall h \in \mathcal{H}, |L_S(h) - L_{\mathcal{D}}(h)| \leq \varepsilon\}) \geq 1 - \delta.$$

Per simplicitat, s'emprarà el canvi de notació $\mathop{\mathbb{P}}_{S\sim\mathcal{D}^m} \equiv \mathcal{D}^m$, obtenint així:

$$\mathcal{D}^m(\{S : \forall h \in \mathcal{H}, |L_S(h) - L_{\mathcal{D}}(h)| \leq \varepsilon\}) \geq 1 - \delta.$$

Provar aquesta desigualtat és equivalent a provar el seu complementari

$$\mathcal{D}^m(\{S : \exists h \in \mathcal{H}, |L_S(h) - L_{\mathcal{D}}(h)| > \varepsilon\}) < \delta.$$

Noti's que l'esdeveniment $\{S : \exists h \in \mathcal{H}, |L_S(h) - L_{\mathcal{D}}(h)| > \varepsilon\}$ es pot reescriure com

$$\{S : \exists h \in \mathcal{H}, |L_S(h) - L_{\mathcal{D}}(h)| > \varepsilon\} = \bigcup_{h\in\mathcal{H}} \{S : |L_S(h) - L_{\mathcal{D}}(h)| > \varepsilon\}.$$

Aplicant la propietat de la σ sub-additivitat

$$\begin{aligned}\mathcal{D}^m(\{S : \exists h \in \mathcal{H}, |L_S(h) - L_{\mathcal{D}}(h)| > \varepsilon\}) &= \mathcal{D}^m\left(\bigcup_{h\in\mathcal{H}} \{S : |L_S(h) - L_{\mathcal{D}}(h)| > \varepsilon\}\right) \\ &\leq \sum_{h\in\mathcal{H}} \mathcal{D}^m(\{S : |L_S(h) - L_{\mathcal{D}}(h)| > \varepsilon\}).\end{aligned}$$

Tot seguit, el següent objectiu és provar que cada sumant del sumatori, donada una m suficientment gran, és suficientment petit; és a dir, per cada predictor $h \in \mathcal{H}$ (el qual

convé tenir present en tot moment que és seleccionat abans de tenir accés a la mostra d'entrenament), la diferència entre l'error real i el risc empíric és petita.

Recordant les definicions de $L_{\mathcal{D}}(h)$ i $L_S(h)$

$$L_{\mathcal{D}}(h) = \mathbb{E}_{z\sim\mathcal{D}}[\ell(h,z)] \quad , \quad L_S(h) = \frac{1}{m}\sum_{i=1}^{m}\ell(h,z_i).$$

Com que les variables z_i són generades i.i.d. d'acord a $\mathcal{D}$, òbviament

$$\mathbb{E}[\ell(h,z_i)] = L_{\mathcal{D}}(h).$$

Per altra banda, per la linearitat de l'esperança

$$\mathbb{E}[L_S(h)] = \mathbb{E}\left[\frac{1}{m}\sum_{i=1}^{m}\ell(h,z_i)\right] = \sum_{i=1}^{m}\frac{1}{m}\mathbb{E}[\ell(h,z_i)] = m\frac{1}{m}L_{\mathcal{D}}(h) = L_{\mathcal{D}}(h).$$

Així doncs, $|L_S(h) - L_{\mathcal{D}}(h)| = |L_S(h) - \mathbb{E}[L_S(h)]|$ i per tant, n'hi haurà prou amb acotar per cada $h \in \mathcal{H}$

$$\mathcal{D}^m(\{S : |L_S(h) - \mathbb{E}[L_S(h)]| > \varepsilon\}).$$

Així doncs, cal provar que, per a qualsevol $h \in \mathcal{H}$, la mesura de $L_S(h)$ està concentrada al voltant de la seva esperança. Per fer-ho s'emprerà la desigualtat de Hoeffding.

Sigui, per a qualsevol $i \in \{1, ..., m\}$, θ_i la variable aleatòria $\ell\ (h,\ z_i)$. Com que h és fixa i les dades z_1, ..., z_m són extretes de manera i.i.d., aleshores com a conseqüència θ_1, ..., θ_m són també variables aleatòries i.i.d.

Per altra banda, se segueix de la definició de $L_S(h)$ que $L_S(h) = \frac{1}{m}\sum_{i=1}^{m}\ell(h,z_i)$; a més a més, existeix $\mu \in \mathbb{R}$ tal que $L_{\mathcal{D}}(h) = \mu$. Assumeixi's també que el rang de ℓ és $[0, 1]$ i, per tant, $\theta_i \in [0, 1]$ per a qualsevol i.

Aleshores s'obté

$$\mathcal{D}^m(\{S : |L_S(h) - L_{\mathcal{D}}(h)| > \varepsilon\}) = \mathbb{P}\left[\left|\frac{1}{m}\sum_{i=1}^{m}\theta_i - \mu\right| > \varepsilon\right] \leq 2e^{(-2m\varepsilon^2)}.$$

On combinant-ho amb

$$\mathcal{D}^m(\{S : \exists h \in \mathcal{H}, |L_S(h) - L_{\mathcal{D}}(h)| > \varepsilon\}) \leq \sum_{h\in\mathcal{H}}\mathcal{D}^m(\{S : |L_S(h) - L_{\mathcal{D}}(h)| > \varepsilon\}).$$

S'obté

$$\begin{aligned}\mathcal{D}^m(\{S : \exists h \in \mathcal{H}, |L_S(h) - L_{\mathcal{D}}(h)| > \varepsilon\}) &\leq \sum_{h\in\mathcal{H}} 2e^{(-2m\varepsilon^2)} \\ &= 2|\mathcal{H}|e^{(-2m\varepsilon^2)}.\end{aligned}$$

Prenent $m \geq \frac{\log(2|\mathcal{H}|/\delta)}{2\varepsilon^2}$ s'obté finalment

$$\mathcal{D}^m(\{S : \exists h \in \mathcal{H}, |L_S(h) - L_{\mathcal{D}}(h)| > \varepsilon\}) \leq \delta.$$

Tal i com es volia provar □

D'aquesta demostració se'n pot extreure el següent resultat:

Corol·lari 31. *Sigui $\mathcal{H}$ una classe d'hipòtesis finita, sigui $\mathcal{Z}$ un domini, sigui ℓ: $\mathcal{H}$ $\times$ $Z \to$ [0, 1] una funció de pèrdua. Aleshores, $\mathcal{H}$ posseeix la propietat de convergència uniforme amb complexitat de mostra*

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

De fet, la classe $\mathcal{H}$ és aprensible PAC agnòsticament fent servir l'algoritme ERM amb complexitat de la mostra

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

La demostració d'aquest corol·lari està inclosa en les demostracions de la proposició 25 i del corol·lari 24.

3.4 El teorema No-Free-Lunch

Retornant per un moment al capítol 1, allí es va veure que si no s'era prudent a l'hora d'aplicar l'algoritme ERM es podia obtenir un predictor que no funcionés correctament d'acord amb la realitat. A aquest fenomen se'l va anomenar sobreajustament i, per tal de solventar-lo, es va posar de manifest la necessitat d'introduir algun tipus de coneixment previ; així doncs, des d'ençà s'ha treballat sobre una classe d'hipòtesis $\mathcal{H}$, la qual reflecteix aquest coneixement previ a mode d'una restricció sobre l'espai de cerca de la funció de predicció.

No obstant, en cap moment s'ha qüestionat si realment és necessari aquest coneixement previ per tal que sigui possible aprendre. Fins ara tan sols s'ha assenyalat la incapacitat d'aprendre de l'algoritme ERM de forma fiable i satisfactòria en absència de coneixement previ, però no s'ha plantejat l'existència d'un altre algoritme que si pugui oferir garanties de fiabilitat i precisió a l'hora d'aprendre un problema qualsevol en absència de coneixement previ. És a dir, no s'ha plantejat la possibilitat de l'existència d'un aprenent universal. Així doncs, en aquesta secció s'abordarà justament aquesta qüestió.

Formalitzant el problema; consideri's una distribució desconeguda $\mathcal{D}$ sobre una parella de conjunt de domini i conjunt de classificacions $\mathcal{X} \times \mathcal{Y}$. La qüestió és la següent: existeix un algoritme d'aprenentatge A tal que, donats ε i $\delta \in (0, 1)$, donada una mida de la mostra m (els exemples de la qual són i.i.d.), amb probabilitat igual o major a $1 - \delta$, retorni una funció de predicció h: $\mathcal{X} \to \mathcal{Y}$ tal que el risc $L_{\mathcal{D}}(h)$ sigui menor que ϵ; és a dir, amb alta probabilitat tingui un eror baix.

La resposta a la pregunta anterior, malauradment, és negativa; no existeix cap aprenent universal. Es pot demostrar mitjançant el teorema "No-Free-Lunch"(NFL) que per a problemes de predicció en classificació binària, per a cada aprenent, hi ha una distribució concreta en la qual fracasa (i, com no se sap en cap moment res sobre la distribució $\mathcal{D}$ latent, no es pot afirmar mai l'existència de tal aprenent). A continuació, s'enunciarà i demostrarà aquest resultat.

Teorema 32. (teorema "No-Free-Lunch"o NFL) *Sigui A un algoritme d'aprenentatge per a la tasca de classificació binària respecte la funció de pèrdua $0 - 1$ sobre un domini $\mathcal{X}$. Sigui m un nombre qualsevol menor que $|\mathcal{X}|/2$, el qual representa la mida del conjunt d'entrenament. Aleshores, existeix una distribució $\mathcal{D}$ sobre $\mathcal{X} \times \mathcal{Y}$ tal que*

1. *Existeix una funció* $f\colon \mathcal{X} \to \{0, 1\}$ *amb* $L_{\mathcal{D}}(f) = 0$.
2. *Amb probabilitat major o igual a 1/7 sobre l'elecció de* $S \sim \mathcal{D}^m$ *es té* $L_{\mathcal{D}}(A(S)) \geq 1/8$.

Per tal de provar el teorema NFL, cal el següent resultat:

Lema 33. *Sigui Z una variable aleatòria que pren valors a l'interval [0, 1]. Assumeixi's que* $\mathbb{E}[Z] = \mu$. *Aleshores, per a qualsevol* $a \in (0, 1)$,

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

La qual cosa implica alhora que, per a qualsevol $a \in (0, 1)$,

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

Demostració:
Sigui $Y = 1 - Z$. Aleshores Y és una variable aleatòria no negativa amb $\mathbb{E}[Y] = 1 - \mathbb{E}[Z] = 1 - \mu$. Aplicant la desigualtat de Markov (Lema 29) sobre Y s'obté

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

Així doncs

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

Com es volia demostrar □

Ara sí, ja es pot procedir a demostrar el teorema NFL.

Demostració del teorema 32.
Sigui $C \subset \mathcal{X}$ un subconjunt de $\mathcal{X}$ de mida $2m$. Com que cada element de C pot estar classificat com a 0 o com a 1 independentment de com estiguin classificats els altres, hi ha un total de $T = 2 \cdot 2 \cdot \overset{2m}{\cdots} \cdot 2 = 2^{2m}$ combinacions de classificacions possibles i, per tant, de funcions de C a $\{0, 1\}$ diferents. Des d'ara fins que conclogui la demostració es denotaran aquestes funcions com $f_1, ..., f_T$.

Per cadascuna d'aquestes funcions f_i, $i \in \{1, ..., 2^{2m}\}$, es defineix la distribució $\mathcal{D}_i$ sobre $C \times \{0, 1\}$ com

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

La qual indica que la probabilitat de cada parella de punts (x, y) és $1/|C|$ si, d'acord a la funció f_i, el punt x està correctament classificat com a y i 0 si la classificació no és correcta. És a dir, per cada funció f_i, $\mathcal{D}_i$ és la distribució uniforme sobre totes les parelles de punts del domini i la seva imatge (classificació correcta) i distribució 0 sobre tots els altres. Clarament, $L_{\mathcal{D}_i}(f_i) = 0$.

L'objectiu consisteix en demostrar que, donat qualsevol algoritme A tal que, donada una mostra qualsevol d'elements de $C \times \{0, 1\}$ de mida m, retorni una funció $A(S)\colon C \to \{0, 1\}$ es complirà

$$\max_{i \in [T]} \mathbb{E}_{S \sim \mathcal{D}_i^m}[L_{\mathcal{D}_i}(A(S))] \geq 1/4.$$

On la notació $[i]$ fa referència al conjunt de nombres naturals menors o iguals a T.

Per fer-ho, en primer lloc, com C té $2m$ elements diferents i S és una seqüència de m elements qualssevol (repetits o no) de C, hi ha un total de $k = 2m \cdot ... \cdot 2m = (2m)^m$ possibles seqüències S de m exemples de C. Des d'ara fins que acabi aquesta demostració es denotaran aquestes seqüències com S_1, ..., S_k. A més a més, es denotarà com S_j^i la seqüència consistent en les instàncies $(x_1, ..., x_m)$ de la seqüència S_j etiquetades segons la funció f_i; és a dir, $S_j^i = ((x_1, f_i(x_1)), ..., (x_m, f_i(x_m)))$.

Sigui $\mathcal{D}_i$ la distribució, aleshores les mostres que pot rebre A són S_1^i, ..., S_k^i, les quals tenen totes la mateixa probabilitat de ser seleccionades. Així doncs, es té

$$\mathop{\mathbb{E}}_{S\sim\mathcal{D}_i^m}[L_{\mathcal{D}_i}(A(S))] = \frac{1}{k}\sum_{j=1}^{k} L_{\mathcal{D}_i}(A(S_j^i)). \tag{3.4.1}$$

Trivialment, com que el màxim sempre és major o igual que el promig i el mínim sempre és igual o menor al promig es té

$$\begin{aligned}\max_{i\in[T]}\frac{1}{k}\sum_{j=1}^{k} L_{\mathcal{D}_i}(A(S_j^i)) &\geq \frac{1}{T}\sum_{i=1}^{T}\frac{1}{k}\sum_{j=1}^{k} L_{\mathcal{D}_i}(A(S_j^i)) = \\ = \frac{1}{k}\sum_{j=1}^{k}\frac{1}{T}\sum_{i=1}^{T} L_{\mathcal{D}_i}(A(S_j^i)) &\geq \min_{j\in[k]}\frac{1}{T}\sum_{i=1}^{T} L_{\mathcal{D}_i}(A(S_j^i)).\end{aligned} \tag{3.4.2}$$

Fixant ara una $j \in [k]$, es denota $S_j = (x_1, ..., x_m)$ i es prenen v_1, ..., v_p els elements de C que no apareixen a S_j, on $p \in \mathbb{N}$. Clarament $p \geq m$, ja que $|C| = 2m$, $|S_j| = m$ i S_j pot contenir elements repetits. Així doncs, per qualsevol funció h: $C \to \{0, 1\}$ i per a qualsevol $i \in [T]$ es tindrà

$$L_{\mathcal{D}_i}(h) = \frac{1}{2m}\sum_{x\in C}\mathbb{1}_{[h(x)\neq f_i(x)]} \geq \frac{1}{2m}\sum_{r=1}^{p}\mathbb{1}_{[h(v_r)\neq f_i(v_r)]} \geq \frac{1}{2p}\sum_{r=1}^{p}\mathbb{1}_{[h(v_r)\neq f_i(v_r)]}.$$

On la primera desigualtat es deu que $|C| = 2m > p$ i la segona es deu que $p > m$. Vistes aquestes desigualtats, aleshores

$$\begin{aligned}\frac{1}{T}\sum_{i=1}^{T} L_{\mathcal{D}_i}(A(S_j^i)) &\geq \frac{1}{T}\sum_{i=1}^{T}\frac{1}{2p}\sum_{r=1}^{p}\mathbb{1}_{[A(S_j^i)(v_r)\neq f_i(v_r)]} \\ &= \frac{1}{2p}\sum_{r=1}^{p}\frac{1}{T}\sum_{i=1}^{T}\mathbb{1}_{[A(S_j^i)(v_r)\neq f_i(v_r)]} \\ &\geq \frac{1}{2}\cdot\min_{r\in[p]}\frac{1}{T}\sum_{i=1}^{T}\mathbb{1}_{[A(S_j^i)(v_r)\neq f_i(v_r)]}.\end{aligned} \tag{3.4.3}$$

Fixant ara un $r \in [p]$, es poden separar les funcions f_1, ..., f_T en $T/2$ parelles disjuntes tals que, per cada parella $(f_i, f_{i'})$, per a qualsevol $c \in C$, $f_i \neq f_{i'}$ si i només si $c = v_r$. És a dir, un cop fixat $r \in [p]$, s'aparella cada funció f_i amb la funció $f_{i'}$ $(i, i' \in [T])$ tal que coincideixen en la imatge de tots els elements de C excepte en v_r, on difereixen. Per tant, com que $v_r \notin S_j$, $S_j \subset C$ i $f_i(c) = f_{i'}(c)$ per a qualsevol $c \in C$ excepte $c = v_r$, es té $S_j^i =$

$S_j^{i'}$. Conseqüentment

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

la qual cosa implica

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

Per tant, combinant-ho amb (3.4.3), (3.4.2) i (3.4.1) s'obté

$$\begin{aligned}\max_{i\in T} \mathop{\mathbb{E}}_{S\sim\mathcal{D}_i^m}[L_{\mathcal{D}_i}(A(S))] &= \max_{i\in[T]}\frac{1}{k}\sum_{j=1}^{k} L_{\mathcal{D}_i}(A(S_j^i)) \geq \min_{j\in[k]}\frac{1}{T}\sum_{i=1}^{T} L_{\mathcal{D}_i}(A(S_j^i)) \\ &\geq \frac{1}{2}\cdot\min_{r\in[p]}\frac{1}{T}\sum_{i=1}^{T}\mathbb{1}_{[A(S_j^i)(v_r)\neq f_i(v_r)]} = \frac{1}{2}\times\frac{1}{2} = \frac{1}{4}.\end{aligned}$$

Que és el que es volia provar a (3.4). Aquesta condició significa que, per qualevol algoritme A' que rebi un conjunt d'entrenament de mida m de $\mathcal{X}\times\{0,1\}$, existeix una funció $f\colon \mathcal{X}\to\{0,1\}$ i una distribució $\mathcal{D}$ sobre $\mathcal{X}\times\{0,1\}$ tal que $L_{\mathcal{D}}(f)=0$ i

$$\mathop{\mathbb{E}}_{S\sim\mathcal{D}^m}[L_{\mathcal{D}}(A(S))].$$

Finalment, només resta provar que aquesta condició és suficient per tal que $\mathbb{P}[L_{\mathcal{D}}(A'(S)) \geq 1/8] \geq 1/7$. Per fer-ho, s'emprarà el lema 33.

$$\begin{aligned}\mathbb{P}[L_{\mathcal{D}}(A(S)) \geq 1/8] &= \mathbb{P}[L_{\mathcal{D}}(A(S)) \geq 1 - 7/8] \\ &\geq \frac{\mathbb{E}[L_{\mathcal{D}}(A(S))] - (1-7/8)}{7/8} \\ &\geq \frac{1/8}{7/8} = 1/7.\end{aligned}$$

Tal i com es volia demostrar. □

Capítol 4

La dimensió VC i el teorema fonamental de l'aprenentatge estadístic

Al capítol anterior s'ha plantejat la qüestió sobre quines classes d'hipòtesis són aprensibles, sobre què caracteritza l'aprensibilitat o no d'una classe d'hipòtesis donada. Si bé s'ha donat una resposta parcial a aquesta incògnita al veure que qualsevol classe finita és PAC aprensible agnòsticament, la qüestió de fons segueix sense una resposta concreta, clara i contundent. Així doncs, aquest és justament el tema que s'abordarà en aquest capítol, es determinarà quines classes són aprensibles PAC i es caracteritzarà exactament la complexitat de la mostra d'una classe donada que sigui aprensible.

En primer lloc, ja que prèviament s'ha demostrat que qualsevol classe finita és aprensible, un bon punt de partida és plantejar-se just la pregunta contrària; és a dir, existeixen classes d'hipòtesis de mida infinita que siguin aprensibles? El següent exemple demostra que sí:

Exemple: Sigui $\mathcal{H}$ la classe d'hipòtesis de les funcions semiinterval inferior sobre la recta dels nombres reals; formalment $\mathcal{H} = \{h_a : a \in \mathbb{R}\}$, on h_a: $\mathbb{R} \to \{0, 1\}$ és una funció tal que $h_a(x) = \mathbb{1}_{[x<a]}$. És a dir, funcions que valen 1 si $x < a$ i valen 0 si $x \geq a$. Claramnent, $\mathcal{H}$ té mida infinita; no obstant, es pot demostrar que és aprensible PAC. Aquest resultat queda formalitzat en el següent lema, la demostració del qual servirà per provar aquesta afirmació.

Lema 34. *Sigui $\mathcal{H}$ la classe d'equivalència de les funcions semiinterval inferior definida com $\mathcal{H} = \{h_a : a \in \mathbb{R}\}$, on h_a: $\mathbb{R} \to \{0, 1\}$ és una funció tal que $h_a(x) = \mathbb{1}_{[x<a]}$. Aleshores, $\mathcal{H}$ és aprensible PAC fent servir la regla ERM i la complexitat de la mostra és $m_{\mathcal{H}}(\varepsilon, \delta) \leq \lceil \log(2/\delta)/\varepsilon \rceil$.*

Demostració:
Sigui a^* una cota superior de semiinterval tal que la hipòtesis $h^*(x) = \mathbb{1}_{[x<a^*]}$ assoleix $L_D(h^*) = 0$. Sigui $\mathcal{D}_x$ una distribució marginal sobre el domini $\mathcal{X}$ i siguin a_0, $a_1 \in \mathbb{R}$ tals que $a_0 < a^* < a_1$ i

$$\mathop{\mathbb{P}}_{x\sim\mathcal{D}_x}[x \in (a_0, a^*)] = \mathop{\mathbb{P}}_{x\sim\mathcal{D}_x}[x \in (a^*, a_1)] = \varepsilon.$$

On si $\mathcal{D}_x(-\infty, a^*) \leq \varepsilon$ es pren $a_0 = -\infty$ i, equivalentment, si $(a^*, \infty) \leq \varepsilon$ es pren a_1

$= \infty$. Donada una mostra d'entrenamnet S, es pren $b_0 = \max\{x\colon (x,1) \in S\}$ i $b_1 = \min\{x\colon (x,0) \in S\}$; on $b_0 = -\infty$ si no existeix cap exemple a S etiquetat positivament i $b_1 = \infty$ si no existeix cap exemple a S etiquetat negativament. Sigui ara b_S una cota superior de semiinterval corresponent a una hipòtesi h_S retornada per l'ERM; per definició, $b_s \in (b_0, b_1)$. Per tant, per tal que $L_{\mathcal{D}} \leq \varepsilon$ n'hi ha prou que $b_0 \geq a_0$ i $b_1 \leq a_1$, és a dir

$$\mathop{\mathbb{P}}_{S\sim\mathcal{D}^m}[L_{\mathcal{D}^m}(h_S) > \varepsilon] \leq \mathop{\mathbb{P}}_{S\sim\mathcal{D}^m}[b_0 < a_0 \vee b_1 > a_1].$$

Fent servir la desigualtat de la unió es pot acotar l'anterior probabilitat de la forma:

$$\mathop{\mathbb{P}}_{S\sim\mathcal{D}^m}[L_{\mathcal{D}^m}(h_S) > \varepsilon] \leq \mathop{\mathbb{P}}_{S\sim\mathcal{D}^m}[b_0 < a_0] + \mathop{\mathbb{P}}_{S\sim\mathcal{D}^m}[b_1 > a_1]. \tag{4.0.1}$$

On hom pot observar que l'esdeveniment $b_0 < a_0$ succeeix si i només si tots els exemples de S estan fora de l'interval (a_0, a^*), la massa de probabilitat del qual és, per definició, ε; és a dir

$$\mathop{\mathbb{P}}_{S\sim\mathcal{D}^m}[b_0 < a_0] = \mathop{\mathbb{P}}_{S\sim\mathcal{D}^m}[\forall(x,y) \in S, x \notin (a_0, a^*)] = (1-\varepsilon)^m \leq e^{-\varepsilon m}.$$

Imposant que la mida de la mostra sigui $m > \frac{\log(2/\delta)}{\varepsilon}$, s'obté

$$\mathop{\mathbb{P}}_{S\sim\mathcal{D}^m}[b_0 < a_0] \leq \delta/2. \tag{4.0.2}$$

Anàlogament es pot demostrar que $\mathop{\mathbb{P}}_{S\sim\mathcal{D}^m}[b_1 > a_1] \leq \delta/\ 2$.

L'esdeveniment $b_1 > a_1$ succeeix si i només si tots els exemples de S estan fora de l'interval (a^*, a_1), la massa de probabilitat del qual és, per definició, ε; és a dir

$$\mathop{\mathbb{P}}_{S\sim\mathcal{D}^m}[b_1 > a_1] = \mathop{\mathbb{P}}_{S\sim\mathcal{D}^m}[\forall(x,y) \in S, x \notin (a*, a_1)] = (1-\varepsilon)^m \leq e^{-\varepsilon m}.$$

I, imposant la mateixa cota inferior a la mida de la mostra $m > \frac{\log(2/\delta)}{\varepsilon}$, s'obté

$$\mathop{\mathbb{P}}_{S\sim\mathcal{D}^m}[b_1 > a_1] \leq \delta/2. \tag{4.0.3}$$

Finalment, ajuntant (4.0.1), (4.0.2) i (4.0.3) s'obté finalment

$$\mathop{\mathbb{P}}_{S\sim\mathcal{D}^m}[L_{\mathcal{D}^m}(h_S) > \varepsilon] \leq \mathop{\mathbb{P}}_{S\sim\mathcal{D}^m}[b_0 < a_0] + \mathop{\mathbb{P}}_{S\sim\mathcal{D}^m}[b_1 > a_1] \leq \delta/2 + \delta/2 = \delta.$$

Per a qualsevol $m \geq \frac{\log(2/\delta)}{\varepsilon}$, tal i com es volia provar inicialment. Trivialment, $m_{\mathcal{H}}(\varepsilon, \delta) \leq \frac{\log(2/\delta)}{\varepsilon}$, per la seva definició de mida mínima tal que es compleix la condició d'aprenentatge. □

Així doncs, ha sigut evidenciat que, malgrat la finitut d'una classe d'hipòtesis $\mathcal{H}$ és condició suficient per tal que sigui aprensible PAC agnòsticament, no és condició necessària. Per tant, l'objectiu és determinar quina o quines condicions són necessàries per garantir l'aprensibilitat d'una classe d'hipòtesis $\mathcal{H}$.

4.1 Dimensió VC

Per tal de començar a abordar la situació, convé aturar-se a revisar el teorema NFL, així com la seva demostració. Aquest teorema garanteix, per qualsevol algoritme d'aprenentatge i una mida de la mostra inferior a la meitat de la mida del domini, l'existència d'una distribució de probabilitat en la qual l'algoritme fracassarà (malgrat un altre algoritme

pogués resultar-hi exitós) si no es restringeix la classe d'hipòtesis $\mathcal{H}$. Tal i com es pot veure a la demostració, per tal de trobar aquesta distribució l'adversari empra un subconjunt $C \subset \mathcal{X}$ finit i considera una familia de funcions concentrades en elements de C. Per tal de garantir que per qualsevol algoritme pot trobar una distribució on fracassi, l'adversari aprofita la seva capacitat d'escollir una funció del conjunt de totes les funcions que van de C a $\{0, 1\}$. No obstant, en tot moment l'adversari es veu restringit a construir distribucions on alguna hipòtesi $h \in \mathcal{H}$ tingués risc empíric nul. Així doncs, com que s'estan considerant distribucions concentrades en elements de C, un bon punt de partida és estudiar com es comporta $\mathcal{H}$ a C.

Definició 35. *Sigui $\mathcal{H}$ una classe de funcions de $\mathcal{X}$ a $\{0, 1\}$, sigui $C = \{ c_1, ..., c_m \} \subset \mathcal{X}$. S'anomena restricció de $\mathcal{H}$ a C al conjunt de funcions de C a $\{0, 1\}$ tals que poden ser derivades de $\mathcal{H}$. És a dir:*

$$\mathcal{H}_C = \{(h(c_1), ..., h(c_m)) : h \in \mathcal{H}\}.$$

On cada funció de C a $\{0, 1\}$ es representa com un vector a $\{0, 1\}^{|C|}$.

Quan la restricció de $\mathcal{H}$ a C és el conjunt de totes les funcions de C a $\{0, 1\}$, aleshores es diu que $\mathcal{H}$ separa ("shatters") C.

Formalment:

Definició 36. *Una classe d'hipòtesis $\mathcal{H}$ separa un subconjunt finit $C \subset \mathcal{X}$ si la restricció de $\mathcal{H}$ a C és el conjunt de totes les funcions de C a $\{0, 1\}$; és a dir, $|\mathcal{H}_C| = 2^{|C|}$.*

Per tal de facilitar la compressió d'aquesta definció per part del lector, el següet exemple pot resultar força ilustratiu.

Exemple: Sigui $\mathcal{H}$ la classe de les funcions semiinterval obert acotat superiorment (anteriorment anomenades semiinterval inferior) sobre $\mathbb{R}$. Sigui $C = \{c_1\}$, on $c_1 \in \mathbb{R}$, un subconjunt de $\mathbb{R}$. Prenent $a = c_1 + 1$ es té trivialment $h_a(c_1) = 1$, mentre que prenent $a = c_1 - 1$ es té $h_a(c_1) = 0$. Així doncs, com que $\mathcal{H}_C$ és el conjunt de totes les funcions de C a $\{0, 1\}$, $\mathcal{H}$ separa C. No obstant, si es pren $C = \{c_1, c_2\}$, amb $c_1 \leq c_2$, prenent $a = (c_1 + c_1)/2$ es té $h_a(C) = (h_a(c_1), h_a(c_2)) = (1, 0)$, prenent $a = c_1 - 1$ es té $h_a(C) = (h_a(c_1), h_a(c_2)) = (0, 0)$, prenent $a = c_1 + 1$ es té $h_a(C) = (h_a(c_1), h_a(c_2)) = (1, 1)$; no obstant, l'element $(0, 1)$ és impossible d'obtenir, ja que (com $c_1 \leq c_2$), $\forall\ a \in \mathbb{R}$, si $c_2 \leq a$, aleshores $c_1 \leq$ i, per tant, $h_a(c_2) = 1 \Rightarrow h_a(c_1) = 1$. Així doncs, no totes les funcions de C a $\{0, 1\}$ estan incloses a $\mathcal{H}$ (ja que falta la funció que envia c_1 a 0 i c_2 a 1) i per tant no separa $C = \{c_1, c_2\}$.

De la definició 36, tornant momentàniament a la demostració del teorema NFL, es pot observar que, a l'hora de construir la distribució adversarial, si la classe d'hipòtesis $\mathcal{H}$ separa C, aleshores la classe d'hipòtesis no suposa cap restricció real per l'adversari, ja que aquest segueix podent escollir qualsevol funció de C a $\{0, 1\}$ en la que basar-se a l'hora de construir la distribució $\mathcal{D}$ (conservant en tot moment l'assumpció de realitzabilitat).

Aquesta situació es pot formalitzar en el següent corol·lari.

Corol·lari 37. *Sigui $\mathcal{X}$ un domini, sigui $\mathcal{H}$ una classe d'hipòtesis formada per funcions de $\mathcal{X}$ a $\{0, 1\}$, sigui S una mostra de $m \in \mathbb{N}$ elements de $\mathcal{X} \times \{0, 1\}$ generats i.i.d. Suposi's l'existència d'un subconjunt $C \subset \mathcal{X}$ de mida $2m$ que és separa per $\mathcal{H}$. Aleshores, per a qualsevol algoritme d'aprenentatge A, existeix una distribució $\mathcal{D}$ sobre $\mathcal{X} \times \{0, 1\}$ i un predictor $h \in \mathcal{H}$ tals que $L_{\mathcal{D}}(h) = 0$, però amb probabilitat major o igual a 1/7 sobre l'elecció de $S \sim \mathcal{D}^m$ es té $L_{\mathcal{D}}(A(S)) \geq 1/8$.*

Demostració:

Per l'argument proporcionat previ a l'anunci del corol·lari, l'adversari pot escollir qualsevol funció de C a $\{0, 1\}$ per basar-s'hi a l'hora de construir la distribució $\mathcal{D}$; així doncs, no hi ha cap restricció efectiva sobre $\mathcal{H}$. Per tant, pel teorema NFL es compleix el corol·lari (mateixes presmises, mateix resultat). □

Així doncs, aquest corol·lari diu que, donada una classe d'hipòtesis $\mathcal{H}$, si aquesta separa un conjunt C de mida $2m$ aleshores $\mathcal{H}$ no es pot aprendre amb m exemples.

Un cop introduïdes totes aquestes nocions teòriques prèvies, ja es pot presentar el que és el concepte al voltant del qual girarà tot aquest capítol; la dimensió VC.

Definició 38. *Sigui $\mathcal{X}$ un domini, sigui $\mathcal{H}$ una classe d'hipòtesis. La dimensió VC de la classe d'hipòtesis $\mathcal{H}$ es denota VCdim($\mathcal{H}$) i és la mida màxima d'un subconjunt $C \subset \mathcal{X}$ que pugui ser separa per $\mathcal{H}$. Si $\mathcal{H}$ pot separar conjunts de mida indefinida, es diu que $\mathcal{H}$ té dimensió VC infinita.*

A partir del concepte de dimensió VC i del corol·lari 37 s'obté el següent resultat:

Teorema 39. *Sigui $\mathcal{H}$ una classe d'hipòtesis amb dimensió VC infinita. Aleshores $\mathcal{H}$ no és aprensible PAC.*

Demostració:
Com que $\mathcal{H}$ té dimensió VC infinita, per qualsevol mida m de mostra d'entrenament existeix un conjunt C de mida $2m$ que és separat per $\mathcal{H}$. Pel corol·lari 37 aleshores $\mathcal{H}$ no és aprensible PAC per mostres de mida m; com que aquest argument és cert per valors arbitràriamnet grans de m, $\mathcal{H}$ no és aprensible PAC. □

De la definció es pot extreure que, per tal de demostrar que la dimensió VC d'una classe d'hipòtesis $\mathcal{H}$ és un valor concret (VCdim($\mathcal{H}$) = d, on $d \in \mathbb{Z}$) cal demostrar que existeix un conjunt C de mida d tal que és separat per $\mathcal{H}$ i que qualsevol conjunt C de mida $d + 1$ no és separat per $\mathcal{H}$. Per exemple, anteriorment s'ha provat que la dimensió VC de la classe $\mathcal{H}$ de les funcions semiinterval obert acotat superiorment és 1, ja que s'ha vist que separa qualsevol conjunt de mida 1 però no és capaç de separar-ne cap de mida 2.

El següent exemple pot evidenciar més clarament aquesta noció:

Exemple: Suposi's que es vol calcular la dimensió VC de la classe d'hipòtesis $\mathcal{H}$ dels rectangles alineats amb els eixos cartesians; formalment:

$$\mathcal{H} = \{h_{(a_1,a_2,b_1,b_2)} : a_1 \leq a_2 \text{ i } b_1 \leq b_2\}.$$

on

$$h_{(a_1,a_2,b_1,b_2)}(x_1, x_2) = \begin{cases} 1 & \text{si } \ a_1 \leq x_1 \leq a_2 \text{ i } b_1 \leq x_2 \leq b_2 \ , \\ 0 & \text{altrament .} \end{cases}$$

La dimensió VC de $\mathcal{H}$ és 4; per provar-ho en primer lloc es proporciona un conjunt de mida 4 tal que pugui ser separat. El conjunt C = { c_1 = (-2, 0), c_2 = (0, 2), c_3 = (2, 0), c_4 = (0, -2) } és fàcilment separable[1].

Ara queda provar que no existeix cap conjunt C de 5 element que pugui ser separat

[1]Mitjançant els rectangles $h_{(-1,1,-1,1)}$, $h_{(-3,-1,-1,1)}$, $h_{(-1,1,1,3)}$, $h_{(3,1,-1,1)}$, $h_{(-1,1,-3,-1)}$, $h_{(-3,1,-1,3)}$, $h_{(-3,3,-1,1)}$, $h_{(-3,1,-3,1)}$, $h_{(-1,3,-1,3)}$, $h_{(-1,1,-3,3)}$, $h_{(-1,3,-3,1)}$, $h_{(-1,3,-3,3)}$, $h_{(-3,3,-3,1)}$, $h_{(-3,1,-3,3)}$, $h_{(-3,3,-1,3)}$ i $h_{(-3,3,-3,3)}$; on cadascun equival a l'escenari $\{(h(c_1), h(c_2), h(c_3), h(c_4))\}$ = {(0, 0, 0, 0)}, {(1, 0, 0, 0)}, {(0, 1, 0, 0)}, {(0, 0, 1, 0)}, {(0, 0, 0, 1)}, {(1, 1, 0, 0)}, {(1, 0, 1, 0)}, {(1, 0, 0, 1)}, {(0, 1, 1, 0)}, {(0, 1, 0, 1)}, {(0, 0, 1, 1)}, {(0, 1, 1, 1)}, {(1, 0, 1, 1)}, {(1, 1, 0, 1)}, {(1, 1, 1, 0)}, {(1, 1, 1, 1)} respectivament.

per $\mathcal{H}$. Sigui $C = \{c_1, c_2, c_3, c_4, c_5\}$; s'anomenarà α a la mínima coordenada x d'entre els 5 punts, β a la màxima coordenada x, λ a la mínima coordenada y i γ a la màxima coordenada y. Sigui $p = (p_1, p_2)$ el punt (o un dels punts) de C tal que $p_1 \neq \alpha$, β i $p_2 \neq \lambda$, γ. Aleshores, és impossible construir un rectangle que contingui tots els punts menys p, ja que aquest rectangle hauria de contenir el rectangle $h_{(\alpha,\beta,\lambda,\gamma)}$ (rectangle mínim que conté els altres 4 punts), però per definició de p i dels paràmetres del rectangle p hi pertany. Així doncs, $\mathcal{H}$ no separa C i, per tant, $\text{VCdim}(\mathcal{H}) = 4$ tal i com es volia provar. □

4.2 El teorema fonamental de l'aprenentatge estadístic

Finalment, arriba el resultat més important de la memòria, el teorema que caracteritza definitivament quines classes d'hipòtesis són aprensibles per al cas concret de problema de classificació binària i funció de pèrdua $0-1$. En essència, aquest teorema diu que qualsevol classe amb dimensió VC finita té la propietat de convergència uniforme i, per tant, és aprensible PAC i aprensible PAC agnòsticament. A més a més, es presentarà també una segona versió d'aquest teorema, l'anomenada versió quantitativa, la qual determina en funció de dues constants, de la dimensió VC i dels paràmetres de confiança i precisió la compexitat de la mostra per tal que una classe d'hipòtesis donada tingui la propietat de convergència uniforme i sigui aprensible PAC i aprensible PAC agnòsticament.

No obstant, abans d'enunciar i demostrar els teoremes mencionats, cal introduïr primer el concepte de funció de creixement ("growth function"), ja que serà necessari a l'hora de demostrar una de les dues versions.

Definició 40. *Sigui $\mathcal{X}$ un domini, sigui $\mathcal{Y} = \{0, 1\}$ un conjunt de classificació, sigui $\mathcal{H}$ una classe d'hipòtesis formada per funcions $h{:}\mathcal{X} \to \{0, 1\}$. Aleshores la funció de creixement de $\mathcal{H}$, denotada com $\tau_{\mathcal{H}}{:}\mathbb{N} \to \mathbb{N}$ es defineix com:*

$$\tau_{\mathcal{H}}(m) = \max_{C \subset \mathcal{X}:|C|=m} |\mathcal{H}_C|.$$

És a dir, $\tau_{\mathcal{H}}$ és el nombre de funcions diferents des d'un conjunt C de mida m a $\{0, 1\}$ que es poden obtenir restringint $\mathcal{H}$ a C. Intuïtivament, la funció de creixement mesura la mida efectiva màxima de $\mathcal{H}$ en un conjunt de m exemples.

Ara si, començant per la versió bàsica del teorema, aquesta diu el següent:

Teorema 41. (teorema fonamental de l'aprenentatge estadístic) *Sigui $\mathcal{H}$ una classe d'hipòtesis de funcions d'un domini $\mathcal{X}$ a un conjunt de classificacions $\mathcal{Y} = \{0, 1\}$, prenent $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$, sigui ℓ: $\mathcal{H} \times \mathcal{Z} \to \{0, 1\}$ la funció de pèrdua $0-1$. Aleshores, els següents enunciats són equivalents:*

1· $\mathcal{H}$ té la propietat de convergència uniforme.

2· Qualsevol algoritme ERM és un aprenent PAC agnòstic exitós per a $\mathcal{H}$.

3· $\mathcal{H}$ és aprensible PAC agnòsticament.

4· $\mathcal{H}$ és aprensible PAC.

5· Qualsevol algoritme ERM és un aprenent PAC exitós per a $\mathcal{H}$.

6· $\mathcal{H}$ té dimensió VC finita.

Per tal de demostrar el teorema fonamental de l'aprenentatge estadístic calen tot un seguit de resultats matemàtics, els quals s'enunciaran i demostraran a continuació

prèviament a demostrar el teorema 41.

Lema 42. *Sigui $a > 0$. Aleshores $x \geq 2a\log(a)$ implica $x \geq a\log(x)$.*

Demostració:
En primer lloc, noti's que per a qualsevol $a \in (0, \sqrt{e}\,]$ la desigualtat $x \geq a\log(x)$ es compleix incondicionalment i, per tant, la implicació és trivial; per tant, d'ara en endavant se suposarà en tot moment que $a > \sqrt{e}$. Consideri's ara la funció $f(x) = x - a\log(x)$; és fàcil comprovar que la seva derivada és $f'(x) = 1 - a/x$. Per a qualsevol $x > a$ es té $f'(x) > 0$ i, per tant, la funció creix. A més a més

$$\begin{aligned} f(2a\log(a)) &= 2a\log(a) - a\log(2a\log(a)) \\ &= 2a\log(a) - a\log(a) - a\log(2a\log(a)) \\ &= a\log(a) - a\log(2a\log(a)). \end{aligned}$$

I, com $a - 2\log(a) > 0$, queda provat l'enunciat. □

Lema 43. *Sigui $a \geq 1$ i $b > 0$. Aleshores $x \geq 4a\log(2a) + 2b$ implica $x \geq a\log(x) + b$.*

Demostració:
N'hi ha prou amb provar que $x \geq 4a\log(2a) + 2b$ implica que $x \geq 2a\log(x)$ i $x \geq 2b$; ja que si $b \geq a\log(x) + b$, aleshores $x \geq 2b \geq a\log(x) + b$, i en canvi si $a\log(x) + b \geq b$, aleshores $x \geq 2a\log(x) \geq a\log(x) + b$.

Com que $a \geq 1$, aleshores $4a\log(2a) \geq 0$ i, per tant, $x \geq 4a\log(2a) + 2b \geq 2b$. A més a més, com $b > 0$, es té anàlogament $x \geq 4a\log(2a)$; per tant, fent us del lema 42, s'obté que $x \geq 2a\log(x)$, concloent així la demostració del lema 43. □

Lema 44. *Sigui X una variable aleatòria, sigui $x' \in \mathbb{R}$ un escalar, assumeixi's l'existència de a, $b \in \mathbb{R}$ tals que $a > 0$, $b > e$ i que per a qualsevol $t \geq 0$ es té $\mathbb{P}[|X - x'| > t] \leq 2be^{-t^2/a^2}$. Aleshores:*

$$\mathbb{E}[|X - x'|] \leq a(2 + \sqrt{\log(b)}).$$

Demostració:
Per a qualsevol $i = 0, 1, 2...$ es denota $t_i = a(i + \sqrt{\log(b)})$. Com que t_i és monòtonament creixent, es té

$$\mathbb{E}[|X - x'|] \leq a\sqrt{\log(b)} + \sum_{i=1}^{\infty} t_i \mathbb{P}[|X - x'| > t_{i-1}]. \tag{4.2.1}$$

Fent servir la hipòtesi del propi lema i desenvolupant s'obté

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

Combinant-ho amb (4.2.1) s'obté el resultat desitjat. □

Teorema 45. *Sigui $\mathcal{Z}$ un domini, sigui $\mathcal{H}$ una classe d'hipòtesis i sigui $\tau_{\mathcal{H}}$ la seva funció de creixement. Aleshores, per a qualsevol $\mathcal{D}$ sobre $\mathcal{Z}$ i per a qualsevol $\delta \in (0, 1)$ es*

té

$$\mathop{\mathbb{P}}_{S\sim\mathcal{D}^m}\left(|L_{\mathcal{D}}(h)-L_S(h)|\right)\leq\frac{4+\sqrt{\log(\tau_{\mathcal{H}}(2m))}}{\delta\sqrt{2m}}.$$

Demostració:
En primer lloc, es provarà

$$\mathop{\mathbb{E}}_{S\sim\mathcal{D}^m}\left[\sup_{h\in\mathcal{H}}|L_{\mathcal{D}}(h)-L_S(h)|\right]\leq\frac{4+\sqrt{\log(\tau_{\mathcal{H}}(2m))}}{\sqrt{2m}}. \tag{4.2.2}$$

Com que la variable aleatòria $\sup_{h\in\mathcal{H}}|L_{\mathcal{D}}(h)-L_S(h)|$ és no negativa, aleshores fent ús de la desigualtat de Markov (Lema 27), si es prova la desigualtat (4.2.2) aleshores quedarà provat el teorema 45.

Proseguint amb la prova de (4.2.2), sigui $S'=\{z'_1,...,z'_m\}$ una altra mostra i.i.d. de m elements de $\mathcal{Z}$, per un argument anàleg al proporcionat a la demostració de la proposició 25, s'observa que per a qualsevol $h\in\mathcal{H}$ es pot reescriure $L_{\mathcal{D}}(h)=\mathbb{E}_{S'\sim\mathcal{D}^m}[L_{S'}(h)]$. Per tant

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

Com que la funció valor absolut és convexa, aplicant la desigualtat de Jensen s'obté

$$\left|\mathop{\mathbb{E}}_{S\sim\mathcal{D}^m}\left[L_{\mathcal{D}}(h)-L_S(h)\right]\right|\leq\mathop{\mathbb{E}}_{S\sim\mathcal{D}^m}\left|L_{\mathcal{D}}(h)-L_S(h)\right|.$$

A més a més, com el suprem de les esperances és menor o igual que l'esperança del suprem es té

$$\sup_{h\in\mathcal{H}}\mathop{\mathbb{E}}_{S\sim\mathcal{D}^m}|L_{\mathcal{D}}(h)-L_S(h)|\leq\mathop{\mathbb{E}}_{S\sim\mathcal{D}^m}\sup_{h\in\mathcal{H}}|L_{\mathcal{D}}(h)-L_S(h)|.$$

Ajuntant-ho tot s'obté

$$\begin{aligned}\mathop{\mathbb{E}}_{S\sim\mathcal{D}^m}\left[\sup_{h\in\mathcal{H}}|L_{\mathcal{D}}(h)-L_S(h)|\right]&\leq\mathop{\mathbb{E}}_{S,S'\sim\mathcal{D}^m}\left[\sup_{h\in\mathcal{H}}|L_{S'}(h)-L_S(h)|\right]=\\&=\mathop{\mathbb{E}}_{S,S'\sim\mathcal{D}^m}\left[\sup_{h\in\mathcal{H}}\frac{1}{m}\left|\sum_{i=1}^m(\ell(h,z'_i)-\ell(h,z_i))\right|\right].\end{aligned}$$

On l'esperança al costat dret de la desigualtat és sobre l'elecció de dos mostres i.i.d. de m elements cadascuna. Per tant, com que l'elecció dels $2m$ vectors es fa de manera idèntica completament idenpendent els uns dels altres, és indiferent intercanviar la variable aleatòria vector z_i per la variable aleatòria z'_i per qualsevol $i\in\{1,...,m\}$; intuïtivament, com que totes les v.a. són equivalents, intercanviar el nom d'elles entre si no altera el resultat. En cas de fer una permuta entre z_i i z'_i per a una i donada, aleshores enlloc d'aparèixer el terme $(\ell(h,z'_i)-\ell(h,z_i))$ a l'equació (4.2), apareixerà el terme $-(\ell(h,z'_i)-\ell(h,z_i))$. Per tant, per qualsevol $\sigma\in\{\pm1\}^m$ es té

$$\begin{aligned}&\mathop{\mathbb{E}}_{S,S'\sim\mathcal{D}^m}\left[\sup_{h\in\mathcal{H}}\frac{1}{m}\left|\sum_{i=1}^m(\ell(h,z'_i)-\ell(h,z_i))\right|\right]\\&=\mathop{\mathbb{E}}_{S,S'\sim\mathcal{D}^m}\left[\sup_{h\in\mathcal{H}}\frac{1}{m}\left|\sum_{i=1}^m\sigma_i(\ell(h,z'_i)-\ell(h,z_i))\right|\right].\end{aligned}$$

Com que aquesta igualtat és certa per qualsevol $\sigma \in \{\pm 1\}^m$, aleshores també serà certa si se selecciona cada σ_i de manera aleatòria uniformement de la distribució uniforme sobre $\{\pm 1\}$, la qual es denotarà $U_\pm$. Així doncs, es té

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

Per la linealitat de l'esperança, es poden permutar les esperances s'obté

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

Fixant ara S i S', sigui $C=\{c_1,\ ...,\ c_m,\ c_1',\ ...,\ c_m'\}$ la col·lecció de $2m$ instàncies de S i S', tan sols es pot prendre el suprem sobre $h\in\mathcal{H}_C$. Obtenint així:

$$\begin{aligned}&\mathop{\mathbb{E}}_{\sigma\sim U_\pm^m}\left[\sup_{h\in\mathcal{H}}\frac{1}{m}\left|\sum_{i=1}^m \sigma_i(\ell(h,z_i')-\ell(h,z_i))\right|\right]\\ &=\mathop{\mathbb{E}}_{\sigma\sim U_\pm^m}\left[\max_{h\in\mathcal{H}}\frac{1}{m}\left|\sum_{i=1}^m \sigma_i(\ell(h,z_i')-\ell(h,z_i))\right|\right].\end{aligned}$$

Fixant també $h\in\mathcal{H}_C$, es denotarà a partir d'ara $\theta_h=\frac{1}{m}\sum_{i=1}^m \sigma_i(\ell(h,z_i')-\ell(h,z_i))$. És fàcil veure que $\mathbb{E}[\theta_h]=0$; a més a més, com θ_h és una mitjana aritmètica de variables aleatòries independents que prenen el seu valor en $[-1,1]$, aleshores per la desigualtat de Hoeffding es té que per a qualsevol $\rho>0$:

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

Aplicant la desigualtat de la unió sobre $h\in\mathcal{H}_C$, aleshores per a qualsevol $\rho>0$ es té

$$\mathbb{P}\left[\max_{h\in\mathcal{H}_C}|\theta_h|>\rho\right]\leq 2|\mathcal{H}_C|e^{(-2m\rho^2)}. \tag{4.2.3}$$

Fent servir el lema 44 es té que (4.2.3) implica

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

Combinant-ho amb la definició de la funció de creixement $\tau_\mathcal{H}$ s'ha demostrat

$$\mathop{\mathbb{E}}_{S\sim\mathcal{D}^m}\left[\sup_{h\in\mathcal{H}}|L_\mathcal{D}(h)-L_S(h)|\right]\leq\frac{4+\sqrt{\log(\tau_\mathcal{H}(2m))}}{\sqrt{2m}}.$$

Tal i com es volia provar. □

Lema 46. *Siguin m, $d \in \mathbb{Z}$ tals que $d \leq m-2$. Aleshores*

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

Demostració:
Es provarà per inducció.
Cas inicial, $d = 1$:
En aquest cas, $\sum_{k=0}^{1} \binom{m}{k} = m+1$; mentre que $\left(\frac{em}{1}\right)^1 = em$. Com que $d = 1 \leq m-2$ implica $3 \leq m$, aleshores trivialment es compleix la desigualtat $\sum_{k=0}^{1} \binom{m}{k} = m+1 \leq \left(\frac{em}{1}\right)^1 = em$.
Hipòtesi d'inducció: $\exists\, d \in \mathbb{Z}$ tal que $\sum_{k=0}^{d} \binom{m}{k} \leq \left(\frac{em}{d}\right)^d$.
Ara es vol provar que el cas d implica el cas $d+1$:
Per hipòtesi d'inducció es té:

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

On fent servir l'aproximació de Stirling $\left(n! \sim \sqrt{2\pi n}(\frac{n}{e})^n\right)$ s'obté

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

On a l'última igualtat s'ha fet servir que $d \leq m-2$. A més a més:

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

Que és el que es volia provar i, per tant, queda provat l'argument inductiu. □

Lema 47. (Lema de Sauer) *Sigui $\mathcal{H}$ una classe d'hipòtesis amb VCdim($\mathcal{H}$) $\leq d < \infty$. Aleshores, per a qualsevol m, $\tau_{\mathcal{H}}(m) \leq \sum_{i=0}^{d} \binom{m}{i}$. En concret, si $m > d+1$, aleshores $\tau_{\mathcal{H}}(m) \leq (em/d)^d$.*

Aquest lema també s'anomena lema de Shelah, lema de Perles o lema de Sauer-Shelah-Perles.

Demostració:
Per tal de provar aquest lema, es provarà el següent enunciat, el qual és més fort i implica el lema de Sauer: per a qualsevol $C = \{ c_1, ..., c_m \}$ es té

$$\forall \mathcal{H}, \quad |\mathcal{H}_C| \leq |\{B \subseteq C : \mathcal{H} \text{ separa } B\}|. \tag{4.2.4}$$

Per fer-ho, s'emprarà un argument inductiu.
Cas inicial $m = 1$:
Com que el conjunt buit, per conveni, sempre es considera que és separat per $\mathcal{H}$, aleshores indiferentment de quina sigui la classe $\mathcal{H}$ ambdós costats de la desigualtat (4.2.4) són iguals a 2 o a 1 solidariament.
Hipòtesi d'inducció: Se suposarà com a cert que per a qualsevol $k < m$ se satisfà (4.2.4).
A continuació es provarà que aquesta assumpció implica que (4.2.4) també es compleix per conjunts de mida m.
Fixi's $\mathcal{H}$ i $C = \{ c_1, ..., c_m \}$. Es denota $C' = \{ c_2, ..., c_m \}$ i es defineixen els següents conjunts:

$$Y_0 = \{(y_2, ..., y_m) : (0, y_2, ..., y_m) \in \mathcal{H}_C \vee (1, y_2, ..., y_m) \in \mathcal{H}_C\} \ ,$$

$$Y_1 = \{(y_2, ..., y_m) : (0, y_2, ..., y_m) \in \mathcal{H}_C \wedge (1, y_2, ..., y_m) \in \mathcal{H}_C\}.$$

Els quals són, respectivament, Y_0 el conjunt de les funcions $g{:}C' \to \{0, 1\}$ tals que existeix $h \in \mathcal{H}_C$ tal que g coincideix amb h a C' i $h(c_1) = 0$ o $h(c_1) = 1$; i Y_1 el conjunt de les funcions g: $C' \to \{0, 1\}$ tals que existeix $h \in \mathcal{H}_C$ tal que g coincideix amb h a C' i $h(c_1) = 0$ i existeix $h' \in \mathcal{H}_C$ tal que g coincideix amb h' a C' i $h'(c_1) = 1$.
És facil comprovar que $Y_0 = \mathcal{H}_{C'}$ i que $|\mathcal{H}_C| = |Y_0| + |Y_1|$. Aplicant la hipòtesi d'inducció al conjunt C' de mida $m - 1$:

$$|Y_0| = |H_{C'}| \leq |\{B \subseteq C' : \mathcal{H} \text{ separa } B\}| = |\{B \subseteq C' : c_1 \notin B \ i \ \mathcal{H} \text{ separa } B\}|.$$

Es defineix a continuació el conjunt $\mathcal{H}' \subset \mathcal{H}$ de la forma

$$\mathcal{H}' = \{h \in \mathcal{H} : \exists h \in \mathcal{H} \text{ tal que } (1 - h'(c_1), h'(c_2), ..., h'(c_m)) = (h(c_1), h(c_2), ..., h(c_m))\}.$$

És a dir, $\mathcal{H}'$ és el conjunt dels parells d'hipòtesis d'$\mathcal{H}$ tals que coincideixen en C' però que no coincideixen en c_1.
De la definició d'$\mathcal{H}'$ resulta evident que per qualsevol conjunt $B \subseteq C'$ que sigui separat per $\mathcal{H}'$, aleshores el conjunt $\{ c_1 \} \cup B$ també és separat per $\mathcal{H}'$ i viceversa.
De la mateixa manera que $Y_0 = \mathcal{H}_{C'}$, es pot observar que $Y_1 = \mathcal{H}'_{C'}$. Com que C' té mida $m - 1$, anàlogament aplicant la hipòtesi d'inducció

$$\begin{aligned} |Y_1| = |\mathcal{H}'_{C'}| &\leq |\{B \subseteq C' : \mathcal{H}' \text{ separa } B\}| \\ &= |\{B \subseteq C' : \mathcal{H}' \text{ separa } B \cup \{c_1\}\}| \\ &= |\{B \subseteq C : c_1 \in B \text{ i } \mathcal{H}' \text{ separa } B \cup \{c_1\}\}| \\ &\leq |\{B \subseteq C : c_1 \in B \text{ i } \mathcal{H} \text{ separa } B \cup \{c_1\}\}|. \end{aligned}$$

Així doncs, ajuntant les cotes de $|Y_0|$, $|Y_1|$ i que $|\mathcal{H}_C| = |Y_0| + |Y_1|$ s'ha provat

$$\begin{aligned} |\mathcal{H}_C| &= |Y_0| + |Y_1| \\ &\leq |\{B \subseteq C : c_1 \notin B \text{ i } \mathcal{H} \text{ separa } B \ \cup \ \{c_1\}\}| \\ &\qquad + |\{B \subseteq C : c_1 \in B \text{ i } \mathcal{H} \text{ separa } B \ \cup \ \{c_1\}\}| \\ &= |\{B \subseteq C : \mathcal{H} \text{ separa } B\}|. \end{aligned}$$

Concloent així la prova de (4.2.4).
No obstant, si bé s'ha provat (4.2.4), encara resta provar per què la veracitat d'aquest enunciat implica la veracitat del lema de Sauer.

Si $\text{VCdim}(\mathcal{H}) \leq d$, aleshores cap conjunt de mida major que d és separat per $\mathcal{H}$; per tant

$$|\{B \subseteq C : \mathcal{H} \text{ separa } B\}| \leq \sum_{i=0}^{d} \binom{m}{i}.$$

Es pot provar que el sumatori que hi ha a la banda dreta de la desigualtat és menor o igual a $(em/d)^d$ per qualsevol $m > d + 1$, com a conseqüència del lema 46. Així doncs, es té el resultat que es volia provar i, per tant, queda demostrat el lema de Sauer. □

Ara sí, ja estan disponibles totes les eines necessàries per demostrar el teorema 41.

Demostració:
Per tal de provar que els 6 enunciats són equivalents, cal provar una cadena d'implicacions d'on s'obtingui que qualsevol enunciat implica tots els altres. En concret, es consideraran dues cadenes d'implicacions, les quals al combinar-les proporcionen el resultat desitjat. En primer lloc, es probarà la cadena d'implicacions $1 \Rightarrow 2$, $2 \Rightarrow 3$, $3 \Rightarrow 4$, $4 \Rightarrow 6$ i $6 \Rightarrow 1$. En segon lloc, es provarà la cadena $1 \Rightarrow 2$, $2 \Rightarrow 5$, $5 \Rightarrow 6$ i $6 \Rightarrow 1$. Combinant-les, s'obté un esquema com el mostrat a la figura 4.1:

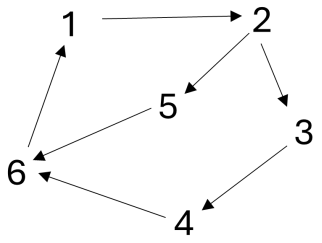

Figura 4.1: Esquema de les implicacions dels diferents enunciats del teorema

On hom pot comprovar que es pot construir una cadena d'implicacions (i per tant una implicació) des de qualsevol enunciat fins a qualsevol altre.

1 ⇒ 2:

Així doncs, començant amb les proves de les diferents implicacions, s'observa que la implicació $1 \Rightarrow 2$ és una reformulació del corol·lari 24, el qual ja ha sigut provat anteriorment. Per tant, aquesta implicació ja ha sigut provada previament.

2 ⇒ 3:

Tot seguit, s'observa que la implicació $2 \Rightarrow 3$ és trivial. Si qualsevol algoritme ERM és un aprenent PAC agnòstic exitós per a la classe $\mathcal{H}$, aleshores $\mathcal{H}$ és aprensible PAC agnòsticament per qualsevol algoritme ERM i, per tant, és aprensible PAC agnòsticament.

3 ⇒ 4:

En següent lloc, pel que fa a la implicació $3 \Rightarrow 4$, es vol veure que qualsevol classe aprensible PAC agnòsticament és també aprensible PAC. Recordant les definicions d'aprensibilitat PAC agnòstica (definició 15) i d'aprensibilitat PAC (definició 11), es pot observar (com ja s'ha indicat en aquesta mateixa memòria a la secció corresponent) que

l'aprensibilitat PAC és un cas concret de l'aprensibilitat PAC agnòstica; més concretament, sota les condicions de l'assumpció de realitzabilitat (la qual és una condició necessària per parlar d'aprenentatge PAC), la noció d'aprensibilitat PAC agnòstica es converteix en la d'aprensibilitat PAC. Així doncs, trivialment si $\mathcal{H}$ és aprensible PAC agnòsticament, aleshores $\mathcal{H}$ també serà, en concret, aprensible PAC.

4 ⇒ 6:

A fi de provar ara la implicació $4 \Rightarrow 6$, convé recordar el teorema NFL (teorema 32) així com el corol·lari 37, el qual és un resultat que se'n deriva. Volem veure que qualsevol classe $\mathcal{H}$ que sigui aprensible PAC tindrà dimensió VC finita, és a dir que existeix $m \in \mathbb{N}$ tal que no existirà cap subconjunt C de mida major o igual que m del domini $\mathcal{X}$ tal que sigui separat per $\mathcal{H}$.

Com que $\mathcal{H}$ és aprensible PAC, per definició, donats ε, $\delta \in (0, 1)$ qualssevol existeix $m_{\mathcal{H}}(\varepsilon, \delta) \in \mathbb{N}$ tal que per a qualsevol $m \geq m_{\mathcal{H}}(\varepsilon, \delta)$, essent m la mida d'una mostra S d'elements de $\mathcal{X} \times \{0, 1\}$ generats i.i.d., aleshores per un algoritme ERM es compleix

$$\mathop{\mathbb{P}}_{S \sim \mathcal{D}^m}(L_{\mathcal{D}}(A(S)) > \varepsilon) < \delta.$$

Aleshores, prenent en concret $\varepsilon < 1/8$ i $\delta < 1/7$, es té per la definició d'aprensibilitat PAC que existeix $m \in \mathbb{N}$ tal que per a qualsevol $m' \geq m$, donada una mostra qualsevol S de mida m' d'elements de $\mathcal{X} \times \{0, 1\}$ generats i.i.d., aleshores per un algoritme ERM es compleix

$$\mathop{\mathbb{P}}_{S \sim \mathcal{D}^m}(L_{\mathcal{D}}(A(S)) > 1/8) < 1/7.$$

Que no és altra cosa que la negació, sota les mateixes hipòtesis, del resultat del corol·lari 37. Per tant, pel contrarrecíproc es té que no existeix cap subconjunt C de mida $2m$ de $\mathcal{X}$ tal que $\mathcal{H}$ separi C. Per tat, per definició de la dimensió VC es té $\text{VCdim}(\mathcal{H}) < 2m < \infty$.

Tal i com es volia provar.

6 ⇒ 1:

A continuació. es provarà la implicació $6 \Rightarrow 1$, la qual és, sens dubte, la més extensa i complicada de totes.

La demostració d'aquesta implicació constarà de dues parts ben diferenciades; en primer lloc, es provarà que si la dimensió VC de $\mathcal{H}$ és finita aleshores la seva mida efectiva al restringir-la sobre un conjunt $C \subset \mathcal{X}$ ($|\mathcal{H}_C|$ és tan sols $O(|C|^d)$, malgrat que $\mathcal{H}$ sigui infinita. És a dir, la mida de $\mathcal{H}_C$ no creix exponencialment segons $|C|$, sinó que ho fa polinòmicament. Tot seguit, es provarà que tota classe d'hipòtesis $\mathcal{H}$ de mida efectiva petita posseeix la propietat de convergència uniforme, on mida efectiva petita fa referència a classes per les quals $|\mathcal{H}_C|$ creix polinòmicament segons $|C|$.

Així doncs, a fi de provar en primer lloc la dependència polinòmica de $|\mathcal{H}_C|$ respecte $|C|$ per a qualsevol classe $\mathcal{H}$ amb $\text{VCdim}(\mathcal{H}) = d < \infty$, es consideraran els casos $m \leq d$ i $m > d$ per separat, on m és la mida de C.

En primer lloc, pel que fa al cas $m \leq d$, trivialment aleshores C és finit, $\mathcal{H}_C$ és finita i per tant té la propietat de convergència uniforme.

Passant ara al cas $m > d$, n'hi ha prou amb emprar el lema de Sauer , és a dir el lema 47, el qual posa de manifest que quan la mida m del conjunt C esdevé major que la dimensió VC d de la classe $\mathcal{H}$, aleshores la funció de creixement incrementa polinòmicament respecte m.

Proseguint ara amb la segona part de la demostració, el nou objectiu és veure que qualsevol classe $\mathcal{H}$ de mida efectiva petita gaudeix de la propietat de convergència uniforme.

Finalment, només resta provar perquè aquest teorema implica que tota classe d'hipòtesis $\mathcal{H}$ de mida efectiva petita té la propietat de convergència uniforme per tal de concloure la prova de la implicació $6 \Rightarrow 1$.

Per fer-ho, es provarà que la complexitat de la mostra $m_{\mathcal{H}}^{UC}(\varepsilon, \delta)$ està acotada d'una de les següents formes

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

O pel contrari

$$m_{\mathcal{H}}^{UC}(\varepsilon, \delta) \leq \frac{32}{\delta\varepsilon}.$$

Pel lema de Sauer (Lema 47), per $m > d$ es té $\tau_{\mathcal{H}}(2m) \leq (2em/d)^d$; aquesta desigualtat, en combinació amb el teorema 45 implica

$$\mathop{\mathbb{P}}_{S\sim\mathcal{D}^m}\left(|L_S(h) - L_{\mathcal{D}}(h)| \leq \frac{4 + \sqrt{d\log(2em/d)}}{\delta\sqrt{2m}}\right) \geq 1 - \delta. \tag{4.2.5}$$

Suposi's en primer lloc que $4 \geq \sqrt{d\log(2em/d)}$; aleshores

$$|L_S(h) - L_{\mathcal{D}}(h)| \leq \frac{4\sqrt{2}}{\delta\sqrt{m}} \leq \varepsilon \Leftrightarrow \frac{32}{\delta^2\varepsilon^2} < m.$$

Per tant, com que (4.2.5) es compleix per a qualsevol $m > \frac{32}{\delta^2\varepsilon^2}$ i $m_{\mathcal{H}}^{UC}(\varepsilon, \delta)$ és la mida mínima de la mostra tal que es compleix, es té $m_{\mathcal{H}}^{UC}(\varepsilon, \delta) \leq \frac{32}{\delta^2\varepsilon^2}$.

Per altra banda, suposi's ara $4 \leq \sqrt{d\log(2em/d)}$; aleshores

$$|L_S(h) - L_{\mathcal{D}}(h)| \leq \frac{1}{\delta}\sqrt{\frac{2d\log(2em/d)}{m}} \leq \varepsilon \Leftrightarrow$$

$$\Leftrightarrow m \geq \frac{2d\log(m)}{(\delta\varepsilon)^2} + \frac{2d\log(2e/d)}{(\delta\varepsilon)^2}. \tag{4.2.6}$$

A fi de concloure finalment la demostració de la implicació $6 \Rightarrow 1$, fent us del lema 43, aleshores (4.2.6) implica

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

I, per un argument anàleg al cas $4 \geq \sqrt{d\log(2em/d)}$, aleshores $m_{\mathcal{H}}^{UC}(\varepsilon, \delta) \leq 4\frac{2d}{(\delta\varepsilon)^2}\log\left(\frac{2d}{(\delta\varepsilon)^2}\right) + \frac{4d\log(2e/d)}{(\delta\varepsilon)^2}$ i, per tant $\mathcal{H}$ té la propietat de convergència uniforme. Així doncs, ha quedat provat $6 \Rightarrow 1$.

2 ⇒ 5:

Proseguint amb la demostració del teorema 41, el següent pas és provar la implicació $2 \Rightarrow 5$.

De manera anàloga a la demostració de la implicació $3 \Rightarrow 4$, per les definicions d'aprensibilitat PAC (definició 11) i d'aprensibilitat PAC agnòstica (definició 15), s'observa

que l'aprensibilitat PAC és un cas concret de l'aprensibilitat PAC agnòstica (quan se satisfà l'assumpció de realitzabilitat); per tant, si qualsevol agoritme ERM és un aprenent PAC agnòstic exitós per a $\mathcal{H}$, aleshores també serà un aprenent PAC exitós per a $\mathcal{H}$. Queda provada així la implicació $2 \Rightarrow 5$.

5 ⇒ 6:

Finalment, ja només resta provar la implicació $5 \Rightarrow 6$;

Anàlogament a la prova de la implicació $4 \Rightarrow 6$, si qualsevol algoritme ERM és un aprenent PAC exitós per a $\mathcal{H}$, com a conseqüència del teorema NFL i el corol·lari 37, per qualsevol algoritme ERM es compleix:

$$\underset{S\sim\mathcal{D}^m}{P}(L_{\mathcal{D}}(A(S)) > 1/8) < 1/7.$$

I tal i com s'ha vist a la prova de $4 \Rightarrow 6$, això implica que la dimensió VC de $\mathcal{H}$ és finita i, per tant, queda provat $5 \Rightarrow 6$.

Així doncs, com que s'han provat totes les implicacions, ha quedat demostrat el teorema 41. □

La versió que s'acaba de mostrar i provar del teorema fonamental de l'aprenentatge estadístic és l'anomenada versió qualitativa. Aquesta versió posa de manifest que qualsevol classe d'hipòtesis on les seves integrants són funcions de classificació binària sobre un domini (i mesurant el seu error segons la funció de pèrdua $1 - 0$), aleshores són equivalents que la classe sigui aprensible PAC agnòsticament, que tingui la propietat de convergència uniforme i que tingui dimensió VC finita. És a dir, per a qualssevol valors ε i $\delta \in (0, 1)$, existeixen funcions $m_{\mathcal{H}}(\varepsilon, \delta)$: $(0, 1)^2 \to \mathbb{N}$ que determinen la complexitat de la mostra S per tal de garantir les condicions d'aprensibilitat PAC agnòstica, aprensibilitat PAC i convergència uniforme respectivament. Així doncs, se l'anomena versió qualitativa ja que demostra l'existència d'aquestes funcions (i per tant, de les cotes sobre el nombre d'elements a la mostra), però no proporciona cap informació sobre el valor de la complexitat. No obstant, existeix una altra versió del teorema fonamental de l'aprenentatge estadístic, la qual sí que proporciona unes cotes inferiors i superiors a aquests valors; aquesta s'anomena la versió quantitativa del teorema fonamental de l'aprenentage estadístic, i s'enunciarà i demostrarà a continuació:

Teorema 48. (teorema fonamental de l'aprenentatge estadístic; versió quantitativa) *Sigui $\mathcal{H}$ una classe d'hipòtesis d'un domini $\mathcal{X}$ a $\{0, 1\}$ (classiifcació binària) i sigui la funció de pèrdua $0 - 1$ la funció de pèrdua. Suposi's que VCdim($\mathcal{H}$) = $d < \infty$. Aleshores, existeixen constants absolutes C_1 i C_2 tals que:*

1. *$\mathcal{H}$ té la propietat de convergència uniforme amb una complexitat de la mostra tal que:*
$$C_1\frac{d + \log(1/\delta)}{\varepsilon^2} \leq m^{UC}_{\mathcal{H}}(\varepsilon, \delta) \leq C_2\frac{d + \log(1/\delta)}{\varepsilon^2}.$$
2. *$\mathcal{H}$ és aprensible PAC agnòsticament amb una complexitat de la mostra tal que:*
$$C_1\frac{d + \log(1/\delta)}{\varepsilon^2} \leq m_{\mathcal{H}}(\varepsilon, \delta) \leq C_2\frac{d + \log(1/\delta)}{\varepsilon^2}.$$
3. *$\mathcal{H}$ és aprensible PAC amb una complexitat de la mostra tal que:*
$$C_1\frac{d + \log(1/\delta)}{\varepsilon} \leq m_{\mathcal{H}}(\varepsilon, \delta) \leq C_2\frac{d\log(1/\varepsilon) + \log(1/\delta)}{\varepsilon}.$$

Malgrat pugui resultar d'interés la demostració d'aquesta versió del teorema, donades la seva extensió i la quantitat de resultats externs requerits, s'ha optat per ometre-la. En qualsevol cas, si algun lector desitja veure aquesta prova, pot consultar-la al llibre *Understanding machine learning: From theory to algorithms.* [1], entre les pàgines 392 i 401.

Capítol 5

Una aplicació pràctica; una cota de generalització en aprenentatge profund

Al llarg dels capítols previs s'han introduit i desenvolupat tot un seguit de conceptes i resultats matemàtics de l'àmbit de l'aprenentatge automàtic fins arribar al teorema fonamental de l'aprenentatge estadístic, del qual s'han vist dues versions; la versió qualitativa i la versió quantitativa. No obstant, fins ara tan sols s'ha explorat aquest terreny des del punt de vista de l'estudi teòric, essent un àmbit clarament aplicat. Així doncs, en aquesta última secció de la memòria s'introduirà un tema punter de recerca actual, així com els conceptes que es requereixen per presentar-lo, i es possarà de manifest la relació del mateix amb els resultats teòrics obtinguts al llarg del treball.

El resultat de recerca que es preté discutir, així com les definicions i teoremes requerides per la comprensió del mateix (les introduides a la secció 5.2) està extret de l'article *On Generalitzation Bounds for Neural Networks with Low Rank Layers* dels doctors Andrea Pinto, Akshay Rangamani i Tomaso Poggio; publicat a la revista *Proceedings of Machine Learning Research, 30* [2].

L'objecte d'estudi en qüestió és acotar la complexitat Gaussiana de xarxes neuronals de baix rang. Per tal de poder comprendre en detall el resultat que es presentarà, convé primer tractar tots els conceptes teòrics indispensables.

5.1 Xarxes neuronals

Les xarxes neuronals (*Neural networks*) són un model d'aprenentatge automàtic modelat en base al funcionament i estructura del cervell humà. A grans trets, a mode de idea general abans d'entrar més en detall, les xarxes neuronals estan composades per un conjunt de nodes, els quals estan organitzats en diferents capes que estan connectades contiguament entre si (és a dir, els nodes de la capa i estan connectats amb els nodes de les capes $i-1$ i $i+1$), emulant les neurones del cervell humà.

Aquesta modelització de les neurones va ser introduida per primera vegada per Warren McCulloch i Walter Pitts a l'article *A logical calculus of ideas immanent in nervous activity* [11].En aquest, es defineix la neurona com la unitat de càlcul bàsica sobre la qual es construeix una xarxa neuronal, replicant el funcionament de la seva homòloga biològica.

És a dir, rep una informació d'entrada (una sinapsi), efectua un càlcul i aplica una funció no lineal sobre el resultat obtingut per tal de determinar la informació de sortida. Aquesta funció no lineal és la denominada *funció d'activació*, que es denotarà com ϕ. Durant molts anys, a fi de ser fidedignes al funcionament real de les neurones, la funció d'activació va ser la funció llindar; és a dir, la neurona emet o no una informació de sortida en funció de si l'estímul rebut és suficient. No obstant, des de ja fa uns anys, per motius d'optimització, s'ha abandonat aquesta idea i s'ha reemplaçat la funció llindar per altre funcions no lineals; un exemple famós és la denominada funció d'activació ReLU.

La funció d'activació es communa per tots els nodes de la xarxa, excepte pel nodes de la última capa, i pot prendre multitut de formes diferents en funció de la xarxa que es vulgui construir. Pel cas que es vol estudiar en aquest treball, es tracta d'una funció linear a troços i 1-Lipschitz.

Un cop vista la noció intuitiva darrere d'una neurona, així com s'ha introduït el concepte de funció d'activació, és hora de definir formalment què és un node.

Definició 49: *Sigui $\mathcal{N}$ una xarxa neuronal. Un node (neurona) és una funció $\varphi : \mathbb{R}^n \to \mathbb{R}$ definida com:*

$$\varphi(\mathbf{x}) = \phi(\mathbf{w}\mathbf{x}^\top), \tag{5.1.1}$$

on $\mathbf{x} = (x_1, ..., x_n)$ és el vector que actua com la informació d'entrada ("input") que rep el node, essent n la dimensió d'aquest vector; $\mathbf{w} = (w_1, ..., w_n)$ és un vector de pesos propi de cada node; i ϕ és la funció d'activació.

En una xarxa neuronal profunda els nodes s'agrupen en capes, les quals en conjunt constitueixen la xarxa neuronal. Sigui $\mathcal{N}$ una xarxa neuronal, s'anomena capa número i al conjunt $\mathcal{M}_i$ format per l'i-èssim grup de nodes que reben la mateixa informació d'entrada entre ells. En concret, la capa $\mathcal{M}_1$ està formada per tots els nodes que reben com a informació d'entrada les pròpies dades que tracta la xarxa; la capa $\mathcal{M}_2$ per tots els nodes que reben com a informació d'entrada el vector de les informacions de sortida dels nodes de la capa $\mathcal{M}_1$ i així successivament. Formalment:

Definició 50: *Sigui $\mathcal{N}$ una xarxa neuronal. Es defineix una capa com la funció $F : \mathbb{R}^a \to \mathbb{R}^b$ de la forma:*

$$F(\mathbf{x}) = \phi(\mathbf{W}\mathbf{x}^\top),$$

on $\mathbf{W} = (\mathbf{w}_1^\top, ..., \mathbf{w}_b^\top)^\top \in \mathbb{R}^{b\times a}$ és la matriu de pesos d'aquesta capa, la qual està formada pels b vectors de pesos colocats en files dels b nodes d'aquesta capa. El terme $b \in \mathbb{R}$ indica la dimensió d'aquesta capa, és a dir el nombre de nodes que la formen. $\mathbf{x}$ és la informació d'entrada que rep la capa; si és la primera, aleshores $\boldsymbol{x}$ són els paràmetres[1] *de la dada concreta de totes les de la mostra que s'estigui processant en aquell instant; si no és la primera, aleshores $\boldsymbol{x}$ és el vector amb les informacions de sortida dels nodes de la capa anterior.*

Tal i com s'ha posat de manifest, cada capa rep com a informació d'entrada la informació de sortida de la capa anterior; és a dir, definint la capa i-èssima com la funció $F_i : \mathbb{R}^{n_{i-1}} \to \mathbb{R}^{n_i}$, la seva informació d'entrada és la imatge de la funció F_{i-1}; on n_{i-1}, n_i

[1] és a dir, les coordenades del vector informació d'entrada. En argot d'informàtica es denominen característiques.

són el nombre de nodes de les cape $i-1$, i respectivament. Així doncs, la operació per la qual les capes interaccionen entre elles és la composició de funcions. Aquesta noció condueix a la definició formal de xarxa neuronal de transmissió endavant ("feedforward").

Definició 51: [2] *Sigui $\mathcal{N}$ una xarxa neuronal. Aleshores $\mathcal{N}$ es defineix com una funció $f : \mathbb{R}^{n_0} \to \mathbb{R}^{n_L}$ de la forma:*

$$f(\mathbf{x}) = \mathbf{W}_L\phi(\mathbf{W}_{L-1}(...\phi(\mathbf{W}_1\mathbf{x})...)), \tag{5.1.2}$$

on $\mathbf{W}_i$ és la matriu de la capa i; ϕ és la funció d'activació; $\mathbf{x}$ és el vector de paràmetres de la dada que està tractant la xarxa; en un instant donat; L és el nombre de capes que té la xarxa; n_L és la dimensió (nombre de nodes) de l'última capa i n_0 la dimensió de $\mathbf{x}$, és a dir el nombre de característiques que tenen les dades que tracta la xarxa.

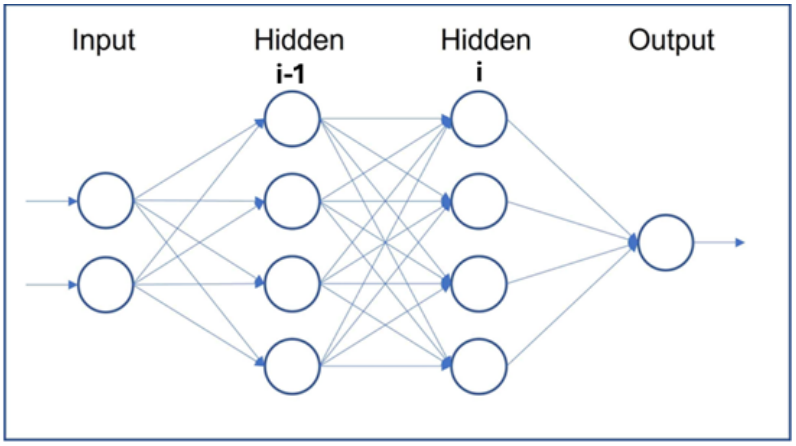

Figura 5.1: Esquema de les capes d'una xarxa neuronal

A la figura 5.1 es pot observar l'esquema general d'una xarxa neuronal i la distribució de les seves capes.

Tal i com es pot observar a (5.1.2), la xarxa neuronal simplement és el resultat d'aplicar ordenadament totes les capes de nodes que la conformen. Ara que ja han sigut definits formalment els conceptes de node, capa i xarxa, resulta convenient donar una idea intuitiva sobre el funcionament de la mateixa. Un bon punt de pàrtida és exposar els tipus de capes que hi ha en una xarxa i com s'estructuren.

Existeixen tres tipus diferents de capes, la capa d'entrada, les capes ocultes i la capa de sortida. En primer lloc, la capa d'entrada és la primera capa de nodes de la xarxa, els seus nodes són el que reben la informació d'entrada ("input"); Aquesta informació d'entrada pren la forma d'un vector les components del qual són les característiques que es prenen

[2] Al veure la definició de la funció d'una xarxa neuronal, hom podria pensar que falta aplicar una última vegada la funció d'activació, en concret, aplicar-la també als nodes de l'última capa de la xarxa. Així doncs, es podria pensar que la deifinició està malament i la funció hauria de ser de la forma:

$$f(\mathbf{x}) = \phi(\mathbf{W}_L\phi(\mathbf{W}_{L-1}(...\phi(\mathbf{W}_1\mathbf{x})...))).$$

No obstant, això no és veritat. Tal i com s'ha expressat just després de la definició de funció d'activació, aquesta és comuna a tots els nodes de la xarxa excepte els de l'última capa. Això es deu que, donada la forma d'algunes funcions d'activació, si aquesta s'apliqués també als nodes de l'última capa, aleshores la xarxa neuronal veuria molt restringit el rang de valors que pot retornar. Així doncs, generalment s'opta per considerar la funció d'activació d'aquests nodes com la funció identitat, obtenint així l'expressió (5.1.2).

en consideració de les dades amb les que es treballen. Per exemple, pel cas de problema de classificació binària discutit al capítol 2 on un veterinari vol classificar exemplars d'una nova espècie de cavall segons si tenen sobrepés o no, les dades serien els exemplars d'aquesta espècie i les característiques les seves mesures de longitut i pes. En segon lloc, les capes ocultes són totes les capes intermitges entre la capa d'entrada i la capa de sortida; els seus nodes reben com a informació d'entrada la informació de sortida proporcionada per nodes de la capa prèvia i emeten una informació de sortida als nodes de la capa posterior, els quals la rebran com a informació d'entrada i així successivament. Aquesta informació d'entrada pren la forma d'un vector on cada component és la informació de sortida d'un dels nodes de la capa anterior. Finalment, la capa de sortida és l'última capa de nodes de tota la xarxa; els seus nodes són els que generen la informació de sortida ("output") que retorna la xarxa.

Tot seguit, tal i com s'ha avançat prèviament, a fi de comprendre millor la naturalesa de la interacció entre els nodes de les diferents capes convé explicar què signifiquen els vectors (i per tant les matrius) de pesos, així com les seves components. El vector de pesos d'un node indica quantitativament el grau "d' importància"que aquest node li dòna a cadascuna de les instàncies de la informació d'entrada que reb. Pel cas d'un node de la capa d'entrada, la rellevància que li assigna a cada paràmetre concret de les dades que rep (tornant a l'exemple del zoòleg, la importància que li dona al pes i a la longitut respectivament). Pel cas d'un node d'una capa oculta o de la capa de sortida, la importància que li dona a la informació de sortida de cada node de la capa anterior respectivament. Aquesta importància es reflecteix de la següent manera; tal i com es pot veure a (5.1.1), quan un node rep el vector amb la informació d'entrada, multiplica cada component del vector (cada paràmetre de les dades / la sortida de cada node de la cap anterior) pel pes corresponent associat a aquella component d'acord al vector de pesos. Finalment, suma tots aquests productes i s'aplica la funció d'activació a aquest valor. Així doncs, en funció del factor multiplicatiu (component del vector pes) que s'hagi aplicat a cada component de la informació d'entrada, aquesta haurà tingut un major o menor impacte en el càlcul del valor del terme al qual se li aplicarà la funció d'activació.

Un cop vist a grans trets l'estructura i el funcionament d'una xarxa neuronal, hom pot observar que el model descrit fins ara és un model estàtic (és a dir, no s'actualitza en base a l'experiència) i preconstruït (elements com les matrius de pesos tenen els seus valors inicialitzats al començar el procés de classificació). Aquestes característiques xoquen directament amb la idea d'un model d'aprenentatge automàtic, ja que la funció d'aquest és poder resoldre cada cop millor un cert tipus de problema a mesura que va rebent nous casos requerint el mínim coneixement previ possible; pel cas de problema de classificació que s'està estudiant en aquest treball, l'objectiu seria classificar correctament els elements del domini que se li subministrin en base a una mostra d'entrenament prèviament classificada i dotant al model d'un cert coneixement previ. Així doncs, és necessari introduir les nocions de com s'inicialitzen i com s'actualitzen els valors de les matrius de pesos de la xarxa.

En primer lloc, pel que fa a la inicialització dels valors de les matrius de pesos, existeixen moltes tècniques diferents, cadascuna amb els seus avantatges i inconvenients, consistents en inicialitzar tots els valors de les matrius segons algun criteri concret; alguns exemples són la inicialització constant (inicialitzar tots els valors a una constant, com per exemple 0) o la inicialització aleatòria (seleccionar els valors segons alguna llei aleatòria, com una distribució gaussiana centrada en 0 i amb desviació típica σ^2) entre d'altres.

Per altra banda, pel que fa a l'actualització dels valors d'aquestes matrius, cada cop que la xarxa neuronal rep una nova instància recalcula les matrius de pesos actualitzant-les

fent servir mètodes d'optimització estocàstica basats en subgradients. Hom pot pensar en les noves matrius com:

$$W = W_0 + \Delta W,$$

on W_0 és la matriu abans de la nova instància i ΔW l'actualització de la mateixa.

Referent a aquest apartat de l'actualització dels valors de les matrius de pesos, per tal d'optimitzar qüestions referents als temps i complexitats de computació, en el desenvolupament del resultat que es pretén estudiar s'empra un procediment anomenat LoRA (*Low Rank Adaptation*). Aquest model hipotitza que l'actualització de la matriu té rang intrínsec baix i es pot descomposar en el producte de dues matrius de rang molt menor. Formalment, sigui $W_0 \in \mathbb{R}^{d\times k}$ una matriu de pesos entrenada prèviament d'una capa d'una xarxa neuronal i ΔW la seva actualització, es restringeix la seva actualització descomposant-la de la següent forma:

$$W_0 + \Delta W = W_0 + BA,$$

on $B \in \mathbb{R}^{d\times r}$, $A \in \mathbb{R}^{r\times k}$ i el rang $r << min(d, k)$.

Un cop feta aquesta introducció al camp d'estudi on s'ubica el resultat que es vol discutir, a la següent secció es presentarà el problema que es preté abordar, s'introduiran tots els conceptes i resultats matemàtics necessaris per tal de desenvolupar el teorema desitjat i s'enunciarà el mateix.

5.2 Presentació del problema i del resultat

En primer lloc, fent una breu descripció del problema, així com dels seus elements, es vol entrenar models d'aprenentatge profund per tal de classificar elements d'un conjunt $\mathcal{X} \subseteq \mathbb{R}$ segons classificacions d'un conjunt $\mathcal{Y} \subseteq \mathbb{R}^C$. L'objecte d'estudi són xarxes neuronals que es denotaran com $f_{\mathbf{W}}(x) = \mathbf{W}_L\phi(\mathbf{W}_{L-1}\phi(...\phi(\mathbf{W}_1\mathbf{x})...))$ essent $\mathbf{W}_i \in \mathbb{R}^{n_i\times n_{i-1}}$ la matriu de pesos corresponent a la capa i de la xarxa i essent ϕ: $\mathbb{R} \to \mathbb{R}$ una funció d'activació 1-Lipschitz que s'aplica coordenada a coordenada.

Definició 52: *Sigui $L \in \mathbb{N}$, es denota la classe de les xarxes neuronals de L capes amb norma espectral i rang acotats com:*

$$\mathcal{F}_L = \{f_{\mathbf{W}}(\mathbf{x}) = \mathbf{W}_L\phi(\mathbf{W}_{L-1}\phi(...\phi(\mathbf{W}_1\mathbf{x})...)) \mid ||\mathbf{W}_i||_2 \leq B_i \ , \ \text{rang}(\mathbf{W}_i) \leq r_i\}.$$

En la teoria de l'aprenentatge estadístic sovint es fa servir la complexitat Gaussiana empírica per tal de computar cotes de generalització. La complexitat Gaussiana mesura la mida d'una classe de funcions $\mathcal{F}$ relacionant-la al conjunt d'imatges d'una mostra $S = \{\mathbf{x}_1, ..., \mathbf{x}_m\}$ i subseqüentment mesurant la seva amplitud Gaussiana.

Definició 53: *Sigui $\mathcal{F}$ una classe de funcions, sigui $S = \{\boldsymbol{x}_1, ..., \boldsymbol{x}_m\}$ una mostra, sigui $\gamma = \{\gamma_i\}_{i=1}^m$ una successió de variables gaussianes $\gamma_i \sim \mathcal{N}(0, 1)$. Aleshores es defineix l'amplitud Gaussiana de la classe de funcions $\mathcal{F}$ com:*

$$\widehat{\mathcal{G}}_S(\mathcal{F}) = \frac{1}{m}\ \mathbb{E}_\gamma \sup_{f\in\mathcal{F}} \sum_{i=1}^{m} \gamma_i f(\mathbf{x}_i).$$

Fent un breu recordatori de l'apartat sobre funcions de pèrdua, donats un conjunt de domini $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$, donada una distribució $\mathcal{D}$ sobre $\mathcal{Z}$, donada una classe de funcions $\mathcal{H}$, donada una funció $h : \mathcal{X} \to \mathcal{Y}$, $h \in \mathcal{H}$, donada una funció de pèrdua $\ell : \mathcal{H} \times \mathcal{Z} \to \mathbb{R}^+$ i

donada una mostra $S = (z_1, ..., z_m) \in \mathcal{Z}^m$ es defineix la funció de risc (Definició 18) de h com:

$$L_{\mathcal{D}}(h) := \mathop{\mathbb{E}}_{z\sim\mathcal{D}}[\ell(h, z)].$$

I es defineix el risc empíric de h com:

$$L_S(h) := \frac{1}{m}\sum_{i=1}^{m} \ell(h, z_i).$$

Pel model que s'està emprant, la funció h és la funció de la xarxa neuronal $f_{\mathbf{W}}$. Amb aquests elements presents, és pot definir el concepte d'error de generalització, el qual no és altra cosa que la diferència entre la funció de risc i el risc empíric.

Definició 54: *Sigui $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ un conjunt de domini, sigui $\mathcal{D}$ una distribució sobre $\mathcal{Z}$, sigui $S = \{(x_i, y_i)\}_{i=1}^m$ una mostra d'elements de $\mathcal{Z}$ seleccionada i.i.d. segons $\mathcal{D}$, sigui $f_{\boldsymbol{W}} : \mathbb{R}^k \to \mathbb{R}^d$ un model, sigui $g(\boldsymbol{z}) = \ell(f_{\boldsymbol{W}}(x), \boldsymbol{y})$. Es defineix l'error de generalització de g com:*

$$\varepsilon_{gen} = \mathop{\mathbb{E}}_{\mathbf{z}\sim\mathcal{D}}[g(\mathbf{z})] - \frac{1}{m}\sum_{i=1}^{m} g(\mathbf{z}_i)).$$

Aquesta magnitud no és altra cosa que la diferència entre el rendiment d'un model de predicció (en aquest cas una xarxa neuronal) en les dades que ha rebut com a entrenament i el seu rendiment en dades noves i desconegudes. Noti's que, en termes de la notació emprada a la secció 3.2, l'error de generalització ε_{gen} no és altra cosa que $L_{\mathcal{D}}(f_W) - L_S(f_W)$.

El següent resultat relaciona entre si els conceptes d'amplitud Gaussiana i d'error de generalització.

Teorema 55: *Sigui $\mathcal{H}$ una classe d'hipòtesis, sigui $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ un conjunt de domini, sigui $\ell : \mathcal{H} \times \mathcal{Z} \to [0, 1]$ una funció de pèrdua. Aleshores, per a qualsevol $\delta > 0$, amb probabilitat major o igual a $1 - \delta$ sobre l'elecció i.i.d. d'una mostra S de m elements de $\mathcal{Z}$, per a qualsevol $g \in \mathcal{H}$ es compleix:*

$$\mathbb{E}[g(\mathbf{z})] \leq \frac{1}{m}\sum_{i=1}^{m} g(\mathbf{z}_i) + \sqrt{2\pi}\widehat{\mathcal{G}}_S(\mathcal{H}) + \sqrt{\frac{9\log(\frac{2}{\delta})}{2m}}.$$

Passant al costat esquerra de la desigualtat el terme del sumatori s'obté:

$$\varepsilon_{gen} = \mathbb{E}[g(\mathbf{z})] - \frac{1}{m}\sum_{i=1}^{m} g(\mathbf{z}_i) \leq \sqrt{2\pi}\widehat{\mathcal{G}}_S(\mathcal{H}) + \sqrt{\frac{9\log(\frac{2}{\delta})}{2m}}.$$

On s'observa la relació prèviament mencionada entre l'error de generalització i l'amplitud Gaussiana.

Ara que ja s'han introduït tots els conceptes necessaris per abordar el resultat que es preté discutir, ja es pot fer la presentació del mateix:

Teorema 56: *Sigui $\mathcal{F}_L$ la classe de les xarxes neuronals amb L capes, rang i norma espectral acotades, amb una funció d'activació lineal per troços 1-Lipschitz ϕ.*

$$\mathcal{F}_L := \{f_{\mathbf{W}}(\mathbf{x}) = \mathbf{W}_L\phi(\mathbf{W}_{L-1}\phi(\dots\phi(\mathbf{W}_1\mathbf{x})\dots)) \mid ||\mathbf{W}_i||_2 \leq B_i \ , \ rang(\mathbf{W}_i) \leq r_i\}.$$

L'amplada de la capa i és n_i, per tant les dimensions de $\boldsymbol{W}_i$ són $n_i \times n_{i-1}$ amb $n_0 = d$. Sigui $\boldsymbol{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m]^T \in \mathbb{R}^{m \times d}$ la matriu corresponent a la mostra de dades S. Assumeixi's que les dades estan totes acotades, amb $max_i\ ||\boldsymbol{x}_i|| \leq R$. La complexitat Gaussiana de $\mathcal{F}_L$ es pot acotar superiorment com:

$$\widehat{\mathcal{G}}_S(\mathcal{F}_L) \lesssim \frac{||\mathbf{X}||_F}{m} \left\{ (||\mathbf{W}_1||_F \sqrt{n_1}) \prod_{i=2}^{L} C_1 ||\mathbf{W}_i||_2 + \sum_{i=2}^{L} C_1^{L-i} C_2 2\sqrt{2\tau_i}\kappa_i ||\mathbf{W}_i||_F \sqrt{n_i} \right\},$$

on C_1, C_1^{L-i} i C_2 són constants, κ_i α $\left(\prod_{j=1, j\neq i}^{d} ||\,\boldsymbol{W}_j||_2\right)$ i $\tau_i = min_{j \leq i}(rang(\,\boldsymbol{W}_j))$.

A més a més, també es poden acotar encara més les normes de Frobenius del teorema anterior de la forma $||\mathbf{W}_i||_F \leq \sqrt{\text{rang}(\mathbf{W}_i)}\ ||\mathbf{W}_i||_2 \leq \sqrt{r} B_i$; adicionalment, prenent $n = \max_i(n_i)$ la màxima amplitud d'entre totes, es pot simplificar la cota de la complexitat Gaussiana com:

$$\mathcal{O}\left(C L_1 \prod_{i=1}^{L} B_i R L r \sqrt{\frac{nh}{m}} \right).$$

No es reproduirà la demostració d'aquest teorema, si el lector desitja saber com es demostra aquest resultat, pot consultar-ho a l'article *Proceedings of Machine Learning Research, 30* [2], entre les pàgines 8 i 9 (Si bé cal consultar també les pàgines de la 4 a la 7 per tenir constància de tots els resultats previs i conceptes emprats).

Observació: Aquest resultat és cert per a problemes de classificació general; és a dir, no noecessàriament classificació binària. No obstant, també és cert per al cas concret de classificació binària, el qual correspon a l'escenari on la capa de sortida està composta per una sola neurona que retorna com a informació de sortida un 0 o un 1. Per tant, a la pròxima secció es restrigirà aquest resultat al cas de classificació binària per tal de poder comparar-lo amb els resultats teòrics presentats al llarg dels capítols 2, 3 i 4.

5.3 Discussió del resultat

Així doncs, el resultat anterior proveeix la següent cota per l'amplitud Gaussiana:

$$\widehat{\mathcal{G}}_S(\mathcal{F}_L) \lesssim C L_1 \prod_{i=1}^{L} B_i R L r \sqrt{\frac{n}{m}}.$$

Recordant la relació que s'havia vist anteriorment entre l'error de generalització ε_{gen} i l'amplitud Gaussiana $\widehat{\mathcal{G}}_S$:

$$\varepsilon \leq \sqrt{2\pi}\widehat{\mathcal{G}}_S(\mathcal{H}) + \sqrt{\frac{9\log(\frac{2}{\delta})}{2m}}.$$

Així doncs, ajuntant-ho amb aquest resultat, el teorema anterior ens proporciona la següent cota per a l'error de generalitzcaió ε_{gen}:

$$\varepsilon_{gen} \leq \sqrt{2\pi} C_1^L \prod_{i=1}^{L} B_i R L r \sqrt{\frac{n}{m}} + \sqrt{\frac{9\log(\frac{2}{\delta})}{2m}}.$$

La qual, agrupant els termes $\sqrt{2\pi}C_1^L \prod_{i=1}^{L} B_i R$ en una sola constant K, s'obté:

$$\varepsilon_{gen} \leq KLr\sqrt{\frac{n}{m}} + \sqrt{\frac{9\log(\frac{2}{\delta})}{2m}} = \sqrt{K'\frac{L^2r^2n}{m}} + \sqrt{\frac{9\log(\frac{2}{\delta})}{2m}}. \tag{5.3.1}$$

Si hom recorda la versió quantitativa del teorema fonamental de l'aprenentatge estadístic (teorema 48), aquest diu que donada una classe d'hipòtesis $\mathcal{H}$ d'un conjunt de domini $\mathcal{Z}$ a $\{0, 1\}$, prenent la funció de pèrdua $0-1$ i assumint que $\text{VCdim}(\mathcal{H}) = d < \inf$; aleshores existeixen unes constants absolutes C_1, C_2 tals que:

1. *$\mathcal{H}$ té la propietat de convergència uniforme amb una complexitat de la mostra tal que:*
$$C_1\frac{d+\log(1/\delta)}{\varepsilon^2} \leq m_{\mathcal{H}}^{UC}(\varepsilon,\delta) \leq C_2\frac{d+\log(1/\delta)}{\varepsilon^2}.$$
2. *$\mathcal{H}$ és aprensible PAC agnòsticament amb una complexitat de la mostra tal que:*
$$C_1\frac{d+\log(1/\delta)}{\varepsilon^2} \leq m_{\mathcal{H}}(\varepsilon,\delta) \leq C_2\frac{d+\log(1/\delta)}{\varepsilon^2}.$$
3. *$\mathcal{H}$ és aprensible PAC amb una complexitat de la mostra tal que:*
$$C_1\frac{d+\log(1/\delta)}{\varepsilon} \leq m_{\mathcal{H}}(\varepsilon,\delta) \leq C_2\frac{d\log(1/\varepsilon)+\log(1/\delta)}{\varepsilon}.$$

Parant atenció exclusivament a les cotes superiors i aïllant l'error enlloc de la complexitat de la mostra s'obtenen les següents cotes superiors:

1. Convergència uniforme:
$$\varepsilon \leq \sqrt{C_2\frac{d+\log(1/\delta)}{m_{\mathcal{H}}^{UC}}}.$$
2. Aprensibilitat PAC agnòstica:
$$\varepsilon \leq \sqrt{C_2\frac{d+\log(1/\delta)}{m_{\mathcal{H}}}}.$$
3. Aprensibilitat PAC:
$$\varepsilon \leq C_2\frac{d\log(1/\varepsilon)+\log(1/\delta)}{m_{\mathcal{H}}}.$$

L'objectiu és comparar aquestes cotes amb (5.3.1). En primer lloc, analitzant la forma de (5.3.1), es pot observar que el producte Lrn denota el nombre màxim de paràmetres lliures que pot tenir la xarxa neuronal que s'està estudiant; per tant, la presència de $\sqrt{L^2r^2n}$ a la cota indica que existeix una dependència de la cota de l'error respecte el nombre de paràmetres lliures del model. Anàlogament, analitzant ara les cotes proporcionades pel TFAE, totes elles presenten una relació entre l'error ε i la dimensió VC d, la qual s'assumeix dins la comunitat de l'aprenetatge automàtic que guarda una relació amb el nombre de paràmetres lliures. Així doncs, les cotes obtingudes per ambdues vies manifesten una dependència respecte el nombre de paràmetres lliures, malgrat el tipus de relació no sigui el mateix. Observem que en els casos de la convergència uniforme i de l'aprensibilitat PAC agnòstica la cota de l'error és inversament proporcional a l'arrel de la mida de la mostra, de la mateixa manera que ho és en la cota proporcionada pel teorema de l'article. A més

a més, malgrat una diferència en les constants, també mostren una dependència respecte l'arrel del logaritme de l'invers del paràmetre de confiança δ.

Per tant, si bé les cotes proporcionades pel TFAE i pel teorema de l'article s'han obtingut per vies diferents, malgrat no siguin ben bé iguals tenen bastantes semblances entre si en la seva forma i implicacions.

De fet, fent uns petits ajustos a ambdues cotes, aquesta semblança i relació entre d i una combinació de L, r i n queda més evidenciada. En primer lloc, aplicant la desigualtat triangular a la cota de l'aprensibilitat PAC agnòstica, i prenent $C' = 2C$, es té:

$$\varepsilon \leq \sqrt{C\frac{d+\log(\frac{1}{\delta})}{m}} \leq \sqrt{C\frac{d+\log(\frac{2}{\delta})}{m}} \leq \sqrt{C'\frac{d}{2m}} + \sqrt{C'\frac{\log(\frac{2}{\delta})}{2m}}. \tag{5.3.2}$$

Per altra banda, definint $K^* := \max(K', 9)$, es pot reescriure la cota del resultat de l'article com:

$$\varepsilon \leq \sqrt{K'\frac{L^2r^2h}{2m}} + \sqrt{\frac{9\log(\frac{2}{\delta})}{2m}} \leq \sqrt{K^*\frac{L^2r^2h}{2m}} + \sqrt{\frac{K^*\log(\frac{2}{\delta})}{2m}}. \tag{5.3.3}$$

Es pot observar que ambdues noves cotes tenen la mateixa forma; ambdues estan composades per la suma de dos termes de forma idèntica, essent el segon exactament igual llevat d'una constant. Així doncs, fent èmfasi en el primer terme d'ambdues cotes, es pot observar que tan sols difereixen en la constant i en el paràmetre del numerador. A la cota (5.3.2), l'error està acotat segons $\sqrt{d}$, mentre que a la (5.3.3) l'error està acotat segons $\sqrt{L^2r^2n}$. Així doncs, es posa de manifest que existeix una relació de proporcionalitat entre d i L^2r^2h. De fet, com que Lrh denota el nombre màxim de paràmtres lliures del model d'una xarxa neuronal, vist que la dimensió VC està relacionada amb una variació del producte d'aquests tres elements, es pot deduir que la dimensió VC guarda relació amb el nombre de paràmetres lliures del model considerat.

Addicionalment, aquest resultat és de gran interés per raons de pragmatisme. En primer lloc, la dimensió VC d és un paràmetre que depén implícitamnet del model (ja que depén de la classe d'hipòtesis $\mathcal{H}$ que s'hagi escollit) però també depén de les dades que s'estiguin tractant. Això es deu que, per la definició d'una classe de funcions separant un conjunt de punts, per tal de calcular la dimensió VC d'una classe de funcions sobre un conjunt, cal proporcionar un conjunt de punts concret pels quals no es compleixi. A més a més, la dimensió VC sovint és molt difícil de calcular, provocant que qualsevol cota que en depengui sigui molt complicada de determinar amb certa exactitut. Per altra banda, els termes n, r i L depenen exclussivament de la xarxa neuronal que s'estigui considerant, així doncs no hi ha cap dependència respecte les dades. Es podria argumentar que, ja que les matrius de pesos es van actualitzant segons les dades que va processant la xarxa, el rang de les matrius $\mathbf{W}_i$ depén en certa manera de les dades i, per tant, també ho fa r. No obstant, el rang d'una $\mathbf{W}_i$ està acotat pel rang màxim r d'entre totes les matrius de pesos, el qual alhora està acotat, en el pitjor dels casos, pel terme r' definit com:

$$r' = \sup_{i\in[L]} \left(\min(n_{i-1}, n_i)\right).$$

El qual ve fixat exclussivament per les mides de les capes i, per tant, per les característiques de la xarxa neuronal. Així doncs, aquests tres termes són independents de les

dades i, per tant, també ho és la cota. A més a més, al ser paràmetres que venen fixats per les característiques de la xarxa que s'està emprant, no suposa cap dificultat computar-los; per tant, suposen un gran avantatge respecte la dimensió VC en termes pragmàtics a l'hora de calcular la cota.

Finalment, convé destacar que la relació de d amb L^2r^2h denota que la complexitat de l'espai d'hipòtesis considerat (en el cas d'aquesta memòria, la classe d'hipòtesis $\mathcal{H}$) està relacionada amb el nombre de paràmetres lliures del model considerat.

Capítol 6

Conclusions

Al llarg d'aquesta memòria s'han introduit i desenvolupat tot un seguit de conceptes i resultats matemàtics de l'àmbit de l'aprenentatge automàtic posant l'èmfasi en els problemes de classificació binària. En el transcurs, s'han obtingut un seguit de resultats amb implicacions molt potents.

En primer lloc, arran del teorema NFL s'ha constatat que no existeix cap aprenent universal; és a dir, en absència de coneixement previ sobre el problema a resoldre, per a cada aprenent concret (algoritme d'aprenentatge), existeix un problema (distribució de probabilitat sobre el producte del conjunt de domini pel conjunt de classificacions) tal que per mostres d'entrenament d'una mida que no sigui gaire gran (menor a la meitat de la cardinalitat del domini) aleshores l'aprenent fracassarà (retornarà una funció de predicció amb paràmetres de precissió i confiança massa grans; en aquest cas 1/8 i 1/7 respectivament). Tot això malgrat pel mateix problema existeixi un altre aprenent que si que sigui capaç de resoldre'l satisfactòriament sense necessitat de coneixement previ. Per tant, la primera conclusió extreta en aquesta memòria és la inexistència d'un aprenent universal.

En segon lloc, com a conseqüència de la versió qualitativa del teorema fonamental de l'aprenentatge estadístic s'ha posat de manifest que la finitut de la dimensió VC d'una classe d'hipòtesis $\mathcal{H}$ és condició necessària i suficient per tal que la mateixa sigui aprensible PAC, aprensible PAC agnòsticament i tingui la propietat de convergència uniforme. Així doncs, s'ha caracteritzat la condició necessària i suficient per tal de garantir les nocions d'aprensibilitat que s'han estudiat al llarg de la memòria.

En tercer lloc, analitzant un resultat recent de recerca en l'àmbit de l'aprenentatge profund mitjançant xarxes neuronals, el qual proporciona una cota a l'error de generalització d'un model d'aprenetatge en funció dels paràmetres d'una xarxa neuronal, s'ha observat que la cota proporcionada a l'article en qüestió és coherent i es comporta de manera raonablement similar a les cotes que proporciona la versió quantitativa del teorema fonamental de l'aprenentatge estadístic.

Finalment, com a línia a seguir després d'aquest treball, la continuació natural d'aquest estudi és abordar la qüestió de la minimització del risc estructural ("Structural Risk Minimitzation"o SRM), la qual fa referència a establir preferències dins de les hipòtesis d'una classe $\mathcal{H}$.

Bibliografia

[1] Shai Shalev-Schwartz and Shai Ben-David. (2014). *Understanding machine learning: From theory to algorithms.* Cambridge University press.

[2] Pinto, Andrea; Rangamani, Akshay and Poggio, Tomaso. (2024). On Generalitzation Bounds for Neural Networks with Low Rank Layers. *Proceedings of Machine Learning Research, 30, 1-16.*

[3] Hu, Edward; Shen, Yelong; Wallis, Phillip; Allen-Zhu, Zeyuan; Li, Yuanzhi; Wang, Shean; Wang, Lu; Chen, Weizhu. (2021) LoRA: Low-Rank Adaptation of Large Language Models. arXiv preprint: Arxiv-2106.09685.

[4] Gonen, Alon. (2014). *Understanding Machine Learing: Solution Manual*

[5] Yu, Huili. (5 d'abril de 2023). Weight Initialization for Deep Neural Network. A *Medium*. https://medium.com/@freshtechyy/weight-initialization-for-deep-neural-network-e0302b6f5bf3

[6] Hardesty, Larry. (14 d'abril de 2017). Explained: Neural networks. A *MIT News on campus and around the world*. https://news.mit.edu/2017/explained-neural-networks-deep-learning-0414

[7] Rallabandi, Srikari. (26 de març de 2023). Activation functions: ReLU vs Leaky ReLU. A *Medium*. https://medium.com/@sreeku.ralla/activation-functions-relu-vs-leaky-relu-b8272dc0b1be

[8] *Lecture 12: Hoeffding's Bound*. ([s.d.]). Recuperat de chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/https://www.cs.purdue.edu/homes/hmaji/teaching/Spring%202019/lectures/12.pdf

[9] Jensen, Johan Ludwig William Valdemar (1906). "Sur les fonctions convexes et les inégalités entre les valeurs moyennes". Acta Mathematica. 30 (1): 175–193. doi:10.1007/BF02418571.

[10] AM-GM Inequality. (2 de desembre de 2024). A la Viquipèdia. https://en.wikipedia.org/wiki/AM%E2%80%93GM_inequality

[11] McCulloch, W.S., Pitts, W. A logical calculus of the ideas immanent in nervous activity. Bulletin of Mathematical Biophysics 5, 115–133 (1943). https://doi.org/10.1007/BF02478259