Teorija statističkog učenja i primjene

Babić, Marija

Teorija statističkog učenja i primjene

Babić, Marija

SVEUČILIŠTE U ZAGREBU

PRIRODOSLOVNO–MATEMATIČKI FAKULTET

MATEMATIČKI ODSJEK

Marija Babić

TEORIJA STATISTIČKOG UČENJA I PRIMJENE

Diplomski rad

Voditelj rada:
prof.dr.sc. Bojan Basrak
dr.sc. Hrvoje Planinić

Zagreb, srpanj 2021.

Ovaj diplomski rad obranjen je dana _____ pred ispitnim povjerenstvom u sastavu:

1. _____, predsjednik
2. _____, član
3. _____, član

Povjerenstvo je rad ocijenilo ocjenom _____.

Potpisi članova povjerenstva:

1. _____
2. _____
3. _____

Sadržaj

Uvod

Zbog velike mogućnosti prilagodbe postojećih modela novim problemima i pružanja brzih, efikasnih i robusnih rješenja, područje strojnog učenja je u stalnom razvoju i prepoznata je njegova važnost. Strojevi uče, pronalaze uzorke i donose odluke zahtijevajući minimalnu ljudsku intervenciju te na taj način industrije koje rade s velikim količinama podataka bilježe brži rad s boljim rezultatima.
Tehnologije strojnog učenja danas se koriste u financijskoj industriji u zaštiti osobnih podataka i detekciji prevare u kartičnim transakcijama; vladine agencije ih koriste za zaštitu osobnih podataka građana; u zdravstvu se koriste za praćenje pulsa, količine kisika i šećera te u poboljšanju dijagnoza i prilagodbi tretmana; u naftnoj i plinskoj industriji za analizu minerala u zemlji te predikcijama kvarova u rafinerijama; u prometu za stvaranje prijevoznih ruta dostavnih službi; u filtriranju e-mailova, detekciji lica na kamerama, prepoznavanju zvučnih naredbi na pametnim telefonima...
Međutim, strojno učenje nije matematički precizno definirano područje, već kolekcija ideja, tehnika i primjena. Iako su statistika i strojno učenje dva odvojena područja, njihovom kombinacijom nastaje statističko učenje. Dok statistika proučava odnos među varijablama, a strojno učenje radi na optimizaciji i poboljšanju točnosti, statističko učenje usmjerava se na razumijevanje podataka i stvaranje prediktivnih statističkih modela.
U ovom radu uvest ćemo osnovne pojmove vezane za modele učenja na podacima i na matematički rigorozan način definirati glavne koncepte statističkog učenja koje ćemo zatim implementirani u neke od algoritama koji se danas uspješno primjenjuju u praksi.
Na kraju uvoda dajemo kratak pregled rada po poglavljima.
Na početku Poglavlja 1 definiramo osnovne dijelove modela za učenje, a to su domena, kodomena i skup za učenje te rezultat učenja, odnosno hipotezu. Nakon što smo naveli sve potrebne pretpostavke, uvodimo dvije vrste grešaka, pravu i empirijsku grešku. Definiramo najjednostavniju vrstu učenja koja se naziva minimizacija empirijskog rizika ili kraće ERM i na primjeru ilustriramo kako ta vrsta učenja ponekad ne daje niti približno točno rješenje. U posljednjem dijelu bavimo se traženjem uvjeta koji će nam garantirati da ERM algoritam daje dobro pravilo predviđanja, a uz uvođenje RA pretpostavke kojom se pretpostavlja da u promatranoj klasi hipoteza postoji hipoteza čija je prava greška jednaka 0, kao jedno rješenje prikazujemo primjenu ERM algoritma na ograničen (konačan) una-

prijed odabran skup hipoteza koji se naziva klasa hipoteza.
U Poglavlju 2 upoznajemo se sa formalnim modelom za učenje, PAC modelom, koji je baziran na RA pretpostavci. Nakon definiranja parametra točnosti, parametra pouzdanosti i složenosti učenja, dajemo primjere klasa hipoteza koje je moguće naučiti u PAC smislu. Uvođenjem pojma agnostičkog PAC učenja koje ne zahtijeva RA pretpostavku te pojma generalizirane funkcije gubitka, poopćili smo PAC model kako bi se mogao koristiti na većoj skupini zadataka za učenje.
U Poglavlju 3 dokazujemo No Free Lunch teorem koji nam govori da ako nemamo nikakvih dodatnih pretpostavki na klasu hipoteza, ne postoji algoritam kojim možemo uspješno riješiti sve probleme učenja. Nadalje pokazujemo da je pri biranju klase hipoteza važno pronaći ravnotežu između greške aproksimacije i greške procjene, što je u literaturi poznato kao *bias-complexity tradeoff*. Naime, ako klasa hipoteza sadrži veliki broj hipoteza može doći do *overfittinga*, dok restringiranje klase na mali skup hipoteza može dovesti do *underfittinga*. Na kraju poglavlja definiramo Vapnik-Chervonenkisovu dimenziju, ili kraće VC dimenziju, čiju konačnost povezujemo s pojmom PAC učenja te za izabrane primjere klasa hipoteza računamo vrijednost njihove VC dimenzije.
U Poglavlju 4 definiramo pojam uniformne konvergencije za općenitu funkciju gubitka. Pokazujemo da svaku klasu koja zadovoljava svojstvo uniformne konvergencije možemo naučiti u PAC smislu te da konačne klase hipoteza čine podskup klasa koje imaju svojstvo uniformne konvergencije.
U Poglavlju 5 bavimo se glavnim teoremom ovog rada, Fundamentalnim teoremom statističkog učenja, kojeg dokazujemo korištenjem Sauer-Shelah-Perles leme i Massartove leme. Njime smo pokazali ekvivalenciju između posjedovanja svojstva uniformne konvergencije, mogućnosti učenja u (agnostičkom) PAC smislu (uz korištenje ERM algoritma) te konačnosti VC dimenzije za određenu klasu hipoteza za probleme binarne klasifikacije. Navodimo i kvantitativnu verziju teorema koja nam daje gornje i donje ograde za složenost učenja koje, između ostalog, ovise o VC dimenziji promatrane klase hipoteza.
U Poglavlju 6 promatramo klasu poluprostora koja spada u jednu od najkorištenijih familija klasa hipoteza, linearne prediktore, te korištenjem Radonovog teorema dokazujemo konačnost te klase. Upoznajemo se sa problemima linearnog programiranja poluprostora i dajemo implementaciju ERM algoritma u obliku iterativnog algoritma, Perceptron ili PLA algoritma, za koji nalazimo gornju ogradu na broj iteracija.
U posljednjem poglavlju fokusirali smo se na jedan od Boosting algoritama, a to je AdaBoost algoritam. Svi Boosting algoritmi kreću od nekih loših prediktora, stoga definiramo pojam γ-slabog učenja koje vraća hipotezu samo malo bolju od slučajnog pogađanja. AdaBoost je iterativni algoritam koji kao rezultat daje hipotezu koja ovisi o linearnoj kombinaciji nekih jednostavnih hipoteza, a njena empirijska greška opada rastom broja iteracija. Ilustracija algoritma dana je jednostavnim primjerom.

Poglavlje 1

Minimizacija empirijskog rizika

Pretpostavimo da imamo problem filtriranja e-mailova, odnosno želimo odrediti spada li neki e-mail u neželjenu poštu ili ne. Potreban nam je skup primjera, odnosno e-mailovi koji su već analizirani te raspoređeni u dvije kategorije: normalna i neželjena pošta. Nakon toga želimo analizirati ove dvije kategorije te stvoriti algoritam koji će na temelju karakteristika svake kategorije biti u mogućnosti s velikom vjerojatnosti prepoznati neželjenu poštu među novopridošlim e-mailovima. Ovaj primjer spada u probleme statističkog učenja, a pomoću njega možemo definirati formalni model i njegove komponente.

1.1 Elementi modela za učenje

Osnovna pretpostavka statističkog učenja je da znamo kako izgledaju domena i skup oznaka te da imamo pristup skupu za učenje.

- Skup objekata nad kojma će biti izvedeno učenje nazivamo *domena* i označavamo ga s $\mathcal{X}$, a njegove elemente nazivamo *primjerima*. Elementi domene mogu biti pojedinačne vrijednosti ili mogu biti reprezentirani vektorom značajki. U primjeru iz uvodnog dijela domena bi bila skup svih e-mailova.
- Skup $\mathcal{Y}$ predstavlja skup oznaka elemenata domene. Najčešće će biti ograničen na dva slučaja, a to su {0,1} i {-1, 1}. Susrest ćemo se i sa slučajevima učenja u kojima će $\mathcal{Y}$ poprimiti druge vrijednosti, kao na primjer u regresijskom modelu gdje je $\mathcal{Y}$ skup $\mathbb{R}$. U primjeru s e-mailovima, $\mathcal{Y}$ će biti skup {0,1}. Oznaku 0 dobivaju e-mailovi koji spadaju u neželjenu poštu, a ostali dobivaju oznaku 1.

- Skup $S = \{(x_1, y_1)...(x_m, y_m)\}$ koji predstavlja konačan niz uređenih parova iz $\mathcal{X} \times \mathcal{Y}$ nazivamo *skup za učenje*. To je skup označenih elemenata domene koji nam pomaže u pronalasku pravila označavanja primjera iz $\mathcal{X}$. U primjeru s e-mailovima, u skup S bi spadali e-mailovi koji su već analizirani i razdvojeni u dvije kategorije.

Rezultat učenja je pravilo predviđanja, $h : \mathcal{X} \to \mathcal{Y}$. Ova funkcija se najčešće naziva *klasifikator* , *hipoteza* ili *prediktor*, a koristi se za predviđanje oznaka novih točaka domene. Sa $A(S)$ označavamo hipotezu koju algoritam za učenje A vraća nakon dobivanja skupa za učenje S.

Pretpostavit ćemo da su elementi domene $\mathcal{X}$ generirani nekom vjerojatnosnom distribucijom $\mathcal{D}$. To može biti bilo koja vjerojatnosna distribucija i nama je nepoznata. Također pretpostavljamo da postoji funkcija $f : \mathcal{X} \to \mathcal{Y}$ koja svakoj točki domene pridružuje točnu oznaku. Funkcija f je nepoznata, a cilj nam je odrediti ju ili barem što bolje aproksimirati.

Oznaku $x \sim \mathcal{D}$ koristimo kako bi označili da je x generiran distribucijom $\mathcal{D}$ i označen funkcijom f. Za skup za učenje $S = \{(x_1, y_1), ..., (x_m, y_m)\}$ definiramo skup $S|_x = \{x_1, ..., x_m\}$. Uobičajeno je pretpostaviti da su elementi od $S|_x$ međusobno nezavisni i jednako distribuirani (kraće n.j.d.). To znači da je svaki $x_i \in S|_x$ generiran distribucijom $\mathcal{D}$ i označen funkcijom f, tj. $y_i = f(x_i)$. Ovu pretpostavku označavamo s $S|_x \sim \mathcal{D}^m$ gdje je m veličina skupa S, a $\mathcal{D}^m$ označava da smo m puta generirali element iz distribucije $\mathcal{D}$ nezavisno jedan od drugoga. Ubuduće ćemo uvijek pretpostavljati da su elementi skupa $S|_x$ generirani nezavisno jedan od drugog, a ukoliko to u nekom slučaju nije nužno, bit će posebno naglašeno.

Definicija 1.1.1. *Neka je x slučajna točka domene $\mathcal{X}$ generirana distribucijom $\mathcal{D}$, $\mathcal{Y}$ skup oznaka te $f : \mathcal{X} \to \mathcal{Y}$ funkcija točnih oznaka. Greška klasifikatora h ili prava greška od h, u oznaci $L_{\mathcal{D},f}(h)$, je vjerojatnost da će vrijednost od $h(x)$ biti različita od točne vrijednosti $f(x)$. Odnosno,*

$$L_{\mathcal{D},f}(h) = \mathbb{P}_{x \sim \mathcal{D}} \left(h(x) \neq f(x) \right) . \tag{1.1}$$

Drugim riječima, prava greška klasifikatora h je vjerojatnost da h neće predvidjeti točnu oznaku slučajne točke x generirane distribucijom $\mathcal{D}$.

1.2 Minimizacija empirijskog rizika

Algoritam za učenje dobije skup za učenje S koji je generiran nepoznatom distribucijom $\mathcal{D}$ i označen nekom nepoznatom funkcijom f te nam vraća funkciju $h_S : \mathcal{X} \to \mathcal{Y}$ (pišemo h_S umjesto h zato što ta funkcija ovisi o skupu za učenje S). Cilj algoritma je pronaći funkciju h_S koja minimizira pravu grešku. Međutim, distribucija $\mathcal{D}$ i funkcija f nam nisu poznate,

pa nismo u mogućnosti odrediti vrijednost prave greške. Zbog toga ćemo promatrati drugu vrstu greške - grešku na skupu za učenje, koju nazivamo *empirijska greška klasifikatora h*:

$$L_S(h) = \frac{1}{m} \sum_{i=1}^{m} \mathbb{1}_{\{h(x_i) \neq y_i\}}\,, \tag{1.2}$$

gdje je $[m] = \{1, ..., m\}$, a $S = \{(x_1, y_1), ..., (x_m, y_m)\}$ skup za učenje.

$L_S(h)$ zapravo označava postotak krivo označenih primjera skupa S. Kako je S jedini dio domene čije oznake su nam dostupne, ima smisla tražiti ono rješenje koje je zadovoljavajuće na tom skupu. Vrsta učenja kojom tražimo prediktor h_S koji minimizira $L_S(h)$ naziva se ***minimizacija empirijskog rizika*** (engl. *empirical risk minimization*) ili kraće, ERM.
Ovakav način određivanja prediktora h čini se dosta prirodan, ali postoje situacije u kojima ERM algoritam ne mora dati niti približno točno rješenje. Jednu takvu situaciju demonstrirat ćemo na sljedećem primjeru.

Primjer 1.2.1. *Neka je $S = \{(x_1, y_1), ..., (x_m, y_m)\}$ fiksan skup za učenje takav da elementi od $S|_x$ nisu nužno generirani nezavisno jedan od drugog, $f : \mathcal{X} \to \{0, 1\}$ nepoznata funkcija oznaka te prediktor h_S zadan s*

$$h_S(x) = \begin{cases} y_i & \text{ako } \exists\, i \in [m] \text{ t.d. } x_i = x \\ 0 & \text{inače}\,. \end{cases} \tag{1.3}$$

Bez obzira na to koji uzorak S imamo, vrijedi da je $L_S(h_S) = 0$. Kako prediktor h_S minimizira vrijednost empirijske greške, slijedi da može biti odabran ERM algoritmom. Po definiciji prave greške i prediktora h_S imamo

$$L_{\mathcal{D},f}(h_S) = \mathop{\mathbb{P}}_{x\sim\mathcal{D}}(h_S(x) \neq f(x)) = \mathop{\mathbb{P}}_{x\sim\mathcal{D}}(f(x) = 1, x \notin S|_x)\,.$$

Naime, prediktor h_S točkama koje nisu elementi skupa $S|_x$ dodjeljuje oznaku 0, a točkama skupa $S|_x$ dodjeljuje identičnu oznaku kao f. Iz toga slijedi da će se oznake od h_S i f razlikovati jedino na točkama domene koje nisu elementi od $S|_x$, a točna oznaka im je 1.
Ako uzmemo da je $\mathcal{X} \subset \mathbb{R}^k$ i da je $\mathcal{D}$ neka neprekidna distribucija, slijedi

$$L_{\mathcal{D},f}(h_S) = \mathop{\mathbb{P}}_{x\sim\mathcal{D}}(f(x) = 1, x \notin S|_x) = \mathop{\mathbb{P}}_{x\sim\mathcal{D}}(f(x) = 1) = p\,.$$

Cilj nam je pokazati da za prozvoljan fiksan $p \in [0, 1]$ možemo izabrati $\mathcal{D}$ i f tako da vrijedi $\mathbb{P}(L_{\mathcal{D},f}(h_S) = p) = 1$. Neka je onda $\mathcal{X} \subset \mathbb{R}^k$ proizvoljan kvadrat površine 1, $\mathcal{Y} \subseteq \mathcal{X}$ proizvoljan kvadrat površine p, $\mathcal{D}$ uniformna distribucija na $\mathcal{X}$ te funkcija f takva da je $f(x) = 1$ za $x \in \mathcal{Y}$ i $f(x) = 0$ za $x \notin \mathcal{Y}$. Tada vrijedi

$$\mathop{\mathbb{P}}_{x\sim\mathcal{D}}(f(x) = 1) = \mathop{\mathbb{P}}_{x\sim\mathcal{D}}(x \in \mathcal{Y})\,.$$

Iz konstrukcije funkcije f i distribucije $\mathcal{D}$ slijedi

$$\mathbb{P}_{S|_x \sim \mathcal{D}^m}(L_{\mathcal{D},f}(h_S) = p) = 1\,.$$

Na početku smo pokazali da je $L_S(h_S) = 0$. Ukoliko vrijedi da je $p = 1$, dobijemo da je $L_{\mathcal{D},f}(h_S) = 1$ gotovo sigurno. Odnosno, gotovo sigurno ćemo u potpunosti pogriješiti oznaku. Iz ovog vidimo da smo pronašli prediktor koji odlično predviđa oznake na skupu za učenje, ali na ostalim elementima domene predviđanja nisu tako dobra. Ova pojava se naziva overfitting*, a najčešće se pojavljuje kad prediktor "predobro" predviđa oznake skupa za učenje.*

1.3 Minimizacija empirijskog rizika restringiranjem na klasu $\mathcal{H}$

U prošlom primjeru pokazali smo kako korištenje ERM algoritma može dovesti do *overfittinga*, stoga želimo pronaći uvjete koji će nam garantirati da će algoritam pronaći dobro rješenje i na skupu za učenje i na cjelokupnoj domeni. Jedno od čestih rješenja je primjena ERM algoritma na ograničen skup funkcija koji treba biti unaprijed odabran. Taj skup nazivamo *klasa hipoteza* i označavamo ga s $\mathcal{H}$. Svaka funkcija $h \in \mathcal{H}$ je preslikavanje sa $\mathcal{X}$ u $\mathcal{Y}$. Za svaku klasu $\mathcal{H}$ i za svaki skup za učenje S, korištenjem ERM algoritma dolazimo do prediktora $ERM_{\mathcal{H}}(S)$ koji minimizira empirijsku grešku:

$$ERM_{\mathcal{H}}(S) \in \operatorname*{argmin}_{h \in \mathcal{H}} L_S(h)\,, \tag{1.4}$$

gdje funkcija *argmin* vraća one hipoteze $h \in \mathcal{H}$ u kojima $L_S(h)$ postiže minimum nad $\mathcal{H}$. U nekim slučajevima u *argmin* može biti više prediktora. Za svaki od njih možemo konstruirati različite algoritme za koje vrijedi da kao rezultat učenja daju baš taj prediktor. Kako svaki takav algoritam ima svojstvo da daje hipotezu koja minimizira empirijsku grešku, možemo ga zvati ERM algoritmom.

Restrikcija na klasu $\mathcal{H}$ određena je prije nego što smo saznali kako izgleda skup za učenje, što znači da treba biti bazirana na nekom prijašnjem znanju vezanom s problemom koji rješavamo. Pokazat ćemo da biranjem $ERM_{\mathcal{H}}(S)$ iz ove klase sigurno neće doći do *overfittinga*.

Konačne klase hipoteza

Najjednostavniji način restringiranja klase $\mathcal{H}$ je ograničenje broja prediktora u njoj. Pokazat ćemo da, ako je $\mathcal{H}$ konačna i ako je skup za učenje S dovoljno velik, prilikom

određivanja $ERM_{\mathcal{H}}(S)$ neće doći do overfittinga. Pri tome će broj primjera u skupu za učenje ovisiti o veličini klase $\mathcal{H}$.
Neka je $\mathcal{H}$ konačna klasa, S skup za učenje označen nekom funkcijom f te h_S rezultat primjene pravila $ERM_{\mathcal{H}}$ na skup S, tj.:

$$h_S \in \underset{h \in \mathcal{H}}{\operatorname{argmin}} L_S(h)\,. \tag{1.5}$$

Uvodimo dodatnu pretpostavku:

Definicija 1.3.1. *(The Realizability Assumption, RA pretpostavka) Postoji $h^* \in \mathcal{H}$ takav da je $L_{\mathcal{D},f}(h^*) = 0$.*

Zanima nas karakterizacija empirijske greške u slučaju kad vrijedi RA pretpostavka. Prvo što trebamo promotriti je očekivanje empirijske greške. Za skup $S = \{(x_1, y_1), ..., (x_m, y_m)\}$ pretpostavimo da su elementi od $S|_x$ generirani distribucijom $\mathcal{D}$ (ne nužno nezavisno jedan od drugog) i označeni vrijednostima funkcije f. Tada vrijedi:

$$\begin{aligned}
\underset{S|_x \sim \mathcal{D}^m}{\mathbb{E}}\left[L_S(h)\right] &= \underset{S|_x \sim \mathcal{D}^m}{\mathbb{E}}\left[\frac{1}{m}\sum_{i=1}^{m} \mathbb{1}_{\{h(x_i)\neq y_i\}}\right] \\
&= \frac{1}{m}\sum_{i=1}^{m} \underset{x_i \sim \mathcal{D}}{\mathbb{E}}\left[\mathbb{1}_{\{h(x_i)\neq y_i\}}\right] \\
&= \underset{x \sim \mathcal{D}}{\mathbb{E}}\left[\mathbb{1}_{\{h(x)\neq y_i\}}\right] \\
&= \underset{x \sim \mathcal{D}}{\mathbb{P}}\left(h(x) \neq f(x)\right) \\
&= L_{\mathcal{D},f}(h)\,.
\end{aligned}$$

Odnosno, očekivana vrijednost empirijske greške hipoteze h je prava greška hipoteze h. Ako za h uzmemo hipotezu h^* danu RA pretpostavkom, dobijemo

$$\underset{S|_x \sim \mathcal{D}^m}{\mathbb{E}}[L_S(h^*)] = L_{\mathcal{D},f}(h^*) = 0\,.$$

Po definiciji $L_S(h)$ poprima samo nenegativne vrijednosti. Znamo da za svaku slučajnu varijablu $X \geq 0$ za koju vrijedi da je $E[X] = 0$ slijedi $X = 0$, pa možemo zaključiti da vrijedi $L_S(h^*) = 0$ gotovo sigurno. Neka je h_S hipoteza koju vraća ERM algoritam. Kako vrijedi da je $h^* \in \mathcal{H}$ te $L_S(h_S) \leq L_S(h)$ za svaki $h \in \mathcal{H}$, slijedi

$$L_S(h_S) \leq L_S(h^*) = 0\,.$$

To znači da za svaku hipotezu koju vraća ERM algoritam vrijedi da je $L_S(h_S) = 0$. Međutim, nas zanima prava greška prediktora h_S, odnosno $L_{\mathcal{D},f}(h_S)$. Ako algoritam ima pristup samo

dijelu domene, tj. skupu za učenje S, onda njegova greška ovisi o načinu na koji su elementi skupa $S|_x$ generirani. Skup S nam daje djelomičan uvid u to na koji način funkcija f označava elemente domene. Što je S veći, to imamo više informacija o načinu preslikavanja.

Ne možemo očekivati da će nas skup S uvijek dovesti do dobrog prediktora zato što postoji vjerojatnost da taj skup nije reprezentativan, tj. postoji vjerojatnost da nam skup S ne daje realan prikaz karakteristika cijele domene i načina označavanja primjera, već samo nekog dijela domene. Naime, ako se vratimo na primjer s e-mailovima, vidimo da nam se može dogoditi da skup za učenje sadrži samo neželjenu poštu. Zbog toga je moguće da će algoritam koji uči na tom skupu svaki novi e-mail označiti kao neželjenu poštu. Vjerojatnost da uzorak za učenje neće biti reprezentativan označavamo s δ, a $(1-\delta)$ zovemo *parametrom pouzdanosti* procjene. Kako ne možemo garantirati da će se dobiveni prediktor h u potpunosti podudarati s funkcijom f, uvodimo *parametar preciznosti* koji označavamo s ε. Događaj $L_{\mathcal{D},f}(h_S) > \varepsilon$ interpretiramo kao neuspjeh, dok kod događaja $L_{\mathcal{D},f}(h_S) \leq \varepsilon$ prediktor h_S prihvaćamo kao približno točan prediktor.

Želimo odrediti gornju granicu na vjerojatnost odabira skupa S koji će dovesti do neuspjeha, tj. želimo odozgo ograničiti vjerojatnost

$$\mathop{\mathbb{P}}_{S|_x \sim \mathcal{D}^m} [L_{\mathcal{D},f}(h_S) > \varepsilon] .$$

Sa $\mathcal{H}_B$ označit ćemo skup svih loših hipoteza, odnosno

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

S M ćemo označiti skup svih uzoraka koji nas dovode do pogrešnog zaključka, tj. skup svih uzoraka čija je empirijska greška 0, a prava greška veća od ε:

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

Kako po RA pretpostavci vrijedi $L_S(h_S) = 0$ gotovo sigurno, događaj $L_{\mathcal{D},f}(h_S) > \varepsilon$ se može dogoditi samo ako vrijedi da je $L_S(h) = 0$ za neki $h \in \mathcal{H}_B$, odnosno ako je naš uzorak iz skupa M. Slijedi:

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

M možemo zapisati kao

$$M = \bigcup_{h \in \mathcal{H}_B} \{S|_x : L_S(h) = 0\} . \tag{1.6}$$

Kako je $\mathcal{H}$ konačan, pa stoga i prebrojiv, vrijedi

$$\mathop{\mathbb{P}}_{S|_x\sim\mathcal{D}^m}[L_{\mathcal{D},f}(h_S) > \varepsilon] \leq \mathop{\mathbb{P}}_{S|_x\sim\mathcal{D}^m}\left[\bigcup_{h\in\mathcal{H}_B}\{L_S(h) = 0\}\right] \leq \sum_{h\in\mathcal{H}_B}\mathop{\mathbb{P}}_{S|_x\sim\mathcal{D}^m}[L_S(h) = 0]\,.$$

$L_S(h) = 0$ znači da se na skupu S hipoteza h podudara s funkcijom f. Iz ovog te zbog nezavisnog generiranja elemenata iz S slijedi

$$\sum_{h\in\mathcal{H}_B}\mathop{\mathbb{P}}_{S|_x\sim\mathcal{D}^m}[L_S(h) = 0] = \sum_{h\in\mathcal{H}_B}\mathop{\mathbb{P}}_{S|_x\sim\mathcal{D}^m}[\{\forall i, h(x_i) = f(x_i)\}] \;= \sum_{h\in\mathcal{H}_B}\prod_{i=1}^{m}\mathop{\mathbb{P}}_{x_i\sim\mathcal{D}}[h(x_i) = f(x_i)]\,.$$

Prisjetimo se da je prava greška definirana s $L_{\mathcal{D},f}(h) = \mathop{\mathbb{P}}_{x\sim\mathcal{D}}[h(x) \neq f(x)]$ te da za hipoteze is skupa $\mathcal{H}_B$ vrijedi da je $L_{\mathcal{D},f}(h) > \varepsilon$. Dobivamo

$$\sum_{h\in\mathcal{H}_B}\prod_{i=1}^{m}\mathop{\mathbb{P}}_{x_i\sim\mathcal{D}}[h(x_i) = f(x_i)] = \sum_{h\in\mathcal{H}_B}\prod_{i=1}^{m}\Big(1 - L_{\mathcal{D},f}(h)\Big) \;\leq \sum_{h\in\mathcal{H}_B}\prod_{i=1}^{m}(1-\varepsilon) \leq \sum_{h\in\mathcal{H}_B}(1-\varepsilon)^m\,.$$

$\mathcal{H}_B$ je podskup konačnog skupa, pa je i on konačan. Korištenjem ocjene $(1 - x) \leq e^{-x}$ dobijemo

$$\sum_{h\in\mathcal{H}_B}(1-\varepsilon)^m \leq \sum_{h\in\mathcal{H}_B}e^{-\varepsilon m} \leq |\mathcal{H}_B|e^{-\varepsilon m} \leq |\mathcal{H}|e^{-\varepsilon m}\,.$$

Ako spojimo sve prethodne dijelove, dobijemo

$$\mathop{\mathbb{P}}_{S|_x\sim\mathcal{D}^m}[L_{\mathcal{D},f}(h_S) > \varepsilon] \leq |\mathcal{H}|e^{-\varepsilon m}\,.$$

Važno je uočiti kako gornja ograda ne ovisi ni o distribuciji $\mathcal{D}$ ni o funkciji f.

Korolar 1.3.2. *Neka je $\mathcal{H}$ konačna klasa hipoteza, $\delta \in (0,1)$, $\varepsilon > 0$ te $m \in \mathbb{N}$ takav da vrijedi*

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

Tada za svaku funkciju f koja elemetima domene pridružuje prave vrijednosti oznaka i za svaku distribuciju $\mathcal{D}$ za koju vrijedi RA pretpostavka vrijedi

$$\mathbb{P}[L_{\mathcal{D},f}(h_S) \leq \varepsilon] \geq 1 - \delta\,, \tag{1.7}$$

gdje je $S|_x \sim \mathcal{D}^m$, a h_S hipoteza koju vraća ERM algoritam.

Drugim riječima, Korolar nam govori da ako imamo dovoljno veliki m, korištenjem ERM algoritma na konačnu klasu hipoteza dobit ćemo prediktor h_S koji je *vjerojatno približno točan*, tj. prediktor koji za proizvoljne $\varepsilon > 0$ i $\delta \in (0, 1)$ ima pravu grešku manju od ε s vjerojatnosti od barem $1 - \delta$, tj.

$$\mathop{\mathbb{P}}_{S|_x \sim \mathcal{D}^m}[L_{\mathcal{D},f}(h_S) \leq \varepsilon] \geq 1 - \delta\,.$$

Iz (1.7) uzimanjem komplementa slijedi $\mathbb{P}[L_{\mathcal{D},f}(h_S) > \varepsilon] \leq \delta$. To znači da za $m \in \mathbb{N}$ za koji vrijedi $m \geq \ln(|\mathcal{H}|/\delta)/\varepsilon$ te skup S od m elemenata parametar pouzdanosti δ je gornja ograda za vjerojatnost da ERM algoritam vraća hipotezu čija prava greška ima vrijednost veću od ε.

Poglavlje 2

Formalni model za učenje

U ovom poglavlju definirat ćemo glavni model za učenje: vjerojatno približno točan model (engl. *Probably Approximately Correct Model*) ili PAC model za učenje. Model je prvi put definiran u [7], a nastao je potragom za modelom u kojem s visokom pouzdanošću uspješno učenje povlači pronalaženje hipoteze koja predstavlja dobru aproksimaciju traženog koncepta. Ne zahtijeva se nalaženje potpuno preciznog opisa nepoznatog koncepta, već se samo traži da greška aproksimacije bude po volji mala i to s određenom vjerojatnosti.

2.1 PAC učenje

U prošlom poglavlju pokazali smo da za konačnu klasu hipoteza $\mathcal{H}$ vrijedi da ukoliko ERM algoritam upotrijebimo na dovoljno velikom skupu za učenje, dobiveni prediktor h_S će biti vjerojatno približno točan. To nas dovodi do definicije PAC učenja.

Definicija 2.1.1. *Klasu hipoteza $\mathcal{H}$ moguće je naučiti u PAC smislu ako postoji funkcija $m_{\mathcal{H}} : (0,1)^2 \to \mathbb{N}$ i algoritam za učenje sa sljedećim svojstvom: za sve $\varepsilon, \delta \in (0,1)$, za svaku distribuciju $\mathcal{D}$ nad $\mathcal{X}$ te za svaku ciljnu funkciju $f : \mathcal{X} \to \{0,1\}$ slijedi da ako vrijedi RA pretpostavka za $\mathcal{H}$, $\mathcal{D}$ i f, korištenjem algoritma na skupu za učenje koji se sastoji od m nezavisnih primjera generiranih distribucijom $\mathcal{D}$ i označenih s f, gdje je $m > m_{\mathcal{H}}(\varepsilon, \delta)$, algoritam vraća hipotezu h za koju vrijedi*

$$\mathbb{P}[L_{\mathcal{D},f}(h) \leq \varepsilon] \geq 1 - \delta\,.$$

PAC princip učenja nam zapravo govori da ako je neka hipoteza izrazito pogrešna, to će biti vidljivo već na malom podskupu primjera s velikom vjerojatnosti. Također vrijedi i da je za bilo koju hipotezu konzistentnu s dovoljno velikim brojem primjera malo vjerojatno da je izrazito pogrešna, tj. vjerojatno je približno točna.

U definiciji se pojavljuju dva parametra: ε i δ. Parametar točnosti (preciznosti) ε određuje koliko daleko hipoteza h može biti od točne funkcije cilja f, dok parametar pouzdanosti δ određuje vjerojatnost da zahtjev točnosti bude zadovoljen. Znamo da je skup za učenje generiran na slučajan način, pa postoji vjerojatnost da je nereprezentativan (npr. u našem primjeru s e-mailovima može se dogoditi da se skup za učenje sastoji samo od *spam* e-mailova). Čak i u slučaju da dobijemo reprezentativan uzorak, posjedujemo samo konačan broj elemenata domene kojima su pridružene vrijednosti funkcije cilja. Stoga taj uzorak ne može u potpunosti opisati preslikavanje funkcije f. Zbog toga je i uveden parametar ε koji nam dopušta da napravimo grešku na onim elementima domene na kojima nismo u mogućnosti dobro procijeniti vrijednost funkcije cilja.

Već smo spomenuli da je uvjet $\mathbb{P}[L_{\mathcal{D},f}(h) \leq \varepsilon] \geq 1 - \delta$ ekvivalentan uvjetu $\mathbb{P}[L_{\mathcal{D},f}(h) > \varepsilon] \leq \delta$. Zbog $L_{\mathcal{D},f}(h) \geq 0$ slijedi:

$$\mathbb{P}\left[L_{\mathcal{D},f}(h) > \varepsilon\right] = \mathbb{P}\left[L_{\mathcal{D},f}(h) - 0 > \varepsilon\right] = \mathbb{P}\left[\left|L_{\mathcal{D},f}(h) - 0\right| > \varepsilon\right].$$

Promotrimo limes gornjeg izraza za $m \to \infty$. Iz uvjeta PAC učenja i definicije konvergencije niza brojeva slijedi da za sve fiksne $\mathcal{D}$ i f te za svaki $\varepsilon > 0$ vrijedi:

$$\lim_{m\to\infty} \mathbb{P}\left[\left|L_{\mathcal{D},f}(h) - 0\right| > \varepsilon\right] = 0\,.$$

Iz ovoga zaključujemo da se uvjet PAC učenja može interpretirati kao konvergencija po vjerojatnosti od $L_{\mathcal{D},f}(h)$ prema 0 kad $m \to \infty$.

Složenost

Složenost učenja klase $\mathcal{H}$ određena je funkcijom $m_{\mathcal{H}}$, odnosno brojem elemenata u skupu za učenje koji će nam garantirati dobivanje vjerojatno aproksimativno točnog rješenja. Funkcija $m_{\mathcal{H}}$ je funkcija dva parametra: parametra pouzdanosti i parametra točnosti. Naravno, ovisi i o svojstvima klase $\mathcal{H}$. U Korolaru 1.3.2 pokazali smo da složenost ovisi o logaritmu veličine klase $\mathcal{H}$.

Funkcija $m_{\mathcal{H}}$ nije jedinstvena, pa ćemo definirati složenost učenja klase $\mathcal{H}$ kao "minimalnu funkciju", tj. za svaki ε i δ, $m_{\mathcal{H}}(\varepsilon, \delta)$ je najmanji prirodni broj koji zadovoljava PAC učenje. Uvodeći ovu definiciju složenosti, možemo preoblikovati Korolar 1.3.2.

Korolar 2.1.2. *Svaku konačnu klasu hipoteza moguće je naučiti u PAC smislu. Složenost u ovom slučaju zadovoljava:*

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

2.2 Primjeri PAC učenja

Intervali

Zamislimo igru između dva igrača, A i B. Igrač A zamisli proizvoljan interval $[a, b]$ gdje su $a, b \in \mathbb{R}$. Nakon toga generira slučajne brojeve x (njih konačno mnogo, odnosno m), te za svaki x mora reći pripada li zamišljenom intervalu ili ne. Kao oznake pripradnosti uzet ćemo 0 i 1. Na temelju generiranih brojeva x i pridruženih oznaka, igrač B treba pokušati odrediti vrijednosti a i b, odnosno početni interval koji je igrač A zamislio. Igrač B na raspolaganju ima samo konačno mnogo primjera, pa nije u mogućnosti pomoću njih odrediti točan interval, pogotovo zato što su rubovi intervala realni brojevi. Zbog toga, koji god interval na kraju izabere, možemo testirati koliko je točan. Odnosno možemo testirati vjerojatnost da će sljedeći generirani broj dobiti krivu oznaku, ukoliko bude označen intervalom koji je izabrao igrač B. Ako je ta vjerojatnost mala, možemo reći da je igrač B "naučio" interval.

Teorem 2.2.1. *Intervale je moguće naučiti u PAC smislu.*

Neka su $\varepsilon, \delta \in (0, 1)$ prozvoljni, $\mathcal{D}$ proizvoljna distribucija nad skupom realnih brojeva i S skup za učenje od m primjera dan s

$$S = \{(x_1, y_1), ..., (x_m, y_m)\},$$

gdje su x_i realni brojevi, a y_i pridružene oznake iz skupa $\{0, 1\}$. Neka je klasa hipoteza $\mathcal{H}$ dana s $\mathcal{H} = \{\mathbb{1}_{[a,b]} : a, b \in \mathbb{R}\}$. Pretpostavimo da vrijedi RA pretpostavka te da je $J = \mathbb{1}_{[a_0,b_0]}$ hipoteza koja minimizira pravu grešku. Definiramo prediktor $I_S = \mathbb{1}_{[a_1,b_1]}$ gdje su a_1 i b_1 dani s

$$a_1 := \min_{i\in\{1,...,m\}} \{x_i : y_i = 1\},$$
$$b_1 := \max_{i\in\{1,...,m\}} \{x_i : y_i = 1\}.$$

Ako ne postoji x_i takav da za pripadni y_i vrijedi $y_i = 1$, slijedi da je $I_S = \mathbb{1}_\emptyset$ te u tom slučaju u svrhu dokaza definiramo a_1 i b_1 kao

$$a_1 := b_0,$$
$$b_1 := a_0.$$

Da bi dokazali da se intervali mogu naučiti u PAC smislu trebamo pokazati da vrijedi

$$\mathbb{P}_{S|_x\sim\mathcal{D}^m}(L_{\mathcal{D},J}(I_S) > \varepsilon) \leq \delta.$$

Oznake u skupu za učenje su uvijek točne, pa vrijedi da je $[a_1, b_1] \subseteq [a_0, b_0]$. Iz ovog zaključujemo da je prava greška zapravo vjerojatnost da primjer x generiran distribucijom $\mathcal{D}$ pripada intervalima $A = [a_0, a_1)$ ili $B = (b_1, b_0]$, odnosno

$$L_{\mathcal{D},J}(I_S) = \mathop{\mathbb{P}}_{x\sim\mathcal{D}}(I_S(x) \neq J(x)) \leq \mathop{\mathbb{P}}_{x\sim\mathcal{D}}(x \in A) + \mathop{\mathbb{P}}_{x\sim\mathcal{D}}(x \in B)\,,$$

stoga vrijedi

$$\mathop{\mathbb{P}}_{S|_x\sim\mathcal{D}^m}(L_{\mathcal{D},J}(I_S) > \varepsilon) \leq \mathop{\mathbb{P}}_{S|_x\sim\mathcal{D}^m}\left(\mathop{\mathbb{P}}_{x\sim\mathcal{D}}(x \in A) > \frac{\varepsilon}{2}\right) + \mathop{\mathbb{P}}_{S|_x\sim\mathcal{D}^m}\left(\mathop{\mathbb{P}}_{x\sim\mathcal{D}}(x \in B) > \frac{\varepsilon}{2}\right). \tag{2.1}$$

U slučaju da vrijedi

$$\mathop{\mathbb{P}}_{x\sim\mathcal{D}}(x \in [a_0, b_0]) \leq \frac{\varepsilon}{2}\,,$$

iz $A, B \subseteq [a_0, b_0]$ nužno slijedi

$$\mathop{\mathbb{P}}_{S|_x\sim\mathcal{D}^m}(L_{\mathcal{D},J}(I_S) > \varepsilon) = 0\,.$$

Pretpostavimo stoga da je

$$\mathop{\mathbb{P}}_{x\sim\mathcal{D}}(x \in [a_0, b_0]) > \frac{\varepsilon}{2}\,.$$

Definiramo intervale $A' = [a_0, w_1]$ i $B' = [w_2, b_0]$ gdje su w_1 i w_2 dani s

$$w_1 = \inf\{w \in \mathbb{R} : w \leq b_0,\ \mathop{\mathbb{P}}_{x\sim\mathcal{D}}(x \in [a_0, w]) \geq \varepsilon/2\}\,,$$
$$w_2 = \sup\{w \in \mathbb{R} : w \geq a_0,\ \mathop{\mathbb{P}}_{x\sim\mathcal{D}}(x \in [w, b_0]) \geq \varepsilon/2\}\,.$$

Iz definicije intervala A' vidimo da vrijede sljedeće nejednakosti:

$$\mathop{\mathbb{P}}_{x\sim\mathcal{D}}(x \in A') \geq \varepsilon/2\,,$$
$$\mathop{\mathbb{P}}_{x\sim\mathcal{D}}(x \in [a_0, w_1)) \leq \varepsilon/2\,.$$

U slučaju da vrijedi $A \subseteq A'$, dobijemo

$$\mathop{\mathbb{P}}_{x\sim\mathcal{D}}(x \in A) \leq \mathop{\mathbb{P}}_{x\sim\mathcal{D}}(x \in [a_0, w_1)) \leq \varepsilon/2\,.$$

Iz tog slijedi da $\mathbb{P}_{x\sim\mathcal{D}}(x \in A) > \varepsilon/2$ nužno povlači da je $A' \subset A$. Kako je taj slučaj moguć jedino ako u skupu za učenje niti jedna točka nije bila iz A', slijedi

$$\begin{aligned}\mathbb{P}_{S|_x\sim\mathcal{D}^m}\left(\mathbb{P}_{x\sim\mathcal{D}}(x \in A) > \frac{\varepsilon}{2}\right) &\leq \mathbb{P}_{S|_x\sim\mathcal{D}^m}(A' \subset A) \\ &= \left(\mathbb{P}_{x\sim\mathcal{D}}(x \notin A')\right)^m \\ &= \left(1 - \mathbb{P}_{x\sim\mathcal{D}}(x \in A')\right)^m \\ &\leq \left(1 - \frac{\varepsilon}{2}\right)^m .\end{aligned}$$

Analogno vrijedi za B i B', pa iz (2.1) slijedi

$$\mathbb{P}_{S|_x\sim\mathcal{D}^m}(L_{\mathcal{D},J}(I_S) > \varepsilon) \leq \mathbb{P}_{S|_x\sim\mathcal{D}^m}(A' \subset A) + \mathbb{P}_{S|_x\sim\mathcal{D}^m}(B' \subset B) \leq 2\left(1 - \frac{\varepsilon}{2}\right)^m .$$

Želimo da vjerojatnost da greška bude velika bude manja od δ, tj. $2(1 - \varepsilon/2)^m \leq \delta$. Korištenjem $1 - x \leq e^{-x}$ vidimo da je dovoljno riješiti

$$2e^{-\varepsilon m/2} \leq \delta ,$$

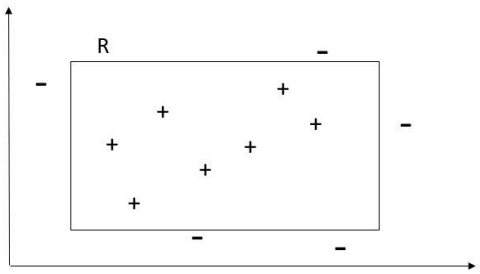
te na kraju dobijemo da za broj primjera treba vrijediti

$$m \geq \frac{2}{\varepsilon}\ln\left(\frac{2}{\delta}\right) .$$

Odnosno, pokazali smo da za m koji zadovoljava gornju nejednakost te za zadane vrijednosti ε i δ, proizvoljnu distribuciju $\mathcal{D}$ i funkciju cilja J, vjerojatnost da će algoritam pridružiti krivu oznaku ε-dijelu novih primjera generiranih distribucijom $\mathcal{D}$ je najviše δ.

Pravokutnici čije su stranice paralelne koordinatnim osima

Pretpostavimo da želimo odrediti pravokutnik u Euklidskoj ravnini čije su stranice paralelne koordinatnim osima. Taj traženi pravokutnik označit ćemo sa R. Dostupan nam je skup za učenje koji se sastoji od m točaka zajedno sa pripadnim oznakama. Točka dobiva oznaku 1 ako se nalazi unutar pravokutnika R, a inače dobiva oznaku 0. Na Slici 2.1 je prikazan traženi pravokutnik zajedno s točkama iz skupa za učenje.



Slika 2.1: Traženi pravokutnik R

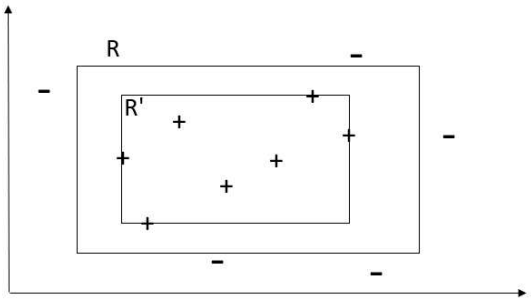
Teorem 2.2.2. *Klasu svih pravokutnika čije su stranice paralelne koordinatnim osima moguće je naučiti u PAC smislu.*

Kako je dokaz ovog teorema sličan dokazu za intervale, bit će objašnjena samo skica dokaza bez većih detalja.

Neka su $\varepsilon, \delta \in (0,1)$ prozvoljni, $\mathcal{D}$ proizvoljna distribucija na skupu $\mathbb{R}^2$, $\mathbb{1}_R$ proizvoljan ciljni (točan) prediktor koji po RA pretpostavci minimizira pravu grešku te neka se skup za učenje sastoji od m točaka (x_{i1}, x_{i2}) s pripadnim oznakama y_i:

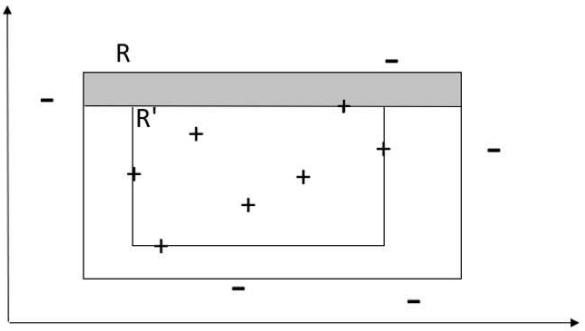
$$S = \{(x_{11}, x_{12}, y_1), ..., (x_{m1}, x_{m2}, y_m)\}.$$

$\mathcal{H} = \{\mathbb{1}_A : A$ je pravokutnik sa stranicama paralelnim koordinatnim osima$\}$ je klasa hipoteza. Koristeći skup za učenje kreiramo novi pravokutnik R'. Neka je $S' \subseteq S$ skup koji čine samo oni elementi od S čija je pripadna oznaka jednaka 1. Ukoliko ne postoji niti jedan y_i takav da je $y_i = 1$, onda je $R' = \emptyset$. Ako takav y_i postoji, definiramo R' kao najuži pravokutnik oko S' takav da su mu stranice paralelne koordinatnim osima. Pravokutnici R i R' prikazani su na sljedećoj slici:



Slika 2.2: Ciljni pravokutnik R zajedno s R'

Prvo što trebamo uočiti je da vrijedi $R' \subseteq R$. Područje na kojem može nastati greška je zapravo dio pravokutnika R koji ne pripada pravokutniku R' i može se rastaviti na 4 manja pravokutnika koja ćemo nazvati A_1, A_2, A_3 i A_4. Neka je A_1 pravokutnik prikazan na sljedećoj slici.



Slika 2.3: Jedan od 4 pravokutnika koji čine područje gdje nastaje greška

Kako vrijedi

$$L_{\mathcal{D},\mathbb{1}_R}(\mathbb{1}_{R'}) = \mathop{\mathbb{P}}_{x\sim\mathcal{D}}(\mathbb{1}_R(x) \neq \mathbb{1}_{R'}(x)) = \mathop{\mathbb{P}}_{x\sim\mathcal{D}}(x \in A_1) + \mathop{P}_{x\sim\mathcal{D}}(x \in A_2) + \mathop{\mathbb{P}}_{x\sim\mathcal{D}}(x \in A_3) + \mathop{P}_{x\sim\mathcal{D}}(x \in A_4),$$

onda slijedi

$$\mathop{\mathbb{P}}_{S|_x\sim\mathcal{D}^m}(L_{\mathcal{D},\mathbb{1}_R}(\mathbb{1}_{R'}) > \varepsilon) \leq \mathop{\mathbb{P}}_{S|_x\sim\mathcal{D}^m}\left(\mathop{\mathbb{P}}_{x\sim\mathcal{D}}(x \in A_1) > \frac{\varepsilon}{4}\right) + \mathop{\mathbb{P}}_{S|_x\sim\mathcal{D}^m}\left(\mathop{P}_{x\sim\mathcal{D}}(x \in A_2) > \frac{\varepsilon}{4}\right) \tag{2.2}$$

$$+ \mathop{\mathbb{P}}_{S|_x\sim\mathcal{D}^m}\left(\mathop{\mathbb{P}}_{x\sim\mathcal{D}}(x \in A_3) > \frac{\varepsilon}{4}\right) + \mathop{\mathbb{P}}_{S|_x\sim\mathcal{D}^m}\left(\mathop{P}_{x\sim\mathcal{D}}(x \in A_4) > \frac{\varepsilon}{4}\right). \tag{2.3}$$

Označimo s w_{RG} vrijednost y-koordinate gornje stranice pravokutnika R te s w_{RD} vrijednost y-koordinate donje stranice pravokutnika R. Neka je R_w oznaka za pravokutnik čija se gornja stranica poklapa s gornjom stranicom od R, a donja stranica ima y-koordinatu jednaku w (podrazumijevamo da vrijedi $w \leq w_{RG}$). Kao i u prethodnom primjeru, možemo pretpostaviti da je $\mathbb{P}_{x\sim\mathcal{D}}(x \in R) > \varepsilon/4$. Definiramo pravokutnik A_1' s $A_1' = R_w$ gdje je w dan s

$$w = \sup\{y \in \mathbb{R} : y \geq w_{RD},\ \mathbb{P}_{x\sim\mathcal{D}}(x \in R_y) \geq \varepsilon/4\}.$$

Budući da iz

$$\mathbb{P}_{x\sim\mathcal{D}}(x \in A_1) > \frac{\varepsilon}{4}$$

nužno slijedi da je $A_1' \subset A_1$, dobijemo

$$\begin{aligned}
\mathbb{P}_{S|_x\sim\mathcal{D}^m}\left(\mathbb{P}_{x\sim\mathcal{D}}(x \in A_1) > \frac{\varepsilon}{4}\right) &\leq \mathbb{P}_{S|_x\sim\mathcal{D}^m}(A_1' \subset A_1) \\
&= \left(\mathbb{P}_{x\sim\mathcal{D}}(x \notin A_1')\right)^m \\
&= \left(1 - \mathbb{P}_{x\sim\mathcal{D}}(x \in A_1')\right)^m \\
&\leq \left(1 - \frac{\varepsilon}{4}\right)^m.
\end{aligned}$$

Analogno se definiraju pravokutnici A_2', A_3' i A_4', pa je zaključak isti za ostala tri područja greške A_2, A_3 i A_4. Iz (2.2) slijedi da je vjerojatnost da novi generirani primjer bude element područja $R \setminus R'$ najviše $4(1 - \varepsilon/4)^m$. Želimo da ova vjerojatnost bude najviše δ, pa korištenjem nejednakosti $1 - x \leq e^{-x}$ dobijemo

$$4(1 - \varepsilon/4)^m \leq 4e^{-\frac{\varepsilon m}{4}} \leq \delta.$$

Iz ovog slijedi da za broj primjera m treba vrijediti

$$m \geq \frac{4}{\varepsilon}\ln\left(\frac{4}{\delta}\right).$$

Sada možemo zaključiti da ako naš algoritam dobije najmanje m primjera u skupu za učenje, onda s vjerojatnosti najmanje $1 - \delta$ konačna hipoteza imat će grešku koja je najviše ε.

Koncentrični krugovi u ravnini

Neka su $\varepsilon, \delta \in (0, 1)$ proizvoljni, $\mathcal{X} = \mathbb{R}^2$, $\mathcal{Y} = \{0, 1\}$, $\mathcal{D}$ prizvoljna distribucija nad $\mathcal{X}$ te neka je $\mathcal{H}$ klasa svih koncentričnih krugova u ravnini, tj.

$$\mathcal{H} = \{h_r : r \in \mathbb{R}_+\},$$

gdje je $h_r(x) = \mathbb{1}_{[\|x\| \leq r]}$. Neka vrijedi RA pretpostavka i neka je h^* ona hipoteza koja minimizira pravu grešku. Njen radijus označimo s r^*. Uzmimo da je skup za učenje dan sa $S = ((x_{i1}, x_{i2}, y_i))_{i=1}^m$ te neka je $\hat{h}$ hipoteza koja predstavlja najuži krug oko onih elemenata skupa za učenje čija je pripadna oznaka 1. Radijus hipoteze $\hat{h}$ označit ćemo s $\hat{r}$. Neka je $\bar{r}$ dan s

$$\bar{r} = \sup\{r : \mathop{\mathbb{P}}_{x\sim\mathcal{D}}[r \leq \|x\| \leq r^*] \geq \varepsilon\}.$$

Definirajmo skup E kao $E = \left\{x \in \mathbb{R}^2 : \bar{r} \leq \|x\| \leq r^*\right\}$. Slično kao i u prethodnim primjerima pokaže se da je vjerojatnost da generiramo skup S za kojeg vrijedi da je $L_{\mathcal{D}}(h_S) > \varepsilon$ odozgo ograničena s vjerojatnosti da niti jedna točka skupa S ne pripada skupu E, tj.

$$\begin{aligned}
\mathop{\mathbb{P}}_{S|_x\sim\mathcal{D}^m}(L_{\mathcal{D},h^*}(\hat{h}) > \varepsilon) &\leq \mathop{\mathbb{P}}_{S|_x\sim\mathcal{D}^m}(\forall x \in S|_x, x \notin E) \\
&= \left(\mathop{\mathbb{P}}_{x\sim\mathcal{D}}(x \notin E)\right)^m \\
&= \left(1 - \mathop{\mathbb{P}}_{x\sim\mathcal{D}}(x \in E)\right)^m \\
&\leq (1-\varepsilon)^m .
\end{aligned}$$

Želimo da ova vjerojatnost bude najviše δ, pa korištenjem $(1-\varepsilon)^m \leq e^{-\varepsilon m}$ dobijemo da za broj primjera m treba vrijediti

$$m \geq \frac{\ln(1/\delta)}{\varepsilon}.$$

2.3 Agnostičko PAC učenje

PAC model je moguće poopćiti kako bi ga mogli primijeniti na veći skup zadataka za učenje. U definiciji PAC učenja zahtijeva se da za danu distribuciju $\mathcal{D}$ i funkciju cilja f, RA pretpostavka bude zadovoljena. Međutim, tokom praktičnog rješavanja zadataka učenja vidljivo je da je ova pretpostavka prejaka. Zbog toga ćemo uvesti pojam agnostičkog PAC učenja u kojem smo se odrekli RA pretpostavke.

Prisjetimo se da RA pretpostavka zahtijeva postojanje hipoteze $h^* \in \mathcal{H}$ za koju vrijedi

$$\mathop{\mathbb{P}}_{x\sim\mathcal{D}}[h^*(x) = f(x)] = 1.$$

U praktičnom rješavanju zadataka ne možemo pretpostaviti da su oznake u potpunosti određene značajkama na kojima gradimo model, odnosno svojstvima koje elementi do-

mene opisuju. Zbog toga ćemo odsada sa $\mathcal{D}$ označavati distribuciju nad $\mathcal{X}\times\mathcal{Y}$, gdje $\mathcal{X}$ predstavlja domenu, a $\mathcal{Y}$ predstavlja skup oznaka. $\mathcal{D}$ možemo rastaviti na dva dijela. Prvi dio je distribucija $\mathcal{D}_x$ nad elementima domene (*marginalna distribucija*), a drugi dio je uvjetna vjerojatnost nad oznakama točaka domene, $\mathcal{D}((x,y)|x)$. Ovakva definicija omogućava postojanje dviju točaka domene koje imaju istu oznaku.

Promjenom definicije distribucije $\mathcal{D}$ došlo je i do promjene u definiciji greške. Empirijska greška ostaje ista te je definirana kao

$$L_S(h) = \frac{1}{m}\sum_{i=1}^{m} \mathbb{1}_{\{h(x_i)\neq y_i\}}\,.$$

Međutim, prava greška se mijenja i definirana je s

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

Cilj nam je opet isti, a to je pronalazak hipoteze $h : \mathcal{X}\to\mathcal{Y}$ koja minimizira pravu grešku.

Bayesov optimalni prediktor

Neka je $\mathcal{X} = \mathbb{R}^k$ za neki $k\in\mathbb{N}$ te neka je $\mathcal{D}$ proizvoljna distribucija nad $\mathcal{X}\times\{0,1\}$ i (X,Y) slučajni vektor s distribucijom $\mathcal{D}$. Pozivajući se na Teorem 2.13 iz [4] koji govori da za svaku slučajnu varijablu Z koja je izmjeriva u odnosu na $\sigma(X)$ postoji Borel izmjeriva funkcija h takva da vrijedi $Z = h(X)$, slijedi da postoji funkcija h takva da je

$$h(X) = \mathbb{P}[Y=1|X] = \mathbb{E}\left[\mathbb{1}_{\{Y=1\}}|X\right]\,.$$

Nadalje, neka je

$$\mathbb{P}[Y=1|X=x] := h(x)\,,\forall x\in X\,.$$

Bayesov optimalni prediktor je preslikavanje dano s

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

Pokazat ćemo da je $f_{\mathcal{D}}$ hipoteza koja minimizira pravu grešku, odnosno da za svaku drugu hipotezu $g:\mathcal{X}\to\{0,1\}$ vrijedi

$$L_{\mathcal{D}}(f_{\mathcal{D}})\leq L_{\mathcal{D}}(g)\,.$$

Koristeći Propoziciju 1.22(b) iz [13], tj. da za svaku σ-algebru $\mathcal{G}$ vrijedi $\mathbb{E}[X] = \mathbb{E}[\mathbb{E}[X|\mathcal{G}]]$

te koristeći svojstvo (f) iz iste propozicije na ograničenu slučajnu varijablu $\mathbb{1}_{\{f_{\mathcal{D}}(X)=0\}}$ izmjerivu u odnosu na $\sigma(X)$, slijedi da za $L_{\mathcal{D}}(f_{\mathcal{D}})$ vrijedi

$$\begin{aligned}
L_{\mathcal{D}}(f_{\mathcal{D}}) &= \mathbb{P}(f_{\mathcal{D}}(X) \neq Y) \\
&= \mathbb{P}(f_{\mathcal{D}}(X) = 0, Y = 1) + \mathbb{P}(f_{\mathcal{D}}(X) = 1, Y = 0) \\
&= \mathbb{E}\left[\mathbb{E}\left[\mathbb{1}_{\{f_{\mathcal{D}}(X)=0, Y=1\}}|X\right]\right] + \mathbb{E}\left[\mathbb{E}\left[\mathbb{1}_{\{f_{\mathcal{D}}(X)=1, Y=0\}}|X\right]\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\mathbb{1}_{\{f_{\mathcal{D}}(X)=0\}}\mathbb{1}_{\{Y=1\}}|X\right]\right] + \mathbb{E}\left[\mathbb{E}\left[\mathbb{1}_{\{f_{\mathcal{D}}(X)=1\}}\mathbb{1}_{\{Y=0\}}|X\right]\right] \\
&= \mathbb{E}\left[\mathbb{1}_{\{f_{\mathcal{D}}(X)=0\}}\mathbb{E}\left[\mathbb{1}_{\{Y=1\}}|X\right]\right] + \mathbb{E}\left[\mathbb{1}_{\{f_{\mathcal{D}}(X)=1\}}\mathbb{E}\left[\mathbb{1}_{\{Y=0\}}|X\right]\right] \\
&= \mathbb{E}\left[\mathbb{1}_{\{h(X)<1/2\}}h(X)\right] + \mathbb{E}\left[\mathbb{1}_{\{h(X)\geq 1/2\}}(1-h(X))\right] \\
&= \mathbb{E}\left[\mathbb{1}_{\{h(X)<1/2\}}h(X) + \mathbb{1}_{\{h(X)\geq 1/2\}}(1-h(X))\right] \\
&= \mathbb{E}\left[\min\{h(X), 1-h(X)\}\right],
\end{aligned}$$

Neka je $Z := \min\{h(X), 1-h(X)\}$. Sličnim raspisom za $L_{\mathcal{D}}(g)$ dobijemo

$$\begin{aligned}
L_{\mathcal{D}}(g) &= \mathbb{P}\left(g(X) \neq Y\right) \\
&= \mathbb{P}\left(g(X) = 0, Y = 1\right) + \mathbb{P}\left(g(X) = 1, Y = 0\right) \\
&= \mathbb{E}\left[\mathbb{E}\left[\mathbb{1}_{\{g(X)=0\}}\mathbb{1}_{\{Y=1\}}|X\right]\right] + \mathbb{E}\left[\mathbb{E}\left[\mathbb{1}_{\{g(X)=1\}}\mathbb{1}_{\{Y=0\}}|X\right]\right] \\
&= \mathbb{E}\left[\mathbb{1}_{\{g(X)=0\}}\mathbb{E}\left[\mathbb{1}_{\{Y=1\}}|X\right]\right] + \mathbb{E}\left[\mathbb{1}_{\{g(X)=1\}}\mathbb{E}\left[\mathbb{1}_{\{Y=0\}}|X\right]\right] \\
&= \mathbb{E}\left[\mathbb{1}_{\{g(X)=0\}}h(X)\right] + \mathbb{E}\left[\mathbb{1}_{\{g(X)=1\}}(1-h(X))\right] \\
&\geq \mathbb{E}\left[\mathbb{1}_{\{g(X)=0\}}Z\right] + \mathbb{E}\left[\mathbb{1}_{\{g(X)=1\}}Z\right] \\
&= \mathbb{E}\left[Z\left(\mathbb{1}_{\{g(X)=0\}} + \mathbb{1}_{\{g(X)=1\}}\right)\right] \\
&= \mathbb{E}\left[Z\right] \\
&= L_{\mathcal{D}}(f_{\mathcal{D}}).
\end{aligned}$$

Iako smo pokazali da Bayesov optimalni prediktor minimizira pravu grešku, zbog nedostatka informacija o distribuciji $\mathcal{D}$ ne možemo ga odrediti. Međutim, iz dokazanog možemo zaključiti da koju god hipotezu h primjenom PAC učenja dobili, njena greška ne može biti manja od minimalne greške $L_{\mathcal{D}}(f_{\mathcal{D}})$. Kasnije ćemo pokazati da ako nemamo nikakvih pretpostavki o distribuciji $\mathcal{D}$, ne možemo garantirati da ćemo dobiti hipotezu koja je jednako dobra kao Bayesov prediktor. Jedino što ćemo zahtijevati je da greška dobivene hipoteze nije puno veća od $L_{\mathcal{D}}(f_{\mathcal{D}})$. Prije toga, uvodimo definiciju agnostičkog PAC učenja.

Definicija 2.3.1. *(Agnostičko PAC učenje) Klasu hipoteza $\mathcal{H}$ moguće je naučiti u agnostičkom PAC smislu ako postoji funkcija $m_{\mathcal{H}} : (0,1)^2 \to \mathbb{N}$ i algoritam za učenje sa sljedećim svojstvom: za sve $\varepsilon, \delta \in (0,1)$ te za svaku distribuciju $\mathcal{D}$ nad $\mathcal{X} \times \mathcal{Y}$ slijedi da korištenjem algoritma na skupu za učenje koji se sastoji od m nezavisnih primjera generiranih distribucijom $\mathcal{D}$, gdje je $m > m_{\mathcal{H}}(\varepsilon, \delta)$, algoritam vraća hipotezu h za koju vrijedi:*

$$\mathbb{P}\left[L_{\mathcal{D}}(h) \leq \min_{h' \in \mathcal{H}} L_{\mathcal{D}}(h') + \varepsilon\right] \geq 1 - \delta\,.$$

Ukoliko vrijedi RA pretpostavka, agnostičko PAC učenje se podudara s običnim PAC učenjem. Međutim, ukoliko RA pretpostavka ne vrijedi, ne možemo garantirati da će hipoteza koju algoritam vraća imati prozvoljno malu grešku. U tom slučaju, korištenjem agnostičkog PAC učenja možemo reći da smo dobili dovoljno dobru hipotezu ako njena greška nije puno veća od najmanje greške po klasi $\mathcal{H}$.

2.4 PAC učenje i funkcija gubitka

U dosadašnjim primjerima koristili smo samo binarnu klasifikaciju (je li e-mail spam ili ne, pripada li točka određenom intervalu ili ne). Međutim, mnogi problemi učenja zahtijevaju klasifikaciju različitu od binarne. Zbog toga želimo naš model poboljšati kako bismo ga mogli primijeniti na veću skupinu zadataka za učenje. Dva najpoznatija problema učenja na kojima ne možemo primijeniti binarnu klasifikaciju su višeklasna klasifikacija i regresija.

Kod višeklasne klasifikacije elemente domene želimo podijeliti u više od dvije skupine. Kao primjer možemo uzeti algoritam kojim računalo prepoznaje rukom napisane znamenke. Kao skup za učenje dobijemo niz skeniranih slika, a cilj nam je "naučiti" računalo da prepozna koje znamenke su zapisane na slici. Imamo 10 znamenki, pa ćemo elemente domene dijeliti u 10 skupina (klasa).

Regresija je metoda koja označava procese procjene povezanosti između $\mathcal{X}$ i $\mathcal{Y}$. Funkcija cilja će biti realna funkcija. Kao primjer možemo uzeti procjenu kreditne sposobnosti pojedinca. Ako želimo procijeniti je li pojedinac kreditno sposoban, procjenu ćemo vršiti klasifikacijom, a ako za pojedinca želimo odrediti koliki maksimalni kredit može dobiti, koristit ćemo regresiju. U tom će slučaju domenu $\mathcal{X}$ činiti vektori u kojima je zabilježena godišnja plaća, dob, spol, zanimanje, itd., a $\mathcal{Y}$ će biti skup $\mathbb{R}$. U ovom slučaju ne možemo koristiti definiciju (1.1) prave greške, pa ju npr. možemo definirati kao

$$L_{\mathcal{D}}(h) = \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}} (h(x) - y)^2\,,$$

odnosno kao očekivani kvadrat razlike između stvarne vrijednosti i procijenjene vrijednosti. Kako bismo izbjegli definiranje prilagođene prave greške za svaki zadatak učenja, uvest ćemo općeniti pojam funkcije gubitka.

Generalizirana funkcija gubitka

Neka je $\mathcal{H}$ skup hipoteza i Z proizvoljna domena. *Funkcija gubitka* je svaka funkcija $\ell : \mathcal{H} \times Z \to \mathbb{R}_+$. U svim dosad prikazanim primjerima Z je $\mathcal{X} \times \mathcal{Y}$, ali postoje i primjeri u kojima može biti drukčije definirana.

Definiramo *funkciju rizika* kao očekivanu grešku hipoteze $h \in \mathcal{H}$ s obzirom na distribuciju $\mathcal{D}$ kao

$$L_{\mathcal{D}}(h) = \mathop{\mathbb{E}}_{z\sim\mathcal{D}}[\ell(h, z)]\,, \tag{2.4}$$

te *funkciju empirijskog rizika* za skup za učenje $S = (z_1, ..., z_m) \in Z^m$ kao

$$L_S(h) = \frac{1}{m}\sum_{i=1}^{m} \ell(h, z_i)\,. \tag{2.5}$$

Funkcija gubitka koju koristimo na problemima klasifikacije (binarne i višeklasne) naziva se **0-1 funkcija gubitka** i dana je s

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

Funkcija kvadratnog gubitka koja se koristi na regresijskim problemima dana je s

$$\ell_{sq}(h, (x, y)) = (h(x) - y)^2\,.$$

Nakon što smo uveli opći pojam funkcije gubitka te pomoću toga definirali funkciju rizika, uvest ćemo pojam agnostičkog PAC učenja za generaliziranu funkciju gubitka.

Definicija 2.4.1. *(Agnostičko PAC učenje za generaliziranu funkciju gubitka) Klasu hipoteza $\mathcal{H}$ moguće je naučiti u agnostičkom PAC smislu s obzirom na skup Z i funkciju gubitka $\ell : \mathcal{H} \times Z \to \mathbb{R}_+$ ako postoji funkcija $m_{\mathcal{H}} : (0, 1)^2 \to \mathbb{N}$ i algoritam za učenje sa sljedećim svojstvom: za sve $\varepsilon, \delta \in (0, 1)$ te za svaku distribuciju $\mathcal{D}$ nad Z slijedi da korištenjem algoritma na skupu za učenje koji se sastoji od m nezavisnih primjera generiranih distribucijom $\mathcal{D}$, gdje je $m > m_{\mathcal{H}}(\varepsilon, \delta)$, algoritam vraća hipotezu h za koju vrijedi*

$$\mathbb{P}\left[L_{\mathcal{D}}(h) \leq \min_{h'\in\mathcal{H}} L_{\mathcal{D}}(h') + \varepsilon\right] \geq 1 - \delta\,,$$

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

Poglavlje 3

No-Free-Lunch teorem i VC dimenzija

Prirodno se postavlja pitanje postojanja univerzalnog algoritma koji bi bio uspješno primjenjiv na sve probleme učenja. Pokazat ćemo da ako nemamo nikakvih pretpostavki na klasu hipoteza $\mathcal{H}$, takav algoritam ne postoji. Iz toga će slijediti da ako ne uvedemo neke dodatne pretpostavke, klasu hipoteza nad beskonačnom domenom ne možemo naučiti u PAC smislu. Na kraju definiramo VC dimenziju za koju ćemo kasnije pokazati da ima ključnu ulogu u karakterizaciji klasa hipoteza koje se mogu naučiti u PAC smislu.

3.1 No-Free-Lunch teorem

No-Free-Lunch teorem (kraće NFL teorem) nam govori da ne postoji algoritam koji je uspješno primjenjiv na sve probleme učenja, tj. za svaki algoritam postoji problem na kojem će prava greška biti velika, dok istovremeno postoje neki drugi algoritmi čija je prava greška za isti problem proizvoljno mala. To bi značilo da svaki problem učenja zahtijeva posebnu analizu, što uključuje provjeru pretpostavki, dodatnih informacija, analizu skupa za učenje, zahtjeve na brzinu izvođenja itd.

Prije nego krenemo na iskaz i dokaz NFL teorema, dokazujemo dvije pomoćne leme. Prva od njih je Markovljeva nejednakost.

Lema 3.1.1. *(Markovljeva nejednakost) Neka je Z nenegativna slučajna varijabla. Tada za svaki $t > 0$ vrijedi*

$$\mathbb{P}(Z \geq t) \leq \frac{\mathbb{E}[Z]}{t}.$$

Dokaz. Primijetimo da vrijedi $\mathbb{P}(Z \geq t) = \mathbb{E}[\mathbb{1}_{\{Z \geq t\}}]$. Ako vrijedi da je $Z \geq t$, slijedi $Z/t \geq 1 \geq \mathbb{1}_{\{Z \geq t\}}$, a ako vrijedi da je $Z < t$, onda opet imamo $Z/t \geq 0 = \mathbb{1}_{\{Z \geq t\}}$. Stoga

$$\mathbb{P}(Z \geq t) = \mathbb{E}[\mathbb{1}_{\{Z \geq t\}}] \leq \mathbb{E}\left[\frac{Z}{t}\right] = \frac{\mathbb{E}[Z]}{t}\,.$$

□

Korištenjem Leme 3.1.1 dokazujemo sljedeću lemu koju ćemo iskoristiti u dokazu NFL teorema.

Lema 3.1.2. *Neka je Z slučajna varijabla koja poprima vrijednosti na skupu $[0,1]$ s očekivanjem $E[Z] = \mu$. Tada za svaki $a \in (0,1)$ vrijedi*

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

Također, za svaki $a \in (0,1)$ vrijedi

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

Dokaz. Definiramo slučajnu varijablu $Y = 1 - Z$. Y je nenegativna slučajna varijabla s očekivanjem $\mathbb{E}[Y] = 1 - E[Z] = 1 - \mu$. Korištenjem Leme 3.1.1 na Y dobijemo

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

Iz ovog slijedi

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

□

Sada prelazimo na iskaz i dokaz NFL teorema.

Teorem 3.1.3. *(No-Free-Lunch) Neka je A proizvoljan algoritam za učenje primjenjiv na probleme binarne klasifikacije uz korištenje 0-1 funkcije gubitka na domeni $\mathcal{X}$. Ako se skup za učenje S sastoji od m elemenata, gdje je $m < |\mathcal{X}|/2$, tada postoje distribucija $\mathcal{D}$ nad $\mathcal{X} \times \{0,1\}$ i funkcija $f : \mathcal{X} \to \{0,1\}$ s pravom greškom $L_{\mathcal{D}}(f) = 0$ takve da vrijedi*

$$\mathbb{P}_{S \sim \mathcal{D}^m}\left[L_{\mathcal{D}}(A(S)) \geq 1/8\right] \geq 1/7\,.$$

Dokaz. Neka je $C \subseteq \mathcal{X}$ bilo koji skup od $2m$ elemenata. Tada postoji $T = 2^{2m}$ mogućih preslikavanja sa skupa C na $\{0,1\}$ i ta preslikavanja ćemo označiti sa $f_1, ..., f_T$. Za svaki $i \in \{1, ..., T\}$ definiramo distribuciju $\mathcal{D}_i$ kao

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

Drugim riječima, D_i je uniformna distribucija na skupu $\{(x, f_i(x)) : x \in C\}$, tj. slučajni vektor s distribucijom D_i dobijemo tako da uniformno izaberemo element $x \in C$ i označimo ga s $y = f_i(x)$. Stoga slijedi

$$L_{\mathcal{D}_i}(f_i) = \mathop{\mathbb{P}}_{(x,y)\sim\mathcal{D}_i}(f_i(x) \neq y) = 0\,.$$

Postoji $k = (2m)^m$ nizova elemenata skupa C duljine m (elementi niza se mogu ponavljati). Označimo ih sa $S_1, ..., S_k$. Za $S_j = (x_1, ..., x_m)$ definiramo $S_j^i = ((x_1, f_i(x_1)), ..., (x_m, f_i(x_m)))$. Za distribuciju $\mathcal{D}_i$ algoritam A prima neki od skupova za učenje $S_1^i, ..., S_k^i$ i vraća funkciju $A(S_i^j)$. Uz to vrijedi da svaki od tih skupova za učenje ima jednaku vjerojatnost da bude odabran, pa slijedi

$$\mathop{\mathbb{E}}_{S\sim\mathcal{D}_i^m}\left[L_{\mathcal{D}_i}(A(S))\right] = \frac{1}{k}\sum_{j=1}^{k} L_{\mathcal{D}_i}(A(S_j^i))\,. \tag{3.1}$$

Također vrijedi

$$\begin{aligned}\max_{i\in[T]}\frac{1}{k}\sum_{j=1}^{k} L_{\mathcal{D}_i}(A(S_j^i)) &\geq \frac{1}{T}\sum_{i=1}^{T}\frac{1}{k}\sum_{j=1}^{k} L_{\mathcal{D}_i}(A(S_j^i)) \\ &= \frac{1}{k}\sum_{j=1}^{k}\frac{1}{T}\sum_{i=1}^{T} L_{\mathcal{D}_i}(A(S_j^i)) \\ &\geq \min_{j\in[k]}\frac{1}{T}\sum_{i=1}^{T} L_{\mathcal{D}_i}(A(S_j^i))\,.\end{aligned} \tag{3.2}$$

Fiksirajmo neki $j \in [k]$. Neka je $S_j = \{x_1, ..., x_m\}$ i neka su $v_1, ..., v_p$ elementi od C koji nisu u S_j. Kako je $p \geq m$, slijedi da za svaku hipotezu $h : C \to \{0, 1\}$ i za svaki $i = 1, ..., T$ vrijedi

$$\begin{aligned}L_{\mathcal{D}_i}(h) &= \frac{1}{2m}\sum_{x\in C}\mathbb{1}_{[h(x)\neq f_i(x)]} \\ &\geq \frac{1}{2m}\sum_{r=1}^{p}\mathbb{1}_{[h(v_r)\neq f_i(v_r)]} \\ &\geq \frac{1}{2p}\sum_{r=1}^{p}\mathbb{1}_{[h(v_r)\neq f_i(v_r)]}\,.\end{aligned} \tag{3.3}$$

Koristeći prethodnu nejednakost dobijemo

$$\begin{aligned}\frac{1}{T}\sum_{i=1}^{T} L_{\mathcal{D}_i}(A(S_j^i)) &\geq \frac{1}{T}\sum_{i=1}^{T}\frac{1}{2p}\sum_{r=1}^{p}\mathbb{1}_{[A(S_j^i)(v_r)\neq f_i(v_r)]} \\ &= \frac{1}{2p}\sum_{r=1}^{p}\frac{1}{T}\sum_{i=1}^{T}\mathbb{1}_{[A(S_j^i)(v_r)\neq f_i(v_r)]} \\ &\geq \frac{1}{2}\min_{r\in[p]}\frac{1}{T}\sum_{i=1}^{T}\mathbb{1}_{[A(S_j^i)(v_r)\neq f_i(v_r)]}\,. \end{aligned} \tag{3.4}$$

Fiksirajmo neki $r \in [p]$ i particirajmo funkcije $f_1, ..., f_T$ u $T/2$ disjunktnih uređenih parova $(f_i, f_{i'})$ za koje vrijedi

$$(\forall c \in C) \quad f_i(c) \neq f_{i'}(c) \iff c = v_r\,.$$

Kako $v_r \notin S_j = \{x_1, ..., x_m\}$, slijedi da za $\forall x \in S_j$ vrijedi $f_i(x) = f_{i'}(x)$. Stoga za svaki uređeni par $(f_i, f_{i'})$ vrijedi $S_j^i = S_j^{i'}$. Dodatno, zbog $f_i(v_r) \neq f_{i'}(v_r)$, iz $A(S_j^i)(v_r) \neq f_i(v_r)$ slijedi $A(S_j^{i'})(v_r) = f_{i'}(v_r)$ (i obratno) pa možemo zaključiti da vrijedi

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

iz čega dobijemo

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

Kombinirajući ovu jednakost sa (3.4), (3.2) i (3.1) slijedi da za svaki algoritam A koji dobije skup za učenje od m elemenata iz skupa $C \times \{0, 1\}$ i vraća hipotezu $A(S)$ vrijedi

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

To znači da za svaki algoritam A' koji dobije skup za učenje od m elemenata iz skupa $\mathcal{X} \times \{0, 1\}$ postoje funkcija $f : \mathcal{X} \to \{0, 1\}$ i distribucija $\mathcal{D}$ nad $\mathcal{X} \times \{0, 1\}$ takve da vrijedi $L_{\mathcal{D}}(f) = 0$ i

$$\mathop{\mathbb{E}}_{S\sim\mathcal{D}^m}\left[L_{\mathcal{D}}(A'(S))\right] \geq \frac{1}{4}\,.$$

Korištenjem Leme 3.1.2 dobijemo

$$\begin{aligned}\mathop{\mathbb{P}}_{S\sim\mathcal{D}^m}\left[L_{\mathcal{D}}(A(S)) \geq 1/8\right] &= \mathop{\mathbb{P}}_{S\sim\mathcal{D}^m}\left[L_{\mathcal{D}}(A(S)) \geq 1 - 7/8\right] \\ &\geq \frac{\mathop{\mathbb{E}}_{S\sim\mathcal{D}^m}\left[L_{\mathcal{D}}(A(S))\right] - (1 - 7/8)}{7/8} \\ &\geq \frac{1/8}{7/8} \\ &= \frac{1}{7}\,.\end{aligned}$$

□

U Korolaru 2.1.2 pokazali smo da svaku konačnu klasu hipoteza možemo naučiti u PAC smislu. Ako nemamo neke dodatne pretpostavke na ciljnu hipotezu f, klasu $\mathcal{H}$ će činiti sve funkcije sa $\mathcal{X}$ u $\{0,1\}$. Ako je domena konačna i klasa $\mathcal{H}$ će biti konačna, pa se možemo pozvati na Korolar 2.1.2. A ako domena nije konačna, sljedećim ćemo korolarom koristeći NFL teorem pokazati da klasu $\mathcal{H}$ ne možemo naučiti u PAC smislu.

Korolar 3.1.4. *Ako je $\mathcal{X}$ beskonačna domena, onda klasu hipoteza $\mathcal{H} = \{h : \mathcal{X} \to \{0,1\}\}$ ne možemo naučiti u PAC smislu.*

Dokaz. Pretpostavimo suprotno, tj. da klasu $\mathcal{H}$ možemo naučiti u PAC smislu. Neka je $\varepsilon < 1/8$ i $\delta < 1/7$. Po definiciji PAC učenja postoji algoritam A i $m = m(\varepsilon, \delta)$ takvi da za svaku distribuciju $\mathcal{D}$ nad $\mathcal{X} \times \{0,1\}$ te za svaku ciljnu funkciju $f : \mathcal{X} \to \{0,1\}$ slijedi da, ako vrijedi $L_{\mathcal{D}}(f) = 0$, korištenjem algoritma na skupu za učenje S koji se sastoji od m nezavisnih primjera generiranih distribucijom $\mathcal{D}$, algoritam vraća hipotezu $A(S)$ za koju vrijedi

$$\mathbb{P}[L_{\mathcal{D}}(A(S)) > \varepsilon] \leq \delta\,. \tag{3.5}$$

S druge strane, zbog $\left|\mathcal{X}\right| > 2m$ iz NFL teorema slijedi da za svaki algoritam, pa tako i za algoritam A, postoje distribucija $\mathcal{D}$ i funkcija f takve da je $L_{\mathcal{D}}(f) = 0$ za koju vrijedi

$$\mathbb{P}[L_{\mathcal{D},f}(A(S)) > \varepsilon] \geq \mathbb{P}[L_{\mathcal{D},f}(A(S)) > 1/8] \geq 1/7 > \delta\,,$$

što je u kontradikciji s (3.5). □

Prethodni korolar nam govori da ako klasu hipoteza $\mathcal{H}$ nad beskonačnom domenom želimo naučiti u PAC smislu, onda moramo imati neke dodatne pretpostavke o hipotezama u klasi. To ne znači da $\mathcal{H}$ mora biti konačna. Naime, u 2. poglavlju smo pokazali da

je intervale moguće naučiti u PAC smislu. Pri tome je klasa hipoteza $\mathcal{H}$ bila dana s $\mathcal{H} = \{\mathbb{1}_{[a,b]} : a, b \in \mathbb{R}\}$. Vidimo da je $\mathcal{H}$ beskonačna, ali ograničena na funkcije određenog oblika, odnosno na skup indikatora intervala skupa $\mathbb{R}$.

3.2 Dekompozicija greške

Neka je $\mathcal{H}$ klasa hipoteza, S skup za učenje označen nekom funkcijom f te h_S rezultat primjene pravila $ERM_{\mathcal{H}}$ na skup S. Tada za pravu grešku vrijedi

$$L_{\mathcal{D}}(h_S) = \epsilon_{\text{app}} + \epsilon_{\text{est}} ,$$

gdje su ϵ_{app} i ϵ_{est} definirani s

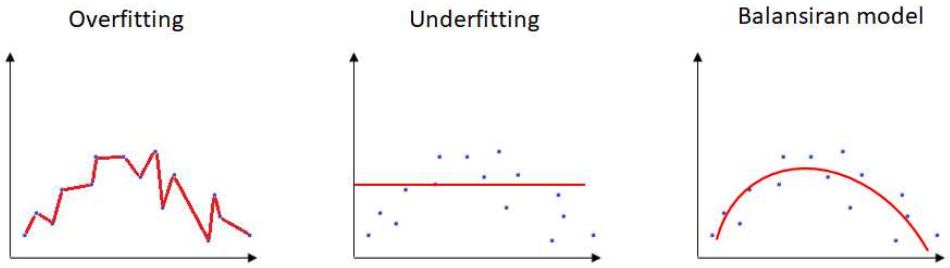
$$\epsilon_{\text{app}} = \min_{h \in \mathcal{H}} L_{\mathcal{D}}(h) ,$$
$$\epsilon_{\text{est}} = L_{\mathcal{D}}(h_S) - \epsilon_{\text{app}} .$$

Greška aproksimacije, u oznaci ϵ_{app}, je minimalna vrijednost prave greške po klasi $\mathcal{H}$. Ne ovisi o veličini skupa S, već samo o klasi $\mathcal{H}$, tj. povećanjem broja hipoteza u klasi $\mathcal{H}$ greška aproksimacije se smanjuje. U praksi, ako model ima veliku grešku aproksimacije to znači da je previše pojednostavljen, zbog čega je prava greška na skupu za učenje i na skupu za testiranje velika.

Greška procjene, u oznaci ϵ_{est}, je razlika između prave greške $ERM_{\mathcal{H}}(S)$ prediktora i greške aproksimacije. Ovisi i o veličini skupa S i o klasi $\mathcal{H}$. U praksi, ako model ima veliku grešku procjene to najčešće znači da je presložen, zbog čega ima jako malu grešku na skupu za učenje, dok je greška na skupu za testiranje velika.

Ako klasu $\mathcal{H}$ izaberemo tako da sadrži veliki broj hipoteza dolazi do smanjenja greške aproksimacije i povećanja greške procjene, pa može doći do *overfittinga*. S druge strane, restringiranjem klase $\mathcal{H}$ na mali skup hipoteza dolazi do povećanja greške aproksimacije i smanjenja greške procjene, što može dovesti do *underfittinga*. Stoga, da bi izgradili dobar model potrebno je pronaći ravnotežu između ove dvije greške tako da prava greška ostane minimalna, što je u literaturi poznato kao *bias-complexity tradeoff*.

Prikaz *underfittinga* i *overfittinga* u odnosu na funkciju koja dobro opisuje model vidi se na sljedećoj slici.



Slika 3.1: Bias-complexity tradeoff

3.3 VC dimenzija

U prvom poglavlju smo pokazali da svaku konačnu klasu hipoteza možemo naučiti u PAC smislu, dok smo u drugom poglavlju pokazali da beskonačnu klasu hipoteza u nekim slučajevima možemo naučiti u PAC smislu, a u nekim slučajevima ne možemo. Iz ovog zaključujemo da broj elemenata klase hipoteza nije dobra karakterizacija mogućnosti učenja u PAC smislu, pa uvodimo pojam VC dimenzije i povezujemo mogućnost učenja u PAC smislu s konačnosti VC dimenzije.

U dokazu NFL teorema koristili smo konačan skup C te činjenicu da možemo birati između svih ciljnih funkcija $f : C \to \{0, 1\}$ kako bi pokazali da postoji distribucija $\mathcal{D}$ takva da prava greška hipoteze koju vraća algoritam A bude velika s velikom vjerojatnosti. Zbog toga ćemo promatrati ponašanje klase $\mathcal{H}$ na skupu C kako bi pronašli dodatne uvjete pomoću kojih ćemo osigurati mogućnost učenja u PAC smislu.

Definicija 3.3.1. *Neka je $\mathcal{H}$ klasa svih funkcija sa $\mathcal{X}$ u $\{0, 1\}$ i neka je $C = \{c_1, ..., c_m\} \subset \mathcal{X}$. Restrikcija klase $\mathcal{H}$ na skup C je skup*

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

Broj elemenata od $\mathcal{H}_C$ može biti najviše $2^{|C|}$, pa uvodimo pojam rastavljanja skupa klasom (engl. *shattering*) kako bi definirali skupove u kojima se taj maksimum postiže.

Definicija 3.3.2. *Kažemo da je skup $C \subset \mathcal{X}$ moguće rastaviti klasom hipoteza $\mathcal{H}$ ako je restrikcija od $\mathcal{H}$ na C skup svih funkcija sa C na $\{0, 1\}$, tj. ako vrijedi $\left|\mathcal{H}_C\right| = 2^{|C|}$.*

Povezivanjem dokaza NFL teorema s definicijom rastavljanja skupa klasom dobijemo sljedeći korolar koji kaže da ako skup C od $2m$ elemenata možemo rastaviti klasom $\mathcal{H}$, onda $\mathcal{H}$ ne možemo naučiti u PAC smislu korištenjem skupa za učenje od m primjera.

Korolar 3.3.3. *Neka je $\mathcal{H}$ klasa hipoteza sa $\mathcal{X}$ u $\{0,1\}$ i m veličina skupa za učenje. Ako postoji skup $C \subset \mathcal{X}$ od $2m$ elemenata kojeg je moguće rastaviti klasom hipoteza $\mathcal{H}$, onda za svaki algoritam A postoji distribucija $\mathcal{D}$ nad $\mathcal{X} \times \{0,1\}$ i hipoteza $h \in \mathcal{H}$ tako da vrijedi $L_{\mathcal{D}}(h) = 0$ te*

$$\mathop{\mathbb{P}}_{S \sim \mathcal{D}^m}[L_{\mathcal{D}}(A(S)) \geq 1/8] \geq 1/7\,.$$

Dokaz. Pretpostavimo da postoji skup $C \subset \mathcal{X}$ od $2m$ elemenata kojeg je moguće rastaviti klasom hipoteza $\mathcal{H}$ te neka je A proizvoljan algoritam. Korištenjem dokaza NFL teorema vidljivo je da za C postoji funkcija $f : C \to \{0,1\}$ i distribucija $\mathcal{D}$ takva da je $\mathcal{D}_X$ koncentrirana na C (tj. x uvijek biramo iz C) tako da vrijedi $L_{\mathcal{D}}(f) = 0$ te $\mathop{\mathbb{P}}_{S \sim \mathcal{D}^m}[L_{\mathcal{D}}(A(S)) \geq 1/8] \geq 1/7$. Budući da je f funkcija na C, a $\mathcal{H}$ rastavlja C, postoji h iz $\mathcal{H}$ takav da je $h(x) = f(x)$ za sve x iz C, tj. restrikcija funkcije h na C je točno f. Budući da je $\mathcal{D}_X$ koncentrirana na C, $L_{\mathcal{D}}(h)$ ovisi samo o vrijednostima funkcije h na skupu C, pa imamo da je$L_{\mathcal{D}}(h) = L_{\mathcal{D}}(f) = 0$. □

Definicija rastavljanja skupa po klasi dana je za proizvoljan skup C. Kako se može dogoditi da $\left|\mathcal{H}_C\right| = 2^{|C|}$ vrijedi za različite skupove s različitim brojem elemenata, definiramo VC dimenziju kao broj elemenata najvećeg podskupa za kojeg ta ista tvrdnja vrijedi.

Definicija 3.3.4. *(Vapnik-Chervonenkisova dimenzija ili kraće VC dimenzija) VC dimenzija klase hipoteza $\mathcal{H}$, u oznaci VCdim($\mathcal{H}$), je veličina najvećeg podskupa $C \subset \mathcal{X}$ kojeg je moguće rastaviti klasom $\mathcal{H}$. Ako je pomoću $\mathcal{H}$ moguće rastaviti po volji velike podskupove od $\mathcal{X}$, onda kažemo da $\mathcal{H}$ ima beskonačnu VC dimenziju.*

Korištenjem definicije VC dimenzije i Korolara 3.3.3 dolazimo do tvrdnje sljedećeg teorema.

Korolar 3.3.5. *Ako klasa $\mathcal{H}$ ima beskonačnu VC dimenziju, onda ju ne možemo naučiti u PAC smislu.*

Dokaz. Pretpostavimo da klasa $\mathcal{H}$ ima beskonačnu VC dimenziju. Tada za svaki $m \in \mathbb{N}$ postoji skup C od $2m$ elemenata kojeg je moguće rastaviti klasom $\mathcal{H}$. Tvrdnja slijedi iz Korolara 3.3.3.

□

Korolar 3.3.5 nam govori da je skup klasa hipoteza koje je moguće naučiti u PAC smislu zapravo podskup skupa klasa hipoteza koje imaju konačnu VC dimenziju. Kasnije ćemo pokazati da su ta dva skupa zapravo jednaka, no prije toga ćemo odrediti VC dimenziju raznih primjera klasa hipoteza.

3.4 Primjeri

U prošlom potpoglavlju definirali smo VC dimenziju i dokazali teorem koji kaže da klasu hipoteza beskonačne dimenzije ne možemo naučiti u PAC smislu. Kako bi pokazali da je VC dimenzija klase hipoteza $\mathcal{H}$ jednaka d, moramo pokazati da postoji skup od d elemenata kojeg je moguće rastaviti klasom $\mathcal{H}$ te da niti jedan skup koji se sastoji od $d + 1$ elemenata nije moguće rastaviti klasom $\mathcal{H}$.

Intervali

Neka je $\mathcal{X} = \mathbb{R}$, $\mathcal{H} = \{\mathbb{1}_{(a,b)} : a, b \in \mathbb{R}, a < b\}$ klasa hipoteza i neka je $C = \{c\}$. Po definiciji slijedi da je restrikcija klase $\mathcal{H}$ na skup C dana s $\mathcal{H}_C = \{h(c) : h \in \mathcal{H}\}$. Za svaki $c \in \mathbb{R}$ postoje $a, b \in \mathbb{R}$ takvi da vrijedi $a, b < c$, iz čega slijedi da za hipotezu $\mathbb{1}_{(a,b)} \in \mathcal{H}$ vrijedi $\mathbb{1}_{(a,b)}(c) = 0$, tj. $0 \in \mathcal{H}_C$. S druge strane, za svaki $c \in \mathbb{R}$ postoje $a, b \in \mathbb{R}$ takvi da vrijedi $a < c$ te $c < b$, iz čega slijedi da za hipotezu $\mathbb{1}_{(a,b)} \in \mathcal{H}$ vrijedi $\mathbb{1}_{(a,b)}(c) = 1$, tj. $1 \in \mathcal{H}_C$. Zbog toga je broj elemenata od $\mathcal{H}_C$ dan s $\left|\mathcal{H}_C\right| = \left|\{0, 1\}\right| = 2^1 = 2^{|C|}$, odnosno $VCdim(\mathcal{H}) \geq 1$.
Neka je $C = \{c_1, c_2\}$ gdje su $c_1, c_2 \in \mathbb{R}$ takvi da vrijedi $c_1 < c_2$. Slijedi

$$\mathcal{H}_C = \{(0, 0), (0, 1), (1, 0), (1, 1)\}.$$

Kako skup $\mathcal{H}_C$ ima $2^2 = 4$ elementa, zaključujemo da vrijedi $VCdim(\mathcal{H}) \geq 2$.
Neka je sad $C = \{c_1, c_2, c_3\}$ gdje su $c_1, c_2, c_3 \in \mathbb{R}$ takvi da vrijedi $c_1 < c_2 < c_3$. Vidimo da biranjem bilo koje funkcije iz $\mathcal{H}$ ne možemo doći do oznaka (1,0,1),tj. ne postoje $a, b \in \mathbb{R}$ takvi da vrijedi

$$\mathbb{1}_{(a,b)}(c_1) = 1 \quad \& \quad \mathbb{1}_{(a,b)}(c_2) = 0 \quad \& \quad \mathbb{1}_{(a,b)}(c_3) = 1\,.$$

Iz ovog slijedi da je $VCdim(\mathcal{H})< 3$, pa zaključujemo da je VC dimenzija klase indikatora intervala jednaka 2.

Polupravci

Neka je $\mathcal{X} = \mathbb{R}$ i klasa indikatora polupravaca dana s

$$\mathcal{H} = \{\mathbb{1}_{(a,\infty)} : a \in \mathbb{R}\}.$$

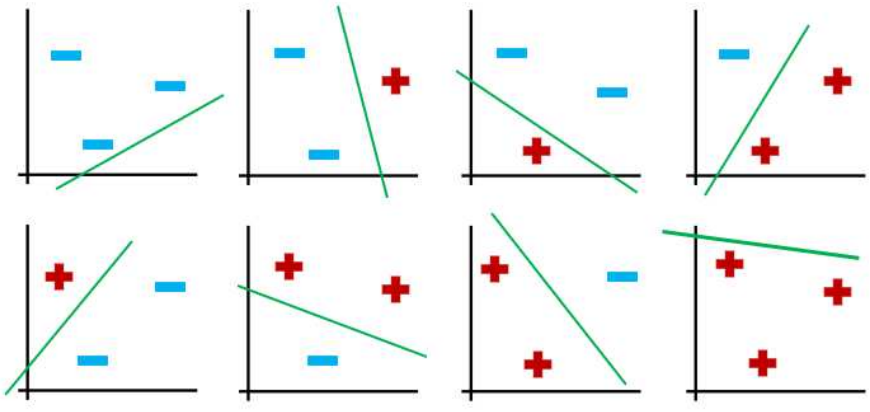
Za skup $C = \{c\}$ postoje točke $a, b \in \mathbb{R}$ takve da vrijedi $a < c < b$. Iz ovog slijedi da je $\mathcal{H}_C = \{0, 1\}$, pa vrijedi da je $VCdim(\mathcal{H}) \geq 1$.
Pretpostavimo da je dan skup $C = \{c_1, c_2\}$ gdje su $c_1, c_2 \in \mathbb{R}$ takvi da vrijedi $c_1 < c_2$. Tada se u skupu $\mathcal{H}_C$ ne može nalaziti element (1,0). Naime, ako za neku hipotezu $\mathbb{1}_{(a,\infty)} \in \mathcal{H}$ vrijedi $\mathbb{1}_{(a,\infty)}(c_1) = 1$, onda mora vrijediti $\mathbb{1}_{(a,\infty)}(c_2) = 1$. Iz ovog slijedi da je $VCdim(\mathcal{H}) < 2$, pa zaključujemo da je VC dimenzija klase indikatora polupravaca jednaka 1.

Poluprostori

Neka je $\mathcal{X} = \mathbb{R}^2$ te $\mathcal{H}$ klasa svih indikatora poluprostora u ravnini, tj.

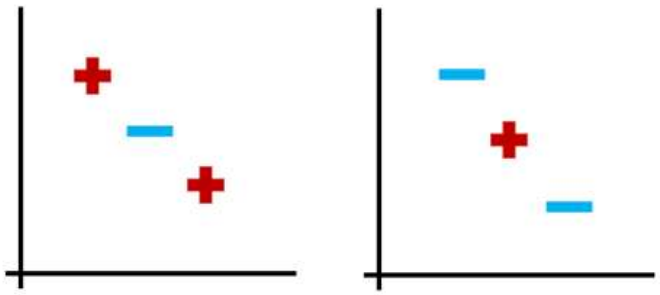
$$\mathcal{H} = \{\mathbb{1}_{\{\alpha_0+\alpha_1 x+\alpha_2 y\geq 0\}} : \alpha_0, \alpha_1, \alpha_2 \in \mathbb{R}\}.$$

Za skupove $C = \{c\}$ i $C = \{c_1, c_2\}$ možemo jednostavno pronaći pravce koji razdvajaju ravninu na dva dijela tako da elementi od C s istom oznakom budu s iste strane pravca.
Promotrimo skup $C = \{c_1, c_2, c_3\}$ za koji postoji 8 mogućih kombinacija oznaka. Za svaku kombinaciju oznaka možemo pronaći pravac koji ravninu dijeli na dva poluprostora tako da točke skupa C s oznakom 1 budu s jedne strane tog pravca, a točke s oznakom 0 s druge strane. Sve kombinacije zajedno s jednim od mogućih pravaca prikazane su na sljedećoj slici, gdje znakom "+" obilježavamo točke s oznakom 1, a znakom "-" točke s oznakom 0.



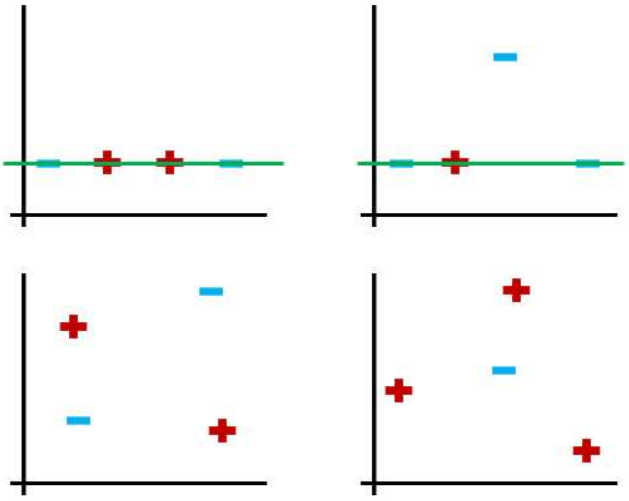
Slika 3.2: Sve kombinacije skupa C s jednim od mogućih pravaca

Raspored točaka u ravnini s prethodne slike nije jedini mogući. Naime, točke skupa C mogu se nalaziti na istom pravcu. Tada postoje dva slučaja za koja ne postoji pravac koji točke razdvaja prema oznakama. Ta dva slučaja prikazana su na sljedećoj slici.



Slika 3.3: Kolinearne točke

Iako u slučaju kolinearnih točaka skup C ne možemo rastaviti klasom $\mathcal{H}$, iz definicije VC dimenzije vidimo da je dovoljno pronaći jedan skup C koji možemo rastaviti klasom $\mathcal{H}$. Kako u prvom slučaju raspored točaka skupa C zadovoljava definiciju, slijedi da je $VCdim(\mathcal{H}) \geq 3$.
Za $C = \{c_1, c_2, c_3, c_4\}$ postoje 4 različita rasporeda točaka u ravnini. Za svaki od tih rasporeda postoji kombinacija oznaka za koju ne postoji pravac koji točke razdvaja prema oznakama, što je prikazano na sljedećoj slici.



Slika 3.4: 4 moguća rasporeda točaka skupa C

Iz ovog slijedi da je $VCdim(\mathcal{H}) < 4$, pa možemo zaključiti da je VC dimenzija klase indikatora poluprostora u ravnini jednaka 3.

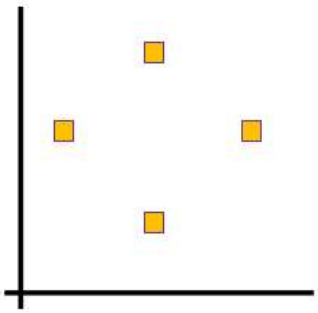
Pravokutnici čije su stranice paralelne koordinatnim osima

Neka je $\mathcal{X} = \mathbb{R}^2$ te $\mathcal{H}$ klasa hipoteza dana s

$$\mathcal{H} = \{\mathbb{1}_A : \text{A je pravokutnik sa stranicama paralelnim koordinatnim osima}\}.$$

Jednostavno je provjeriti da za svaki $k = 1, 2, 3$ postoji skup C od k elemenata za koji postoji raspored točaka u ravnini takav da vrijedi $\left|\mathcal{H}_C\right| = 2^{|C|}$, pa slijedi $VCdim(\mathcal{H}) \geq 3$.

Neka je $C = \{c_1, c_2, c_3, c_4\}$ i neka je raspored točaka skupa C u ravnini dan sljedećom slikom.



Slika 3.5: Raspored točaka skupa C

Vidimo da ako kvadratiće sa slike zamijenimo oznakama ”+” i ”-” u bilo kojem rasporedu, postoji pravokutnik čije su stranice paralelne koordinatnim osima, tako da se točke s oznakom ”+” nalaze u unutrašnjosti pravokutnika. Iz ovog slijedi da je $VCdim(\mathcal{H}) \geq 4$.

Neka je C skup od 5 točaka u ravnini u proizvoljnom fiksnom položaju i neka je A najuži pravokutnik oko skupa C određen najmanjim i najvećim x i y koordinatama točaka skupa C. Primijetimo da je A određen s najmanje dvije i najviše 4 točke, pa rastavimo skup C na dva dijela, U i V, tako da je U skup točaka pomoću kojih smo odredili A, a $V = C \setminus U$. Skupovi U i V su neprazni, pa možemo elementima skupa U pridružiti oznaku 1, a elementima skupa V oznaku 0. Kako se u pravokutniku A nalaze sve točke skupa C, pa i one s oznakom 0, slijedi da skup C ne možemo rastaviti klasom $\mathcal{H}$, tj. $VCdim(\mathcal{H}) < 5$. Već smo prije zaključili da vrijedi $VCdim(\mathcal{H}) \geq 4$, pa slijedi da je VC dimenzija klase indikatora pravokutnika sa stranicama paralelnim koordinatnim osima jednaka 4.

Konačne klase hipoteza

Ako je $\mathcal{H}$ konačna klasa hipoteza, onda za svaki skup C vrijedi

$$\left|\mathcal{H}_C\right| \leq \left|\mathcal{H}\right| .$$

Iz ovog slijedi da ako je $\left|\mathcal{H}\right| < 2^{|C|}$, onda C nije moguće rastaviti klasom $\mathcal{H}$, pa vrijedi

$$VCdim(\mathcal{H}) \leq \log_2(\left|\mathcal{H}\right|) .$$

Za neke klase hipoteza i domene $\mathcal{X}$, VC dimenzija može biti dosta manja od $\log_2(\left|\mathcal{H}\right|)$, ali će u svakom slučaju biti konačna, što će se pokazati ključnim za mogućnost učenja u PAC smislu.

Poglavlje 4

Uniformna konvergencija

U ovom poglavlju cilj nam je uvesti definiciju uniformne konvergencije za općenitu funkciju gubitka. Povezat ćemo taj pojam s agnostičkim PAC učenjem na način da ćemo pokazati da na svaku klasu koja zadovoljava svojstvo uniformne konvergencije možemo naučiti u PAC smislu. Dodatno, pokazat ćemo da su konačne klase podskup klasa koje imaju svojstvo uniformne konvergencije.

4.1 Uniformna konvergencija i učenje

Nakon što ERM algoritam dobije klasu hipoteza $\mathcal{H}$ i skup za učenje S, slijedi procjena greške za svaku hipotezu h. Na kraju kao konačan rezultat dobijemo hipotezu h za koju je greška na skupu za učenje minimalna te se nadamo da ta ista hipoteza minimizira grešku na cijeloj domeni. Zbog toga je dovoljno osigurati da su empirijske greške svih hipoteza h iz $\mathcal{H}$ dobre aproksimacije pravih grešaka tih istih hipoteza.

Propozicija 4.1.1. *Neka je Z proizvoljna domena, $\mathcal{H}$ klasa hipoteza, ℓ funkcija gubitka nad $\mathcal{H} \times Z$, a $\mathcal{D}$ distribucija nad Z. Ako je S skup za učenje takav da $\forall h \in \mathcal{H}$ vrijedi*

$$\left|L_S(h) - L_{\mathcal{D}}(h)\right| \leq \frac{\varepsilon}{2},$$

tada za svaku hipotezu h_S koju dobijemo ERM algoritmom vrijedi

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

Dokaz. Za svaki $h \in \mathcal{H}$ vrijedi

$$\begin{aligned} L_{\mathcal{D}}(h_S) &\leq L_S(h_S) + \frac{\varepsilon}{2} \\ &\leq L_S(h) + \frac{\varepsilon}{2} \\ &\leq L_{\mathcal{D}}(h) + \frac{\varepsilon}{2} + \frac{\varepsilon}{2} \\ &= L_{\mathcal{D}}(h) + \varepsilon\,. \end{aligned}$$

Prva i treća nejednakost vrijede iz pretpostavke propozicije na skup S, a druga nejednakost vrijedi zato što je h_S ERM prediktor. □

Prethodna propozicija nam daje garanciju na dobivanje dobre hipoteze, ali to vrijedi samo za određene vrste skupova za učenje. Zbog toga želimo osigurati da za proizvoljan skup za učenje S, koji je biran na slučajan način distribucijom $\mathcal{D}$, pretpostavke propozicije vrijede s dovoljno velikom vjerojatnosti.

Definicija 4.1.2. *Kažemo da klasa hipoteza $\mathcal{H}$ zadovoljava svojstvo uniformne konvergencije s obzirom na domenu Z i funkciju gubitka ℓ ako postoji funkcija $m^{UC}_{\mathcal{H}} : (0,1)^2 \to \mathbb{N}$ takva da za sve $\varepsilon, \delta \in (0,1)$ i za svaku distribuciju $\mathcal{D}$ nad Z vrijedi: ako je S skup za učenje od $m \geq m^{UC}_{\mathcal{H}}(\varepsilon, \delta)$ nezavisnih primjera dobivenih distribucijom $\mathcal{D}$, onda vrijedi*

$$\mathbb{P}\left[\sup_{h\in\mathcal{H}} \left|L_S(h) - L_{\mathcal{D}}(h)\right| \leq \varepsilon\right] \geq 1 - \delta\,.$$

Termin "uniforman" se odnosi na to da postoji fiksna veličina uzorka koja vrijedi za sve hipoteze iz $\mathcal{H}$.

Funkcija $m^{UC}_{\mathcal{H}}$ mjeri složenost uzorka koji zadovoljava svojstvo uniformne konvergencije, odnosno govori nam koliko primjera se treba nalaziti u skupu za učenje S kako bi za njega s vjerojatnosti od barem $1-\delta$ vrijedilo $\sup_{h\in\mathcal{H}} \left|L_S(h) - L_{\mathcal{D}}(h)\right| \leq \varepsilon$.

Korolar 4.1.3. *Ako klasa $\mathcal{H}$ zadovoljava svojstvo uniformne konvergencije s funkcijom $m^{UC}_{\mathcal{H}}$, onda na $\mathcal{H}$ možemo primijeniti agnostično PAC učenje sa složenosti $m_{\mathcal{H}}(\varepsilon, \delta) \leq m^{UC}_{\mathcal{H}}(\varepsilon/2, \delta)$. U tom slučaju, korištenjem ERM algoritma klasa $\mathcal{H}$ se može naučiti u PAC smislu.*

Dokaz. Slijedi iz Propozicije 4.1.1 i definicije svojstva uniformne konvergencije. □

4.2 Agnostičko PAC učenje konačnih klasa

U prošlom korolaru povezali smo svojstvo uniformne konvergencije sa svojstvom agnostičkog PAC učenja te pokazali odnos složenosti tih dvaju svojstava. Kako smo pokazali da konačne klase hipoteza možemo naučiti u PAC smislu, želimo provjeriti možemo li te iste klase naučiti u agnostičkom PAC smislu. Prije nego krenemo na karakterizaciju konačnih klasa, dokazujemo dvije leme koje će nam biti potrebne u sljedećim dokazima. Prva od njih je Hoeffdingova lema.

Lema 4.2.1. *(Hoeffdingova lema) Neka je X slučajna varijabla koja poprima vrijednosti na intervalu $[a, b]$ za neke $a \leq 0 \leq b$ i za koju vrijedi da je $\mathbb{E}[X] = 0$. Tada za svaki $\lambda > 0$ vrijedi*

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

Dokaz. Funkcija $f(x) = e^{\lambda x}$ je konveksna, pa za svaki $\alpha \in (0, 1)$ i za svaki $x \in [a, b]$ vrijedi

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

Postavljanjem α na $\frac{b-x}{b-a}$ nejednakost glasi

$$e^{\lambda x} \leq \frac{b - x}{b - a} e^{\lambda a} + \frac{x - a}{b - a} e^{\lambda b}.$$

Uzimanjem očekivanja i korištenjem da je $\mathbb{E}[X] = 0$ dobijemo

$$\mathbb{E}[e^{\lambda X}] \leq \frac{b - \mathbb{E}[X]}{b - a} e^{\lambda a} + \frac{\mathbb{E}[X] - a}{b - a} e^{\lambda b} = \frac{b}{b - a} e^{\lambda a} - \frac{a}{b - a} e^{\lambda b}. \tag{4.1}$$

Definirat ćemo h, p i funkciju $L(h)$ sa

$$h := \lambda(b - a), \quad p := \frac{-a}{b - a}, \quad L(h) := -hp + \ln(1 - p + pe^h).$$

Tada slijedi

$$e^{L(h)} = \frac{b}{b - a} e^{\lambda a} - \frac{a}{b - a} e^{\lambda b},$$

pa (4.1) možemo zapisati kao

$$\mathbb{E}[e^{\lambda X}] \leq e^{L(h)}.$$

Stoga, da bi dokazali tvrdnju leme dovoljno je pokazati da vrijedi

$$e^{L(h)} \leq e^{\frac{\lambda^2(b-a)^2}{8}}.$$

Odnosno, dovoljno je pokazati da vrijedi

$$L(h) \leq \frac{\lambda^2(b-a)^2}{8} = \frac{h^2}{8}\,.$$

Funkcija $L(h)$ je dva puta diferencijabilna u svakoj točki $h \geq 0$ pa po Taylorovom teoremu slijedi da za svaki $h \geq 0$ postoji $h_0 \in [0, h]$ takav da vrijedi

$$L(h) = L(0) + hL'(0) + \frac{1}{2}h^2 L''(h_0)\,. \tag{4.2}$$

Kako je $L(0) = 0$, a deriviranjem dobijemo da vrijedi $L'(0) = 0$ te $L''(h) \leq 1/4$ za svaki $h \geq 0$, kombiniranjem sa (4.2) dobijemo

$$L(h) = \frac{1}{2}h^2 L''(h_0) \leq \frac{h^2}{8}\,.$$

□

Korištenjem Hoeffdingove leme i Markovljeve nejednakosti, tj. Leme 3.1.1, dokazujemo sljedeću lemu.

Lema 4.2.2. *(Hoeffdingova nejednakost) Neka je $\theta_1, ..., \theta_m$ niz nezavisnih slučajnih varijabli. Ako su $a, b, \mu_1, ..., \mu_m \in \mathbb{R}$ takvi da za svaki i vrijedi da je $\mathbb{E}(\theta_i) = \mu_i$ te $\mathbb{P}[a \leq \theta_i \leq b] = 1$, onda za svaki $\varepsilon > 0$ vrijedi*

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

Dokaz. Neka je $X_i = \theta_i - \mu_i$ te $\overline{X} = \frac{1}{m}\sum_{i=1}^{m} X_i$. Korištenjem Leme 3.1.1 i monotonosti eksponencijalne funkcije dobijemo da za svaki $\lambda > 0$ i za svaki $\varepsilon > 0$ vrijedi

$$\mathbb{P}[\overline{X} \geq \varepsilon] = \mathbb{P}[e^{\lambda\overline{X}} \geq e^{\lambda\varepsilon}] \leq e^{-\lambda\varepsilon}\,\mathbb{E}[e^{\lambda\overline{X}}]\,.$$

Nezavisnost slučajnih varijabli $\theta_1, ..., \theta_m$ povlači

$$\mathbb{E}[e^{\lambda\overline{X}}] = \mathbb{E}\left[\prod_{i=1}^{m} e^{\lambda X_i/m}\right] = \prod_{i=1}^{m}\mathbb{E}\left[e^{\lambda X_i/m}\right]\,.$$

Kako za svaki $i \in \{1, ..., m\}$ vrijedi $\mathbb{P}[a \leq \theta_i \leq b] = 1$, slijedi

$$\mathbb{P}[a - \mu_i \leq \theta_i - \mu_i \leq b - \mu_i] = \mathbb{P}[a - \mu_i \leq X_i \leq b - \mu_i] = 1\,.$$

Iz nezavisnosti slučajnih varijabli $\theta_1, ..., \theta_m$ dobijemo da za očekivanje slučajne varijable X_i vrijedi

$$\begin{aligned}\mathbb{E}[X_i] &= \mathbb{E}[\theta_i - \mu_i] \\ &= \mathbb{E}[\theta_i] - \mathbb{E}[\mu_i] \\ &= 0\,,\end{aligned}$$

pa su zadovoljene pretpostavke Hoeffdingove leme čijom primjenom na slučajnu varijablu X_i te $\lambda/m > 0$ za svaki $i \in \{1, ..., m\}$ dobijemo

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

Postavljanjem λ na $4m\varepsilon/(b-a)^2$ dobijemo

$$\mathbb{P}[\overline{X} \geq \varepsilon] \leq e^{-\lambda\varepsilon} \prod_{i=1}^{m} e^{\frac{\lambda^2(b-a)^2}{8m^2}} = e^{-\lambda\varepsilon + \frac{\lambda^2(b-a)^2}{8m}} = e^{-\frac{2m\varepsilon^2}{(b-a)^2}}\,.$$

Analogno vrijedi i za varijablu $-\overline{X}$, tj. zaključak prethodnog dokaza primijenimo na varijable $-\theta_i$ koje imaju očekivanja $-\mu_i$ i nalaze se u segmentu $[-b, -a]$, pa dobijemo

$$\begin{aligned}\mathbb{P}[-\overline{X} \geq \varepsilon] &= \mathbb{P}[e^{-\lambda\overline{X}} \geq e^{\lambda\varepsilon}] \\ &\leq e^{-\lambda\varepsilon}\,\mathbb{E}[e^{-\lambda\overline{X}}] \\ &= e^{-\lambda\varepsilon}\,\mathbb{E}\left[\prod_{i=1}^{m} e^{-\lambda X_i/m}\right] \\ &= e^{-\lambda\varepsilon} \prod_{i=1}^{m} \mathbb{E}\left[e^{-\lambda X_i/m}\right] \\ &\leq e^{\lambda\varepsilon} \prod_{i=1}^{m} e^{\frac{\lambda^2(b-a)^2}{8m^2}} \\ &= e^{\lambda\varepsilon}\, e^{\frac{\lambda^2(b-a)^2}{8m}}\,.\end{aligned}$$

Postavljanjem λ na $4m\varepsilon/(b-a)^2$ dobijemo

$$\mathbb{P}[-\overline{X} \geq \varepsilon] \leq e^{-\lambda\varepsilon}\, e^{\frac{\lambda^2(b-a)^2}{8m}} = e^{-\frac{2m\varepsilon^2}{(b-a)^2}}\,.$$

Kombiniranjem prethodnih nejednakosti slijedi tvdnja teorema, tj. vrijedi

$$\begin{aligned}
\mathbb{P}\left[\left|\frac{1}{m}\sum_{i=1}^{m}(\theta_i-\mu_i)\right|>\varepsilon\right] &= \mathbb{P}\left[\left|\overline{X}\right|>\varepsilon\right] \\
&\leq \mathbb{P}\left[\overline{X}\geq\varepsilon\right]+\mathbb{P}\left[\overline{X}\leq-\varepsilon\right] \\
&= \mathbb{P}\left[\overline{X}\geq\varepsilon\right]+\mathbb{P}\left[-\overline{X}\geq\varepsilon\right] \\
&\leq e^{-\frac{2m\varepsilon^2}{(b-a)^2}}+e^{-\frac{2m\varepsilon^2}{(b-a)^2}} \\
&= 2e^{-\frac{2m\varepsilon^2}{(b-a)^2}}.
\end{aligned}$$

□

Korištenjem Korolara 4.1.3 možemo zaključiti da ako dokažemo da proizvoljna klasa ima svojstvo uniformne konvergencije, slijedi da na nju možemo primijeniti agnostičko PAC učenje. Zbog toga je i za konačne klase dovoljno dokazati da imaju svojstvo uniformne konvergencije.

Korolar 4.2.3. *Neka je $\mathcal{H}$ konačna klasa hipoteza, Z domena i neka je funkcija gubitka dana s $\ell : \mathcal{H}\times Z\to[0,1]$. Tada klasa $\mathcal{H}$ zadovoljava svojstvo uniformne konvergencije gdje složenost zadovoljava*

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

Dodatno, $\mathcal{H}$ možemo naučiti u agnostičkom PAC smislu koristeći ERM algoritam. U tom slučaju za složenosti vrijedi

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

Dokaz. Neka su ε i δ proizvoljni i neka je $\mathcal{H}$ konačna klasa hipoteza. Želimo pronaći broj primjera m koji će nam garantirati da za bilo koju distribuciju $\mathcal{D}$ s vjerojatnosti od barem $1-\delta$, za skup za učenje $S=(z_1,...,z_m)$ od m nezavisnih primjera i za svaku hipotezu $h\in\mathcal{H}$ vrijedi $|L_S(h)-L_{\mathcal{D}}(h)|\leq\varepsilon$. Odnosno,

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

Uzimanjem komplementa dolazimo do sljedećeg izraza

$$\mathbb{P}_{S\sim\mathcal{D}^m}[\{\exists h\in\mathcal{H},\ |L_S(h)-L_{\mathcal{D}}(h)|>\varepsilon\}]<\delta. \tag{4.3}$$

Događaj $\{\exists h \in \mathcal{H},\ |L_S(h) - L_{\mathcal{D}}(h)| > \varepsilon\}$ možemo zapisati kao uniju događaja iz $\mathcal{H}$, tj. kao

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

Iskoristimo subaditivnost vjerojatnosti na (4.3) pa dobijemo

$$\mathop{\mathbb{P}}_{S\sim\mathcal{D}^m}\left[\{\exists h \in \mathcal{H},\ |L_S(h) - L_{\mathcal{D}}(h)| > \varepsilon\}\right] \leq \sum_{h\in\mathcal{H}} \mathop{\mathbb{P}}_{S\sim\mathcal{D}^m}\left[\ \left|L_S(h) - L_{\mathcal{D}}(h)\right| > \varepsilon\right]\,. \tag{4.4}$$

Kako je prava greška definirana kao $L_{\mathcal{D}}(h) = \mathbb{E}_{z\sim\mathcal{D}}[\ell(h,z)]$, empirijska greška kao $L_S(h) = \frac{1}{m}\sum_{i=1}^{m}\ell(h,z_i)$, a primjeri z_i dobiveni nezavisnim generiranjem iz iste distribucije $\mathcal{D}$, slijedi

$$\mathop{\mathbb{E}}_{z_1,\ldots,z_m\sim\mathcal{D}}\left[L_S(h)\right] = \mathop{\mathbb{E}}_{z\sim\mathcal{D}}[\ell(h,z)] = L_{\mathcal{D}}(h)\,.$$

Primjenom Hoeffdingove nejednakosti na svaki sumand iz (4.4), uz pretpostavku da funkcija ℓ poprima vrijednosti u skupu $[0,1]$, dobijemo

$$\begin{aligned}
\mathop{\mathbb{P}}_{S\sim\mathcal{D}^m}\left[\{\exists h \in \mathcal{H},\ |L_S(h) - L_{\mathcal{D}}(h)| > \varepsilon\}\right] &\leq \sum_{h\in\mathcal{H}} \mathop{\mathbb{P}}_{S\sim\mathcal{D}^m}\left[\ \left|L_S(h) - L_{\mathcal{D}}(h)\right| > \varepsilon\right] \\
&\leq \sum_{h\in\mathcal{H}} 2\exp(-2m\varepsilon^2) \\
&= |\mathcal{H}|\,2\exp(-2m\varepsilon^2)\,.
\end{aligned}$$

Kako želimo odozgo ograničiti vjerojatnost $\mathop{\mathbb{P}}_{S\sim\mathcal{D}^m}\left[\{\exists h \in \mathcal{H},\ |L_S(h) - L_{\mathcal{D}}(h)| > \varepsilon\}\right]$ s δ, za broj primjera m dovoljno je uzeti

$$m \geq \frac{\ln(2|\mathcal{H}|/\delta)}{2\varepsilon^2}\,.$$

□

U prethodnom korolaru pretpostavili smo da je kodomena funkcije gubitka interval [0,1]. Međutim, to ne mora uvijek vrijediti. Pretpostavimo da je kodomena proizvoljni interval [a,b], gdje su $a, b \geq 0$. Prvo se vratimo na nejednakost koju želimo ograničiti, odnosno nejednakost (4.4). Ponovno ćemo iskoristiti Hoeffdingovu nejednakost, uzimajući u obzir da varijable poprimaju vrijednosti u [a,b]. Tada vrijedi

$$\begin{aligned}
\mathop{\mathbb{P}}_{S\sim\mathcal{D}^m}\left[\{\exists h \in \mathcal{H},\ |L_S(h) - L_{\mathcal{D}}(h)| > \varepsilon\}\right] &\leq \sum_{h\in\mathcal{H}} \mathop{\mathbb{P}}_{S\sim\mathcal{D}^m}\left[\ \left|L_S(h) - L_{\mathcal{D}}(h)\right| > \varepsilon\right] \\
&\leq \sum_{h\in\mathcal{H}} 2\exp(-2m\varepsilon^2/(b-a)^2) \\
&= |\mathcal{H}|2\exp(-2m\varepsilon^2/(b-a)^2)\,.
\end{aligned}$$

Kako želimo s δ odozgo ograničiti vrijednost od

$$\mathop{\mathbb{P}}_{S\sim\mathcal{D}^m}\left[\{\exists h\in\mathcal{H},\ |L_S(h)-L_{\mathcal{D}}(h)|>\varepsilon\}\right],$$

dovoljno je da vrijedi

$$|\mathcal{H}|2\exp(-2m\varepsilon^2/(b-a)^2)<\delta\,.$$

Iz ovog dobijemo da složenost zadovoljava

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

Poglavlje 5

Fundamentalni teorem statističkog učenja

U 3. poglavlju pokazali smo da je skup klasa hipoteza koje je moguće naučiti u PAC smislu zapravo podskup skupa klasa hipoteza koje imaju konačnu VC dimenziju. Sada ćemo uz korištenje *Sauerove leme* preko *Fundamentalnog teorema statističkog učenja* pokazati da su ta dva skupa jednaka te da zadovoljavaju svojstvo uniformne konvergencije definirano u prošlom poglavlju. Iako smo uniformu konvergenciju definirali za općenitu funkciju gubitka, sada ćemo se ograničiti na 0-1 funkciju gubitka.

5.1 Sauerova lema

Prije nego krenemo na iskaz i dokaz Sauerove leme, definirat ćemo pojam funkcije rasta koja za $m \in \mathbb{N}$ daje maksimalnu veličinu restrikcije klase $\mathcal{H}$ po svim podskupovima od $\mathcal{X}$ od m elemenata.

Definicija 5.1.1. *(Funkcija rasta) Neka je $\mathcal{H}$ klasa hipoteza. Funkcija rasta klase $\mathcal{H}$ je funkcija $\tau_{\mathcal{H}} : \mathbb{N} \to \mathbb{N}$ definirana s*

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

Ako je $VCdim(\mathcal{H}) = d$, onda za svaki $m \leq d$ vrijedi $\tau_{\mathcal{H}}(m) = 2^m$. Nadalje, ako je VC dimenzija konačna, tj. ako postoji m takav da vrijedi $m > VCdim(\mathcal{H})$, tada funkcija rasta raste polinomijalno s $m \in \mathbb{N}$. Posljednja tvrdnja je poznata kao *Sauer-Shelah-Perles lema* ili kraće *Sauerova lema*.

Lema 5.1.2. *(Sauer-Shelah-Perles) Neka je $\mathcal{H}$ klasa hipoteza za koju vrijedi $VCdim(\mathcal{H}) \leq d < \infty$. Tada za svaki m vrijedi*

$$\tau_{\mathcal{H}}(m) \leq \sum_{i=0}^{d} \binom{m}{i}.$$

Dodatno, ako je $m \geq d$, onda vrijedi

$$\tau_{\mathcal{H}}(m) \leq (em/d)^d .$$

Dokaz. Za klasu $\mathcal{H}$ za koju je $VCdim(\mathcal{H}) \leq d$, za svaki $m \in \mathbb{N}$ i svaki skup $C = \{c_1, ..., c_m\} \subset X$ vrijedi

$$\left|\{B \subseteq C : \ B \text{ možemo rastaviti klasom } \mathcal{H} \ \}\right| \leq \sum_{i=0}^{d} \binom{m}{i}. \tag{5.1}$$

Naime, kako je $VCdim(\mathcal{H}) \leq d$, onda svaki skup koji ima više od d elemenata ne možemo rastaviti klasom $\mathcal{H}$. Stoga, ako neki skup $B \subseteq C$ možemo rastaviti klasom $\mathcal{H}$, on ima najviše d elemenata. Kako C ima m elemenata, slijedi desna strana od (5.1), pa je za dokaz leme dovoljno je dokazati da za svaki $C = \{c_1, ..., c_m\}$ vrijedi

$$\forall \mathcal{H}, \quad \left|\mathcal{H}_C\right| \leq \left|\{B \subseteq C : \ B \text{ možemo rastaviti klasom } \mathcal{H} \ \}\right| . \tag{5.2}$$

Dokaz provodimo indukcijom po m. Neka je $m = 1$, tj. $C = \{c\}$. Ako je $\left|\mathcal{H}_C\right| = 1$, onda za svaki $h \in \mathcal{H}$ vrijedi $h(c) = 0$ ili za svaki $h \in \mathcal{H}$ vrijedi $h(c) = 1$, odnosno C ne možemo rastaviti klasom $\mathcal{H}$. Kako za prazan skup uvijek vrijedi da se može rastaviti klasom $\mathcal{H}$ zaključujemo da vrijedi

$$\left|\{B \subseteq C : \ B \text{ možemo rastaviti klasom } \mathcal{H} \ \}\right| = \left|\{\emptyset\}\right| = 1 .$$

Ako je $\left|\mathcal{H}_C\right| = 2$, onda postoji $h \in \mathcal{H}$ takav da vrijedi $h(c) = 0$ i postoji $h \in \mathcal{H}$ takav da vrijedi $h(c) = 1$. To znači da C možemo rastaviti klasom $\mathcal{H}$ iz čega slijedi

$$\left|\{B \subseteq C : \ B \text{ možemo rastaviti klasom } \mathcal{H} \ \}\right| = \left|\{\emptyset, C\}\right| = 2 .$$

Pretpostavimo da (5.2) vrijedi za sve $k < m$. Neka je klasa $\mathcal{H}$ fiksna i neka je $C = \{c_1, ..., c_m\}$ proizvoljan skup od m elemenata. Definiramo skup $C' = \{c_2, ..., c_m\}$ te neka su Y_0 i Y_1 skupovi dani s

$$Y_0 = \{(y_2, ..., y_m) : (0, y_2, ..., y_m) \in \mathcal{H}_C \text{ ili } (1, y_2, ..., y_m) \in \mathcal{H}_C\},$$
$$Y_1 = \{(y_2, ..., y_m) : (0, y_2, ..., y_m) \in \mathcal{H}_C \text{ i } (1, y_2, ..., y_m) \in \mathcal{H}_C\}.$$

Općenito, za broj elemenata proizvoljnih skupova A i B vrijedi

$$|A| + |B| = |A \cup B| + |A \cap B| \ ,$$

iz čega slijedi

$$\begin{aligned}
\left|\mathcal{H}_C\right| &= \left|\{(h(c_1), ..., h(c_m)) : h \in \mathcal{H}\}\right| \\
&= \left|\{(h(c_2), ..., h(c_m)) : (0, h(c_2), ..., h(c_m)) \in \mathcal{H}_C\}\right| + \left|\{(h(c_2), ..., h(c_m)) : (1, h(c_2), ..., h(c_m)) \in \mathcal{H}_C\}\right| \\
&= \left|\{(h(c_2), ..., h(c_m)) : (0, h(c_2), ..., h(c_m)) \in \mathcal{H}_C \text{ ili } (1, h(c_2), ..., h(c_m)) \in \mathcal{H}_C\}\right| \\
&\quad + \left|\{(h(c_2), ..., h(c_m)) : (0, h(c_2), ..., h(c_m)) \in \mathcal{H}_C \text{ i } (1, h(c_2), ..., h(c_m)) \in \mathcal{H}_C\}\right| \\
&= |Y_0| + |Y_1| \ .
\end{aligned}$$

Iz definicije rastavljanja skupa klasom slijedi $Y_0 = \mathcal{H}_{C'}$. Korištenjem pretpostavke indukcije na klasu $\mathcal{H}$ i na skup C' dobijemo

$$\begin{aligned}
|Y_0| = \left|\mathcal{H}_{C'}\right| &\leq \left|\{B \subseteq C' : \ B \text{ možemo rastaviti klasom } \mathcal{H} \ \}\right| \\
&= \left|\{B \subseteq C : c_1 \notin B \text{ i } B \text{ možemo rastaviti klasom } \mathcal{H} \ \}\right| \ .
\end{aligned}$$

Definiramo klasu $\mathcal{H}' \subseteq \mathcal{H}$ s

$$\mathcal{H}' = \{h \in \mathcal{H} : \exists h' \in \mathcal{H} \text{ t.d } 1 - h'(c_1) = h(c_1), h'(c_i) = h(c_i) \text{ za } i = 2, ..., m\} \ .$$

Vrijedi $Y_1 = \mathcal{H}'_{C'}$ te klasa $\mathcal{H}'$ sadrži hipoteze iz $\mathcal{H}$ koje se podudaraju na C', a razlikuju u c_1. Zbog toga vrijedi da ako $B \subseteq C'$ možemo rastaviti klasom $\mathcal{H}'$, onda $B \cup \{c_1\}$ također možemo rastaviti klasom $\mathcal{H}'$. Također vrijedi i obrat. Koristeći pretpostavku indukcije sada slijedi

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

Koristeći rastav od $\left|\mathcal{H}_C\right|$ dobijemo

$$\begin{aligned}
\left|\mathcal{H}_C\right| &= |Y_0| + |Y_1| \\
&\leq \left|\{B \subseteq C : c_1 \notin B \text{ i } B \text{ možemo rastaviti klasom } \mathcal{H} \ \}\right| \\
&+ \left|\{B \subseteq C : c_1 \in B \text{ i } B \text{ možemo rastaviti klasom } \mathcal{H} \ \}\right| \\
&= \left|\{B \subseteq C : B \text{ možemo rastaviti klasom } \mathcal{H} \ \}\right| \ .
\end{aligned}$$

Neka je $m \geq d$. Slijedi

$$\begin{aligned}\sum_{i=0}^{d}\binom{m}{i} &\leq \sum_{i=0}^{d}\binom{m}{i}\left(\frac{m}{d}\right)^{d-i}\\ &\leq \sum_{i=0}^{m}\binom{m}{i}\left(\frac{m}{d}\right)^{d-i}\\ &= \left(\frac{m}{d}\right)^{d}\sum_{i=0}^{m}\binom{m}{i}\left(\frac{d}{m}\right)^{i}\\ &= \left(\frac{m}{d}\right)^{d}\left(1+\frac{d}{m}\right)^{m}\\ &\leq \left(\frac{m}{d}\right)^{d}e^{d}\,.\end{aligned}$$

□

5.2 Fundamentalni teorem statističkog učenja

Sljedeća lema, poznata kao *Massartova lema*, u terminima funkcije rasta nam daje gornju granicu na prosječnu vrijednost slučajnih varijabli $\epsilon_1, ..., \epsilon_m$ koje imaju Rademacherovu distribuciju, odnosno za svaki $i \in \{1, ..., m\}$ vrijedi $\mathbb{P}[\epsilon_i = 1] = \mathbb{P}[\epsilon_i = -1] = \frac{1}{2}$.

Lema 5.2.1. *(Massart) Neka je $\mathcal{A}$ konačan podskup od $\mathbb{R}^m$ i neka su $\epsilon_1, ..., \epsilon_m$ nezavisne slučajne varijable takve da za svaki $i \in \{1, ..., m\}$ vrijedi*

$$\mathbb{P}[\epsilon_i = 1] = \mathbb{P}[\epsilon_i = -1] = \frac{1}{2}\,.$$

Za $a = (a_1, ..., a_m) \in \mathcal{A}$ definiramo $\|a\| := \sqrt{a_1^2 + \cdots + a_m^2}$. Neka je $r = \sup_{a\in\mathcal{A}} \|a\|$. Tada vrijedi

$$\mathbb{E}\left[\sup_{a\in\mathcal{A}} \frac{1}{m}\sum_{i=1}^{m}\epsilon_i a_i\right] \leq \frac{r\sqrt{2\ln|\mathcal{A}|}}{m}\,.$$

Dokaz. Definirajmo μ kao

$$\mu := \mathbb{E}\left[\sup_{a\in\mathcal{A}}\sum_{i=1}^{m}\epsilon_i a_i\right].$$

Za konveksnu funkciju φ i slučajnu varijablu X Jensenova nejednakost nam kaže da vrijedi

$$\varphi^{\mathbb{E}[X]} \leq \mathbb{E}\left[\varphi^{X}\right]. \tag{5.3}$$

Kako je eksponencijalna funkcija konveksna, korištenjem (5.3) i nezavisnosti slučajnih varijabli $\epsilon_1, ..., \epsilon_m$ dobijemo da za svaki $\lambda > 0$ vrijedi

$$\begin{aligned} e^{\lambda\mu} &\leq \mathbb{E}\left[\exp\left(\lambda \sup_{a\in\mathcal{A}} \sum_{i=1}^{m} \epsilon_i a_i\right)\right] \\ &= \mathbb{E}\left[\sup_{a\in\mathcal{A}} \exp\left(\lambda \sum_{i=1}^{m} \epsilon_i a_i\right)\right] \\ &\leq \mathbb{E}\left[\sum_{a\in\mathcal{A}} \exp\left(\lambda \sum_{i=1}^{m} \epsilon_i a_i\right)\right] \\ &= \sum_{a\in\mathcal{A}} \mathbb{E}\left[\exp\left(\lambda \sum_{i=1}^{m} \epsilon_i a_i\right)\right] \\ &= \sum_{a\in\mathcal{A}} \prod_{i=1}^{m} \mathbb{E}\left[\exp\left(\lambda\epsilon_i a_i\right)\right] . \end{aligned}$$

Za svaki $i \in \{1, ..., m\}$ vrijedi $\mathbb{P}[\epsilon_i = 1] = \mathbb{P}[\epsilon_i = -1] = 1/2$, iz čega slijedi

$$\mathbb{E}\left[\exp\left(\lambda\epsilon_i a_i\right)\right] = \mathbb{P}[\epsilon_i = 1]e^{\lambda a_i} + \mathbb{P}[\epsilon_i = -1]e^{-\lambda a_i} = \frac{e^{\lambda a_i} + e^{-\lambda a_i}}{2},$$

pa korištenjem nejednakosti $(e^x + e^{-x})/2 \leq e^{x^2/2}$ dobijemo

$$\begin{aligned} e^{\lambda\mu} &\leq \sum_{a\in\mathcal{A}} \prod_{i=1}^{m} \mathbb{E}\left[\exp\left(\lambda\epsilon_i a_i\right)\right] \\ &= \sum_{a\in\mathcal{A}} \prod_{i=1}^{m} \frac{e^{\lambda a_i} + e^{-\lambda a_i}}{2} \\ &\leq \sum_{a\in\mathcal{A}} \prod_{i=1}^{m} e^{\lambda^2 a_i^2/2} \\ &= \sum_{a\in\mathcal{A}} e^{\lambda^2 \|a\|^2/2} \\ &\leq |\mathcal{A}|\, e^{\lambda^2 r^2/2} . \end{aligned}$$

Logaritmiranjem i dijeljenjem s λ dobijemo da za svaki $\lambda > 0$ vrijedi

$$\mu \leq \frac{\ln|\mathcal{A}|}{\lambda} + \frac{\lambda r^2}{2} .$$

Uzimanjem $\lambda = \sqrt{2\ln|\mathcal{A}|/r^2}$ slijedi tvrdnja leme, tj. vrijedi

$$\mu \leq r\sqrt{2\ln|\mathcal{A}|}\,.$$

□

Napomena 5.2.2. *Neka vrijede pretpostavke Leme 5.2.1. Za $\mathcal{A} \subset \mathbb{R}^m$ definiramo skup $-\mathcal{A}$ kao*

$$-\mathcal{A} := \{-a = (-a_1, ..., -a_m) : a \in \mathcal{A}\}\,.$$

Tada vrijedi

$$\begin{aligned}\mathbb{E}\left[\sup_{a\in\mathcal{A}}\left|\frac{1}{m}\sum_{i=1}^{m}\epsilon_i a_i\right|\right] &= \mathbb{E}\left[\sup_{a\in\mathcal{A}}\max\left\{\frac{1}{m}\sum_{i=1}^{m}\epsilon_i a_i, -\frac{1}{m}\sum_{i=1}^{m}\epsilon_i a_i\right\}\right] \\ &= \mathbb{E}\left[\sup_{a\in\mathcal{A}}\max\left\{\frac{1}{m}\sum_{i=1}^{m}\epsilon_i a_i, \frac{1}{m}\sum_{i=1}^{m}\epsilon_i(-a_i)\right\}\right] \\ &= \mathbb{E}\left[\sup_{a\in\mathcal{A}\cup-\mathcal{A}}\frac{1}{m}\sum_{i=1}^{m}\epsilon_i a_i\right].\end{aligned}$$

Za svaki $a \in \mathcal{A}$ vrijedi $\|a\| = \|-a\|$, iz čega slijedi

$$\sup_{a\in\mathcal{A}\cup-\mathcal{A}}\|a\| = \sup_{a\in\mathcal{A}}\|a\| = r\,.$$

Kako je $|\mathcal{A}\cup-\mathcal{A}| \leq 2|\mathcal{A}|$, korištenjem Leme 5.2.1 dobijemo

$$\mathbb{E}\left[\sup_{a\in\mathcal{A}}\left|\frac{1}{m}\sum_{i=1}^{m}\epsilon_i a_i\right|\right] \leq \frac{r\sqrt{2\ln|\mathcal{A}\cup-\mathcal{A}|}}{m} \leq \frac{r\sqrt{2\ln(2|\mathcal{A}|)}}{m}\,. \tag{5.4}$$

Sljedećim teoremom pokazujemo da ako funkcija rasta $\tau_{\mathcal{H}}(m)$ raste polinomijalno s $m \in \mathbb{N}$, onda klasa $\mathcal{H}$ zadovoljava svojstvo uniformne konvergencije.

Teorem 5.2.3. *Neka je $\mathcal{D}$ distribucija nad $\mathcal{X}\times\{0,1\}$, $\mathcal{H}$ klasa hipoteza s domenom $\mathcal{X}$ i kodomenom $\{0,1\}$ i neka je $\tau_{\mathcal{H}}$ pripadajuća funkcija rasta. Tada za svaku distribuciju $\mathcal{D}$, za svaki $m \in \mathbb{N}$ te za svaki $\delta \in (0,1)$ uz 0-1 funkciju gubitka vrijedi*

$$\mathop{\mathbb{P}}_{S\sim\mathcal{D}^m}\left[\sup_{h\in\mathcal{H}}\left|L_{\mathcal{D}}(h) - L_S(h)\right| \leq \frac{2\sqrt{2\ln(2\tau_{\mathcal{H}}(m))}}{\delta\sqrt{m}}\right] \geq 1-\delta\,.$$

Dokaz. Slučajna varijabla $\sup_{h\in\mathcal{H}}\left|L_{\mathcal{D}}(h) - L_S(h)\right|$ je nenegativna, stoga ako pokažemo da vrijedi

$$\mathop{\mathbb{E}}_{S\sim\mathcal{D}^m}\left[\sup_{h\in\mathcal{H}}\left|L_{\mathcal{D}}(h) - L_S(h)\right|\right] \leq \frac{2\sqrt{2\ln\left(2\tau_{\mathcal{H}}(m)\right)}}{\sqrt{m}}\,, \tag{5.5}$$

onda dokaz teorema slijedi primjenom Markovljeve nejednakosti, tj. primjenom Leme 3.1.1.

Neka je $S' = \{z'_1, ..., z'_m\}$ uzorak iz $\mathcal{D}^m$ nezavisan od S. Tada za svaki $h \in \mathcal{H}$ vrijedi

$$L_{\mathcal{D}}(h) = \mathop{\mathbb{E}}_{S'\sim\mathcal{D}^m}\left[L_{S'}(h)\right]\,,$$

iz čega slijedi

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

Primjenom Jensenove nejednakosti na $|\cdot|$ dobijemo

$$\left|\mathop{\mathbb{E}}_{S'\sim\mathcal{D}^m}\left[L_{S'}(h) - L_S(h)\right]\right| \leq \mathop{\mathbb{E}}_{S'\sim\mathcal{D}^m}\left|L_{S'}(h) - L_S(h)\right|\,. \tag{5.6}$$

Za svaki $h' \in \mathcal{H}$ vrijedi

$$\left|L_{S'}(h') - L_S(h')\right| \leq \sup_{h\in\mathcal{H}}\left|L_{S'}(h) - L_S(h)\right|\,. \tag{5.7}$$

Uzimanjem očekivanja dobijemo

$$\mathop{\mathbb{E}}_{S'\sim\mathcal{D}^m}\left|L_{S'}(h') - L_S(h')\right| \leq \mathop{\mathbb{E}}_{S'\sim\mathcal{D}^m}\sup_{h\in\mathcal{H}}\left|L_{S'}(h) - L_S(h)\right|\,, \tag{5.8}$$

iz čega slijedi

$$\sup_{h\in\mathcal{H}}\mathop{\mathbb{E}}_{S'\sim\mathcal{D}^m}\left|L_{S'}(h) - L_S(h)\right| \leq \mathop{\mathbb{E}}_{S'\sim\mathcal{D}^m}\sup_{h\in\mathcal{H}}\left|L_{S'}(h) - L_S(h)\right|\,. \tag{5.9}$$

Iz (5.6) i (5.9) uz primjenu Fubinijevog teorema i nezavisnosti od S i S' dobijemo

$$\begin{aligned}\mathop{\mathbb{E}}_{S\sim\mathcal{D}^m}\left[\sup_{h\in\mathcal{H}}\left|L_{\mathcal{D}}(h) - L_S(h)\right|\right] &\leq \mathop{\mathbb{E}}_{S,S'\sim\mathcal{D}^m}\left[\sup_{h\in\mathcal{H}}\left|L_{S'}(h) - L_S(h)\right|\right] \\ &= \mathop{\mathbb{E}}_{S,S'\sim\mathcal{D}^m}\left[\sup_{h\in\mathcal{H}}\frac{1}{m}\left|\sum_{i=1}^{m}\left(l(h, z'_i) - l(h, z_i)\right)\right|\right]\,.\end{aligned} \tag{5.10}$$

Elementi od S i S' su n.j.d., a S i S' su međusobno nezavisni skupovi, stoga $l(h, z_i') - l(h, z_i)$ iz (5.10) možemo zamijeniti sa $-(l(h, z_i') - l(h, z_i))$. Naime, kako su $z_1, ..., z_m, z_1', ..., z_m'$ n.j.d., distribucija ovog niza se ne mijenja ako ga ispermutiramo na bilo koji način. Iz toga slijedi da za svaki $\sigma = (\sigma_1, ..., \sigma_m) \in \{\pm 1\}^m$ vrijedi

$$\mathop{\mathbb{E}}_{S,S'\sim\mathcal{D}^m}\left[\sup_{h\in\mathcal{H}}\frac{1}{m}\left|\sum_{i=1}^{m}(l(h,z_i')-l(h,z_i))\right|\right] = \mathop{\mathbb{E}}_{S,S'\sim\mathcal{D}^m}\left[\sup_{h\in\mathcal{H}}\frac{1}{m}\left|\sum_{i=1}^{m}\sigma_i(l(h,z_i)-l(h,z_i'))\right|\right]. \quad (5.11)$$

Neka je $U_\pm$ uniformna distribucija nad $\{\pm 1\}$. Kako (5.11) vrijedi za svaki $\sigma \in \{\pm 1\}^m$, uz korištenje Fubinijevog teorema i nezavisnosti od S, S' i σ dobijemo

$$\begin{aligned}
&\mathop{\mathbb{E}}_{S,S'\sim\mathcal{D}^m}\left[\sup_{h\in\mathcal{H}}\frac{1}{m}\left|\sum_{i=1}^{m}\sigma_i(l(h,z_i)-l(h,z_i'))\right|\right]\\
&= \mathop{\mathbb{E}}_{\sigma\sim U_\pm^m}\mathop{\mathbb{E}}_{S,S'\sim\mathcal{D}^m}\left[\sup_{h\in\mathcal{H}}\frac{1}{m}\left|\sum_{i=1}^{m}\sigma_i(l(h,z_i)-l(h,z_i'))\right|\right]\\
&= \mathop{\mathbb{E}}_{S,S'\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{\sigma\sim U_\pm^m}\left[\sup_{h\in\mathcal{H}}\frac{1}{m}\left|\sum_{i=1}^{m}\sigma_i(l(h,z_i)-l(h,z_i'))\right|\right]\\
&\leq \mathop{\mathbb{E}}_{S,S'\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{\sigma\sim U_\pm^m}\left[\sup_{h\in\mathcal{H}}\frac{1}{m}\left|\sum_{i=1}^{m}\sigma_i l(h,z_i)\right|\right] + \mathop{\mathbb{E}}_{S,S'\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{\sigma\sim U_\pm^m}\left[\sup_{h\in\mathcal{H}}\frac{1}{m}\left|\sum_{i=1}^{m}\sigma_i l(h,z_i')\right|\right]. \quad (5.12)
\end{aligned}$$

Kako $\sup_{h\in\mathcal{H}}\frac{1}{m}\left|\sum_{i=1}^{m}\sigma_i l(h,z_i)\right|$ ne ovisi o S', a $\sup_{h\in\mathcal{H}}\frac{1}{m}\left|\sum_{i=1}^{m}\sigma_i l(h,z_i')\right|$ ne ovisi o S, slijedi

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

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

Skupovi S i S' su jednako distribuirani, pa (5.12) možemo zapisati kao

$$\begin{aligned}
&\mathop{\mathbb{E}}_{S,S'\sim\mathcal{D}^m}\left[\sup_{h\in\mathcal{H}}\frac{1}{m}\left|\sum_{i=1}^{m}\sigma_i(l(h,z_i)-l(h,z_i'))\right|\right]\\
&\leq \mathop{\mathbb{E}}_{S,S'\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{\sigma\sim U_\pm^m}\left[\sup_{h\in\mathcal{H}}\frac{1}{m}\left|\sum_{i=1}^{m}\sigma_i l(h,z_i)\right|\right]+\mathop{\mathbb{E}}_{S,S'\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{\sigma\sim U_\pm^m}\left[\sup_{h\in\mathcal{H}}\frac{1}{m}\left|\sum_{i=1}^{m}\sigma_i l(h,z_i')\right|\right]\\
&= \mathop{\mathbb{E}}_{S\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{\sigma\sim U_\pm^m}\left[\sup_{h\in\mathcal{H}}\frac{1}{m}\left|\sum_{i=1}^{m}\sigma_i l(h,z_i)\right|\right]+\mathop{\mathbb{E}}_{S'\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{\sigma\sim U_\pm^m}\left[\sup_{h\in\mathcal{H}}\frac{1}{m}\left|\sum_{i=1}^{m}\sigma_i l(h,z_i')\right|\right]\\
&= 2\mathop{\mathbb{E}}_{S\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{\sigma\sim U_\pm^m}\left[\sup_{h\in\mathcal{H}}\frac{1}{m}\left|\sum_{i=1}^{m}\sigma_i l(h,z_i)\right|\right].
\end{aligned}$$

Neka je $S = \{z_1, ..., z_m\}$ fiksan skup takav da $\forall i \in \{1, ..., m\}$ vrijedi $z_i = (x_i, y_i)$, gdje je $x_i \in \mathcal{X}$, a $y_i \in \{0, 1\}$. Definirajmo skup C sa $C := \{x_1, ..., x_m\}$. 0-1 funkcija gubitka je dana s

$$l(h, z_i) = l(h, (x_i, y_i)) = \mathbb{1}_{\{h(x_i)\neq y_i\}},$$

a kako je za fiksan S bitno samo kako h djeluje na elemente od C, onda vrijedi

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

Primjernom Massartove leme, točnije Napomene 5.2.2, na nezavisne slučajne varijable $\sigma_1, ..., \sigma_m$ i na skup $\mathcal{A}$ dan s

$$\mathcal{A} = \{(l(h, z_1), ..., l(h, z_m)) : h \in \mathcal{H}_C\},$$

dobijemo

$$\mathop{\mathbb{E}}_{\sigma\sim U_\pm^m}\left[\sup_{h\in\mathcal{H}_C}\frac{1}{m}\left|\sum_{i=1}^{m}\sigma_i l(h,z_i))\right|\right] \leq \frac{r\sqrt{2\ln(2|\mathcal{H}_C|)}}{m}. \tag{5.13}$$

Za svaki $i \in \{1, .., m\}$ te za svaki $h \in \mathcal{H}_C$ vrijedi $l(h, z_i) \in \{0, 1\}$, stoga za svaki $h \in \mathcal{H}_C$ imamo

$$\left\|(l(h, z_1), ..., l(h, z_m))\right\| \leq \sqrt{m},$$

iz čega slijedi

$$r = \sup_{a\in\mathcal{A}}\|a\| \leq \sqrt{m}.$$

Uvrštavanjem u (5.13) dobijemo

$$\mathop{\mathbb{E}}_{\sigma\sim U_{\pm}^{m}}\left[\sup_{h\in\mathcal{H}_C}\frac{1}{m}\left|\sum_{i=1}^{m}\sigma_i l(h,z_i))\right|\right]\leq\frac{\sqrt{2\ln\left(2\left|\mathcal{H}_C\right|\right)}}{\sqrt{m}}. \tag{5.14}$$

Iz definicije funkcije rasta vrijedi $\left|\mathcal{H}_C\right|\leq\tau_{\mathcal{H}}(m)$, stoga

$$\mathop{\mathbb{E}}_{\sigma\sim U_{\pm}^{m}}\left[\sup_{h\in\mathcal{H}_C}\frac{1}{m}\left|\sum_{i=1}^{m}\sigma_i l(h,z_i))\right|\right]\leq\frac{\sqrt{2\ln\left(2\tau_{\mathcal{H}}(m)\right)}}{\sqrt{m}}. \tag{5.15}$$

Kako desna strana u (5.15) ne ovisi o S, kombiniranjem svih prethodnih nejednakosti te uzimanjem očekivanja po $S\sim\mathcal{D}^m$ dobijemo

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

□

Nakon iskazivanja i dokazivanja svih potrebnih lema, možemo dokazati *Fundamentalni teorem statističkog učenja* kojim povezujemo mogućnost PAC učenja s VC dimenzijom i svojstvom uniformne konvergencije.

Teorem 5.2.4. *(Fundamentalni teorem statističkog učenja) Neka je $\mathcal{H}$ klasa hipoteza s domenom $\mathcal{X}$ i kodomenom $\{0,1\}$. Uz 0-1 funkciju gubitka, sljedeće tvrdnje su ekvivalentne:*

1. *$\mathcal{H}$ ima svojstvo uniformne konvergencije*
2. *$\mathcal{H}$ možemo naučiti u agnostičkom PAC smislu korištenjem ERM algoritma*
3. *$\mathcal{H}$ možemo naučiti u agnostičkom PAC smislu*
4. *$\mathcal{H}$ možemo naučiti u PAC smislu*
5. *$\mathcal{H}$ možemo naučiti u PAC smislu korištenjem ERM algoritma*
6. *$\mathcal{H}$ ima konačnu VC dimenziju .*

Dokaz. U Korolaru 4.1.3 pokazali smo da vrijedi $1\rightarrow 2$. Tvrdnje $2\rightarrow 3$, $2\rightarrow 5$ i $3\rightarrow 4$ su trivijalne, pa njih nećemo dokazivati, a tvrdnje $4\rightarrow 6$ i $5\rightarrow 6$ slijede iz Korolara

3.3.5. Stoga nam preostaje dokazati da vrijedi 6 $\rightarrow$ 1. Dovoljno je dokazati da ako je VC-dimenzija konačna, onda klasa $\mathcal{H}$ zadovoljava svojstvo uniformne konvergencije.

Iz Sauerove leme slijedi da za $m > d$ vrijedi $\tau_{\mathcal{H}}(m) \leq (em/d)^d$. Koristeći Teorem 5.2.3 slijedi da s vjerojatnosti od barem $1 - \delta$ za svaki $h \in \mathcal{H}$ vrijedi

$$\left|L_S(h) - L_{\mathcal{D}}(h)\right| \leq \frac{2\sqrt{2\ln\left(2\tau_{\mathcal{H}}(m)\right)}}{\delta\sqrt{m}} \leq \frac{2\sqrt{2\ln\left(2(em/d)^d\right)}}{\delta\sqrt{m}}\,. \tag{5.16}$$

Kako je $2 \leq 2^d$, nejednakost (5.16) možemo zapisati kao

$$\left|L_S(h) - L_{\mathcal{D}}(h)\right| \leq \frac{2\sqrt{2d\ln\left(2em/d\right)}}{\delta\sqrt{m}}\,.$$

Bitno je uočiti da za fiksne ε i δ vrijedi

$$\lim_{m\to\infty} \frac{2\sqrt{2d\ln\left(2em/d\right)}}{\delta\sqrt{m}} = 0\,,$$

pa će za svaki dovoljno veliki m vrijediti

$$\frac{2\sqrt{2d\ln\left(2em/d\right)}}{\delta\sqrt{m}} \leq \varepsilon\,.$$

Stoga, da bi $\left|L_S(h) - L_{\mathcal{D}}(h)\right|$ bilo odozgo ograničeno s ϵ dovoljno je da vrijedi

$$m \geq \frac{8d\ln(m)}{(\varepsilon\delta)^2} + \frac{8d\ln(2e/d)}{(\varepsilon\delta)^2}\,.$$

Primjenom Leme 5.3.2 slijedi da je dovoljno da za m vrijedi

$$m \geq \frac{32d}{(\varepsilon\delta)^2}\ln\left(\frac{16d}{(\varepsilon\delta)^2}\right) + \frac{16d}{(\varepsilon\delta)^2}\ln\left(\frac{2e}{d}\right)\,.$$

□

Postoji još jedna verzija *Fundamentalnog teorema statističkog učenja*, a to je njegova kvantitativna verzija u kojoj su dane gornje i donje ograde na funkciju $m_{\mathcal{H}}^{UC}(\varepsilon, \delta)$.

Teorem 5.2.5. *(Fundamentalni teorem statističkog učenja-Kvantitativna verzija) Neka je $\mathcal{H}$ klasa hipoteza s domenom $\mathcal{X}$ i kodomenom $\{0, 1\}$. Ako vrijedi da je $VCdim(\mathcal{H}) = d < \infty$, onda za 0-1 funkciju gubitka postoje konstante C_1 i C_2 tako da vrijedi*

1. *$\mathcal{H}$ ima svojstvo uniformne konvergencije t.d. za složenost vrijedi*

$$C_1 \frac{d + \ln(1/\delta)}{\varepsilon^2} \leq m_{\mathcal{H}}^{UC}(\varepsilon, \delta) \leq C_2 \frac{d + \ln(1/\delta)}{\varepsilon^2}$$

2. *$\mathcal{H}$ možemo naučiti u agnostičkom PAC smislu t.d. za složenost vrijedi*

$$C_1 \frac{d + \ln(1/\delta)}{\varepsilon^2} \leq m_{\mathcal{H}}(\varepsilon, \delta) \leq C_2 \frac{d + \ln(1/\delta)}{\varepsilon^2}$$

3. *$\mathcal{H}$ možemo naučiti u PAC smislu t.d. za složenost vrijedi*

$$C_1 \frac{d + \ln(1/\delta)}{\varepsilon} \leq m_{\mathcal{H}}(\varepsilon, \delta) \leq C_2 \frac{d \ln(1/\varepsilon) + \ln(1/\delta)}{\varepsilon}\,.$$

Za dokaz ovog teorema potrebni su pojmovi koji neće biti obrađeni u ovom radu, stoga se navodi iskaz bez dokaza.

5.3 Dodatak

Slijede iskazi i dokazi dviju tehničkih lema korištenih pri dokazivanju *Fundamentalnog teorema statističkog učenja*.

Lema 5.3.1. *Neka je $a > 0$. Ako vrijedi*

$$x \geq 2a \ln(a)\,,$$

onda

$$x \geq a \ln(x)\,.$$

Dokaz. Ako je $a \in (0, \sqrt{e}]$, tada $x \geq a \ln(x)$ vrijedi za svaki x. Stoga pretpostavimo da je $a > \sqrt{e}$. Definiramo funkciju $f : \mathbb{R} \to \mathbb{R}$ s $f(x) := x - a \ln(x)$. Kako je $f'(x) = 1 - a/x$, slijedi da je za $x > a$ $f'(x) > 0$ i funkcija f raste. Dodatno vrijedi

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

Kako za svaki a vrijedi $a - 2 \ln(a) > 0$, slijedi $f(2a \ln(a)) > 0$, iz čega slijedi tvrdnja leme. □

Lema 5.3.2. *Neka je $a \geq 1$ i $b > 0$. Ako vrijedi*

$$x \geq 4a\ln(2a) + 2b\,,$$

onda

$$x \geq a\ln(x) + b\,.$$

Dokaz. Dovoljno je dokazati da ako vrijedi $x \geq 4a\ln(2a) + 2b$ onda vrijedi $x \geq 2a\ln(x)$ te $x \geq 2b$. Kako je $a > 1$, iz $x \geq 4a\ln(2a)$ slijedi $x > 2b$. Također, kako je $b > 0$, iz $x \geq 4a\ln(2a)$ slijedi $x > 4a\ln(2a)$, iz čega korištenjem Leme 5.3.1 slijedi $x > 2a\ln(x)$. □

Poglavlje 6

Poluprostori kao linearni prediktori

Linearni prediktori su jedna od najkorištenijih familija klasa hipoteza. Jednostavni su za korištenje, lako ih je interpretirati, a u problemima učenja su se pokazali kao familija koja u većini slučajeva jako dobro opisuje ponašanje podataka. U ovom poglavlju usredotočit ćemo se na jedan dio te familije, a to je klasa poluprostora koja se koristi za binarnu klasifikaciju podataka gdje funkcija cilja poprima vrijednosti iz skupa $\{-1, +1\}$.

6.1 Poluprostori

Definirajmo klasu afinih funkcija sa

$$\mathcal{L}_d = \{h_{w,b} : w \in \mathbb{R}^d, b \in \mathbb{R}\},$$

gdje je $h_{w,b} : \mathbb{R}^d \to \mathbb{R}$ zadana sa

$$h_{w,b}(x) = \langle w, x \rangle + b\,,$$

a funkcija $\langle \cdot\,, \cdot \rangle : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$, odnosno skalarni produkt na $\mathbb{R}^d$, sa

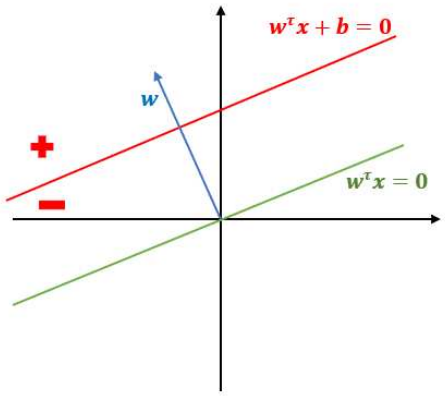
$$\langle w, x \rangle = \sum_{i=1}^{d} w_i x_i\,.$$

Kako poluprostore koristimo za probleme binarne klasifikacije, klasu hipoteza poluprostora možemo definirati kao

$$\mathcal{H}_{HS} = \operatorname{sign} \circ \mathcal{L}_d = \{\operatorname{sign} \circ h_{w,b} : h_{w,b} \in \mathcal{L}_d\}\,,$$

gdje je funkcija $\operatorname{sign} : \mathbb{R} \to \{\pm 1\}$ definirana kao

$$\operatorname{sign}(x) = \begin{cases} -1, & x < 0 \\ 1, & x \geq 0 \end{cases}.$$



Slika 6.1: Poluprostori i razdvajajuća hiperravnina u $\mathbb{R}^2$

U 3. poglavlju pokazali smo da je VC dimenzija poluprostora u ravnini jednaka 3. Sada ćemo pokazati da za proizvoljan $d \in \mathbb{N}$ vrijedi da je VC dimenzija poluprostora u $\mathbb{R}^d$ jednaka $d + 1$, a za to će nam biti potreban Radonov teorem čiji se dokaz može pronaći u [8].

Teorem 6.1.1. *(Radonov teorem) Neka je $\mathcal{X}$ skup koji sadrži $d + 2$ točke iz $\mathbb{R}^d$ za neki $d \in \mathbb{N}$. Skup $\mathcal{X}$ možemo particionirati na dva skupa, $\mathcal{X}_1$ i $\mathcal{X}_2$, tako da se konveksne ljuske tih skupova sijeku.*

Teorem 6.1.2. *VC dimenzija poluprostora u $\mathbb{R}^d$ je $d + 1$.*

Dokaz. Neka je $\mathcal{H}_{HS}$ klasa svih poluprostora u $\mathbb{R}^d$. Neka je $\mathcal{X} = \{x_0, ..., x_d\}$ skup točaka iz $\mathbb{R}^d$ takav da je x_0 ishodište te za svaki $i \in \{1, ..., d\}$ vrijedi da x_i na svim koordinatama ima vrijednost 0, osim na i-toj na kojoj ima vrijednost 1. Neka je $\{y_0, ..., y_d\}$ proizvoljan skup oznaka iz $\{-1, +1\}$, tj. za svaki $i \in \{0, ..., d\}$ vrijedi $y_i \in \{-1, +1\}$. Definirajmo vektor w sa

$$w = (y_1, ..., y_d)\,.$$

Za svaki $i \in \{0, ..., d\}$ vrijedi

$$\operatorname{sign}\left(\langle w, x_i\rangle + \frac{y_0}{2}\right) = \operatorname{sign}\left(y_i + \frac{y_0}{2}\right) = y_i\,. \tag{6.1}$$

Naime, za $i \in \{1, ..., d\}$, iz definicije od x_i i w slijedi $\langle w, x_i\rangle = y_i$. Kako je $y_i \in \{-1, +1\}$, a $y_0/2 \in \{-1/2, +1/2\}$, slijedi

$$\operatorname{sign}\left(y_i + \frac{y_0}{2}\right) = \operatorname{sign}(y_i) = y_i\,,$$

odnosno (6.1) vrijedi za $i \in \{1, ..., d\}$.

Promotrimo slučaj kada je $i = 0$. Kako je $x_0 = (0, ..., 0)$, slijedi $\langle w, x_0 \rangle = 0$. Dodatno vrijedi

$$\text{sign}\left(\frac{y_0}{2}\right) = y_0 \,,$$

što znači da i u slučaju $i = 0$ vrijedi (6.1).

Slijedi da hiperravnina $\langle w, x \rangle + \frac{y_0}{2} = 0$ razdvaja pozitivne i negativne primjere, pa zaključujemo da skup $\mathcal{X}$ možemo rastaviti klasom $\mathcal{H}_{HS}$. Iz ovog slijedi da je VC dimenzija barem $d + 1$.

Neka je $\mathcal{X}$ skup od $d + 2$ točke iz $\mathbb{R}^d$ i pretpostavimo da $\mathcal{X}$ možemo rastaviti klasom $\mathcal{H}_{HS}$. Primjenom Radonovog teorema slijedi da $\mathcal{X}$ možemo particionirati na dva skupa, $\mathcal{X}_1$ i $\mathcal{X}_2$, tako da se konveksne ljuske ta dva skupa sijeku. Kada bi skupove $\mathcal{X}_1$ i $\mathcal{X}_2$ mogli razdvojiti hiperravninom, onda bi i njihove konveksne ljuske također bile razdvojene tom istom hiperravninom. Stoga vrijedi da ta dva skupa ne možemo razdvojiti hiperravninom, iz čega slijedi da skup $\mathcal{X}$ ne možemo rastaviti klasom hipoteza $\mathcal{H}_{HS}$. Naime, kako $\mathcal{X}$ možemo rastaviti klasom $\mathcal{H}_{HS}$, postoji poluravnina koja pravilno klasificira sve točke tako da se točke prvog skupa nalaze s jedne strane, a točke drugog skupa s druge strane pripadajuće hiperavnine. Stavimo stoga oznaku 1 na sve točke iz $\mathcal{X}_1$, a oznaku -1 na sve točke iz $\mathcal{X}_2$. Tada se konveksne ljuske tih skupova nalaze na suprotnim stranama te hiperavnine, tj. ne sijeku se, što je kontradikcija s Radonovim teoremom. Stoga zaključujemo da je VC dimenzija poluprostora u $\mathbb{R}^d$ jednaka $d + 1$. □

Kako je VC dimenzija poluprostora u $\mathbb{R}^d$ konačna, iz Teorema 5.2.4 slijedi da klasu $\mathcal{H}_{HS}$ možemo naučiti u PAC smislu korištenjem ERM algoritma. U nastavku poglavlja pretpostavljat ćemo da postoji hipoteza iz $\mathcal{H}_{HS}$ koja svakoj točki iz S pridružuje točnu oznaku i da primjere možemo razdvojiti hiperavninom koja prolazi kroz ishodište, tj hiperravninom za koju je $b = 0$.

6.2 Linearno programiranje poluprostora

U probleme linearnog programiranja (LP problemi) spadaju optimizacijski problemi u kojima se linearna funkcija cilja mora maksimizirati uz uvjete ili ograničenja dana u obliku linearnih nejednadžbi. Općenito to možemo zapisati kao

$$\max_{w \in \mathbb{R}^d} \langle u, w \rangle$$
$$\text{uz uvjet } Aw \geq v \,,$$

gdje su $u \in \mathbb{R}^d$, $v \in \mathbb{R}^m$ i $A \in M_{m \times d}$.

Pokušajmo ERM algoritam za poluprostore zapisati kao problem linearnog programiranja. Neka je $S = \{(x_1, y_1), ..., (x_m, y_m)\}$ skup za učenje od m elemenata. Kako za svaki $i \in \{1, ..., m\}$ vrijedi $y_i \in \{-1, +1\}$, zapravo želimo pronaći vektor $w \in \mathbb{R}^d$ takav da vrijedi

$$\text{sign}\left(\langle w, x_i \rangle\right) = y_i, \quad \forall i = 1, ..., m\,, \tag{6.2}$$

što je ekvivalentno traženju $w \in \mathbb{R}^d$ tako da vrijedi

$$y_i \langle w, x_i \rangle > 0, \quad \forall i = 1, ..., m\,. \tag{6.3}$$

Uočite da sa (6.3) pretpostavljamo da postoji w takav da $\langle w, x_i \rangle$ nije 0 za sve i, tj. da postoji razdvajajuća hiperavnina, ali takva da se niti jedna točka ne nalazi točno na hiperavnini. Neka je $w_1 \in \mathbb{R}^d$ vektor koji zadovoljava (6.3). Definirajmo w_2 kao

$$w_2 = \frac{w_1}{\min\limits_{i \in \{1,...,m\}} \left(y_i \langle w, x_i \rangle\right)}\,.$$

Tada za svaki $i \in \{1, ..., m\}$ vrijedi

$$y_i \langle w_2, x_i \rangle = \frac{1}{\min\limits_{i \in \{1,...,m\}} \left(y_i \langle w, x_i \rangle\right)} y_i \langle w_1, x_i \rangle \geq 1\,, \tag{6.4}$$

odnosno, pokazali smo da postoji ERM prediktor takav da vektor w zadovoljava

$$y_i \langle w, x_i \rangle \geq 1\,, \quad \forall i = 1, ..., m\,.$$

Ako za svaki $i \in \{1, ..., m\}$ vektor x_i napišemo kao $x_i = (x_{i,1}, ..., x_{i,d})$, onda vektor w možemo pronaći rješavanjem LP problema danog sa

$$\begin{aligned} &\max_{w \in \mathbb{R}^d} \langle w, (0, ..., 0) \rangle \\ &\text{t.d. vrijedi } Aw \geq v\,, \end{aligned} \tag{6.5}$$

gdje je $v = (1, ..., 1) \in \mathbb{R}^m$, a matrica A dana sa

$$A = \begin{bmatrix} y_1 x_{1,1} & \dots & y_1 x_{1,d} \\ y_2 x_{2,1} & \dots & y_2 x_{2,d} \\ \vdots & \vdots & \vdots \\ y_m x_{m,1} & \dots & y_m x_{m,d} \end{bmatrix}.$$

6.3 Perceptroni za poluprostore

Nešto drugačija implementacija ERM algoritma dana je u obliku iterativnog algoritma, Perceptron algoritma ili PLA, zadanog algoritmom ispod.

Perceptron algoritam

Ulazni podaci: $(x_1, y_1), ..., (x_m, y_m)$
Inicijalizacija: $w^{(1)} = (0, ..., 0)$
Za $t = 1, 2, ...$:
  ako ($\exists\, i$ t.d. $y_i\langle w, x_i\rangle \leq 0$) onda
    $w^{(t+1)} = w^{(t)} + y_i x_i$
  inače
    ispiši $w^{(t)}$

PLA algoritam stvara niz vektora $w^{(1)}, w^{(2)}, ...$ tako da u iteraciji t pronalazi primjer i koji $w^{(t)}$ krivo označava, tj. za koji vrijedi $\text{sign}(\langle w^{(t)}, x_i\rangle) \neq y_i$. Nakon toga ažurira vektor $w^{(t)}$ definirajući $w^{(t+1)}$ kao $w^{(t+1)} = w^{(t)} + y_i x_i$, jer na taj način dolazimo do rješenja koje je točnije na primjeru i. To jest, vrijedi

$$y_i\langle w^{(t+1)}, x_i\rangle = y_i\langle w^{(t)} + y_i x_i, x_i\rangle = y_i\langle w^{(t)}, x_i\rangle + \|x_i\|^2 \ .$$

Sljedeći teorem nam garantira da algoritam staje nakon što su sve točke skupa za učenje točno označene.

Teorem 6.3.1. *Neka je $S = \{(x_1, y_1), ..., (x_m, y_m)\}$ skup za učenje od m elemenata takav da postoji hipoteza iz $\mathcal{H}_{HS}$ koja svakoj točki iz S pridružuje točnu oznaku. Neka je B definiran sa*

$$B = \min\{\|w\| : w \in \mathbb{R}^d, \langle w, x_i\rangle y_i \geq 1, \forall i \in \{1, ..., m\}\},$$

te neka je R definiran sa

$$R = \max_{i\in\{1,...,m\}} \|x_i\| \ .$$

Tada se algoritam za učenje perceptrona zaustavlja nakon najviše $(RB)^2$ iteracija te u trenutku zaustavljanja vrijedi

$$y_i\langle w^{(t)}, x_i\rangle > 0, \quad \forall i \in \{1, ..., m\}.$$

Dokaz. Pokazat ćemo da kosinus kuta između w^* i $w^{(T+1)}$ nakon T iteracija iznosi najmanje $\sqrt{T}/(RB)$, tj. da vrijedi

$$\cos(\angle(w^*, w^{(T+1)})) = \frac{\langle w^*, w^{(T+1)}\rangle}{\|w^*\|\left\|w^{(T+1)}\right\|} \geq \frac{\sqrt{T}}{RB},$$

iz čega korištenjem $\cos(\angle(w^*, w^{(T+1)})) \leq 1$ slijedi

$$T \leq (RB)^2 .$$

Pretpostavimo da je algoritam napravio T iteracija.. Definirajmo vektor w^* kao

$$w^* = \operatorname{argmin} \{\|w\| : \langle w, x_i \rangle y_i \geq 1, \forall i \in \{1, ..., m\}\} .$$

Za $w^{(1)} = (0, ..., 0) \in \mathbb{R}^d$ dobijemo $\langle w^*, w^{(1)} \rangle = 0$, a za iteraciju t, gdje je $w^{(t+1)} = w^{(t)} + y_i x_i$, vrijedi

$$\begin{aligned} \langle w^*, w^{(t+1)} \rangle - \langle w^*, w^{(t)} \rangle &= \langle w^*, w^{(t)} + x_i y_i \rangle - \langle w^*, w^{(t)} \rangle \\ &= \langle w^*, x_i y_i \rangle \\ &\geq 1 . \end{aligned}$$

Slijedi da nakon T iteracija vrijedi

$$\langle w^*, w^{(T+1)} \rangle = \sum_{t=1}^{T} \left(\langle w^*, w^{(t+1)} \rangle - \langle w^*, w^{(t)} \rangle \right) \geq T . \tag{6.6}$$

Kako za $t \leq T$ postoji i takav da vrijedi $y_i \langle w^{(t)}, x_i \rangle \leq 0$, iz definicije od R dobijemo da za svaku iteraciju t vrijedi

$$\begin{aligned} \left\| w^{(t+1)} \right\|^2 &= \left\| w^{(t)} + x_i y_i \right\|^2 \\ &= \left\| w^{(t)} \right\|^2 + 2 y_i \langle w^{(t)}, x_i \rangle + y_i^2 \|x_i\|^2 \\ &\leq \left\| w^{(t)} \right\|^2 + R^2 . \end{aligned} \tag{6.7}$$

Korištenjem definicije od $w^{(1)}$ slijedi $\left\| w^{(1)} \right\|^2 = 0$, pa rekurzivno iz (6.7) nakon T iteracija dobijemo

$$\left\| w^{(T+1)} \right\|^2 \leq T R^2 , \tag{6.8}$$

što je ekvivalentno sa

$$\left\| w^{(T+1)} \right\| \leq \sqrt{T} R . \tag{6.9}$$

Stoga, iz definicije od B te nejednakosti (6.9) i (6.6) dobijemo

$$\frac{\langle w^*, w^{(T+1)} \rangle}{\|w^*\| \left\| w^{(T+1)} \right\|} \geq \frac{T}{B \sqrt{T} R} = \frac{\sqrt{T}}{BR} .$$

□

Napomena 6.3.2. *B smo u Teoremu 6.3.1 definirali sa*

$$B = \min\{\|w\| : w \in \mathbb{R}^d, \langle w, x_i\rangle y_i \geq 1, \forall i \in \{1, ..., m\}\}.$$

Kako zbog $\langle w, x_i\rangle y_i \geq 1$ *slijedi* $\|w\| > 0$*, B možemo zapisati drugačije, tj.*

$$\begin{aligned} B &= \min\{\|w\| : w \in \mathbb{R}^d, \langle w, x_i\rangle y_i \geq 1, \forall i \in \{1, ..., m\}\} \\ &= \min\{\|w\| : w \in \mathbb{R}^d, \langle w, x_i\rangle \frac{y_i}{\|w\|} \geq \frac{1}{\|w\|}, \forall i \in \{1, ..., m\}\} \\ &= \min\{\|w\| : w \in \mathbb{R}^d, \langle \frac{w}{\|w\|}, x_i\rangle y_i \geq \frac{1}{\|w\|}, \forall i \in \{1, ..., m\}\} \\ &= \min\{\tau > 0 : w \in \mathbb{R}^d \text{ t.d. } \|w\| = 1, \langle w, x_i\rangle y_i \geq \frac{1}{\tau}, \forall i \in \{1, ..., m\}\}. \end{aligned}$$

Definirajmo P kao

$$P = B^{-1} = \max\{\tau > 0 : w \in \mathbb{R}^d \text{ t.d. } \|w\| = 1, \langle w, x_i\rangle y_i \geq \tau, \forall i \in \{1, ..., m\}\}.$$

Tada za P vrijedi

$$P = \max_{\|w\|=1} \left(\min_{i=1,...,m} y_i \langle w, x_i\rangle \right).$$

Kako je $\min\limits_{i=1,...,m} y_i\langle w, x_i\rangle = \min\limits_{i=1,...,m} \left|\langle w, x_i\rangle\right|$ *zapravo udaljenost najbliže točke skupa S do hiperravnine određene sa w, slijedi da je P najveća moguća udaljenost koju možemo postići. Kombiniranjem s Teoremom 6.3.1 dobijemo da vrijedi*

$$(RB)^2 = \left(\frac{R}{P}\right)^2.$$

Sada možemo zaključiti da što je P veći, algoritam brže konvergira jer smo u Teoremu 6.3.1 pokazali da je broj iteracija T odozgo ograničen sa $(RB)^2$.

Poglavlje 7

Boosting

Boosting je naziv za metodu koja kombinacijom nekih lošijih prediktora ili pravila koja često vrijede u nekim općenitim slučajevima stvara točan prediktor za početni problem. Iako postoji više Boosting algoritama (XGBoost, Gradient Boosting Machines, LightGBM...), u ovom poglavlju fokusirat ćemo se na AdaBoost algoritam koji kao rezultat daje hipotezu nastalu linearnom kombinacijom nekih jednostavnih hipoteza te nam omogućava kontroliranje greške procjene i greške aproksimacije variranjem samo jednog parametra.

7.1 Slabo učenje

Kako Boosting algoritmi kreću od nekih lošijih prediktora, prirodno se javlja potreba definiranja učenja koje vraća hipotezu koja je samo malo bolja od slučajnog pogađanja.

Definicija 7.1.1. *Neka je $\mathcal{H}$ klasa hipoteza sa $\mathcal{X}$ u $\{-1, 1\}$. Klasu $\mathcal{H}$ moguće je naučiti γ-slabo ako postoji funkcija $m_{\mathcal{H}} : (0,1) \to \mathbb{N}$ i algoritam $\mathcal{A}$ tako da za svaki $\delta \in (0,1)$, za svaku distribuciju $\mathcal{D}$ nad $\mathcal{X}$ te za svaku ciljnu funkciju $f : \mathcal{X} \to \{\pm 1\}$, slijedi da ako je $f \in \mathcal{H}$, korištenjem algoritma $\mathcal{A}$ na skupu za učenje koji se satoji od m nezavisnih primjera generiranih distribucijom $\mathcal{D}$ i označenih s f, gdje je $m \geq m_{\mathcal{H}}(\delta)$, algoritam $\mathcal{A}$ vraća hipotezu h (koja ne mora biti iz klase $\mathcal{H}$) i za koju vrijedi*

$$\mathbb{P}\left[L_{\mathcal{D},f}(h) \leq 1/2 - \gamma\right] \geq 1 - \delta\,.$$

Iako definicija γ-slabog učenja podsjeća na definicju PAC učenja, postoji jedna ključna razlika, a to je da u definiciji PAC učenja algoritam pronalazi hipotezu sa proizvoljno malom greškom, dok je u slučaju slabog učenja dovoljno pronaći hipotezu čija prava greška nije veća od $1/2 - \gamma$ (tj. hipoteza je samo malo bolja od slučajnog pogađanja).

Slabo učenje možemo povezati i sa VC dimenzijom. Naime, Teorem 5.2.5 nam govori da ako klasa $\mathcal{H}$ ima VC dimenziju jednaku d, onda složenost PAC učenja zadovoljava

$$m_{\mathcal{H}}(\varepsilon, \delta) \geq C_1 \frac{d + \log(1/\delta)}{\varepsilon}\,.$$

Ako definiramo $\varepsilon = 1/2 - \gamma$, slijedi da ako je $d = \infty$, onda klasu $\mathcal{H}$ ne možemo naučiti γ-slabo.

Primjer 7.1.2. *Neka je $\mathcal{X} = \mathbb{R}$, $\mathcal{Y} = \{\pm 1\}$ te neka je klasa $\mathcal{H}$ dana sa*

$$\mathcal{H} = \{h_{\theta_1,\theta_2,b} : \theta_1, \theta_2 \in \mathbb{R} \cup \{\pm\infty\}, \theta_1 < \theta_2, b \in \{\pm 1\}\}\,,$$

gdje je funkcija $h_{\theta_1,\theta_2,b} : \mathbb{R} \to \{\pm 1\}$ definirana sa

$$h_{\theta_1,\theta_2,b}(x) = \begin{cases} +b & za\ x < \theta_1\ ili\ x > \theta_2 \\ -b & za\ \theta_1 \leq x \leq \theta_2\,. \end{cases}$$

Ilustrativno, vrijednosti funkcije $h_{\theta_1,\theta_2,b}$ dane su Slikom 7.1.

Slika 7.1: Vrijednosti funkcije $h_{\theta_1,\theta_2,b}$

Neka je klasa $\mathcal{B}$ definirana sa

$$\mathcal{B} = \{f_{\theta,b} : \theta \in \mathbb{R} \cup \{\pm\infty\}, b \in \{\pm 1\}\}\,,$$

gdje je funkcija $f_{\theta,b} : \mathbb{R} \to \{\pm 1\}$ definirana sa

$$f_{\theta,b}(x) = \operatorname{sign}(x - \theta) \cdot b\,.$$

Cilj nam je pokazati da za $\gamma = 1/12$, sa $ERM_{\mathcal{B}}$ klasu $\mathcal{H}$ možemo naučiti γ-slabo. Kao prvo, svaka hipoteza $h_{\theta_1,\theta_2,b} \in \mathcal{H}$ dijeli $\mathbb{R}$ na 3 područja. Za svaki par područja (od 3 postojeća) postoji $f_{\theta,b} \in \mathcal{B}$ takva da se vrijednosti od $h_{\theta_1,\theta_2,b}$ i $f_{\theta,b}$ popudaraju na ta dva područja. U slučaju da izabrana područja imaju različit predznak, funkciju iz $\mathcal{B}$ izabrali bi tako da je $\theta \in \{\theta_1, \theta_2\}$, a u slučaju da odaberemo prvo i treće područje, onda $\theta = -\infty$. Uočimo da za proizvoljnu distribuciju $\mathcal{D}$ nad $\mathbb{R}$ i za bilo koju particiju od $\mathbb{R}$ na 3 područja vrijedi da jedno od tih područja po distribuciji $\mathcal{D}$ ima vjerojatnost najviše $1/3$.

Neka je $f : \mathcal{X} \to \{\pm 1\}$ proizvoljna ciljna funkcija. U definiciji slabog učenja pretpostavljamo da vrijedi $f \in \mathcal{H}$, pa slijedi da postoje $\theta_1, \theta_2 \in \mathbb{R}$, $\theta_1 < \theta_2$, te $b \in \{\pm 1\}$ takvi da vrijedi $f = h_{\theta_1,\theta_2,b}$. Tada funkcija f dijeli $\mathbb{R}$ na 3 područja koja ćemo označiti sa D_1, D_2 i D_3. Već smo zaključili kako jedno od ta 3 područja po distribuciji $\mathcal{D}$ ima vjerojatnost najviše $1/3$, pa bez smanjenja općenitosti možemo pretpostaviti da je to područje D_1. Tada za f postoji $f_{\theta,b} \in \mathcal{B}$ takva da se vrijednosti od f i $f_{\theta,b}$ popudaraju na D_2 i D_3.

Dodatno, zbog $f \in \mathcal{H}$ vrijedi $L_{\mathcal{D}}(f) = 0$. Definirajmo $g := f_{\theta,b}$. Kako vrijedi $f|_{D_2} = g|_{D_2}$ i $f|_{D_3} = g|_{D_3}$ te zbog $\mathbb{P}_{\mathcal{D}}(D_1) \leq 1/3$ slijedi

$$\begin{aligned} L_{\mathcal{D},f}(g) &= \mathop{\mathbb{P}}_{x\sim\mathcal{D}}(g(x) \neq f(x)) \\ &= \mathop{\mathbb{P}}_{x\sim\mathcal{D}}(x \in D_1)\ . \end{aligned}$$

Dakle, $\forall f \in \mathcal{H}$ vrijedi

$$\min_{g\in\mathcal{B}} L_{\mathcal{D},f}(g) \leq \frac{1}{3}\,. \tag{7.1}$$

Kako je $VCdim(\mathcal{B}) = 2$, primjenom Teorema 5.2.5 slijedi da $\mathcal{B}$ možemo naučiti u agnostičkom PAC smislu tako da za složenost vrijedi

$$C_1\frac{2+\ln(1/\delta)}{\varepsilon^2} \leq m_{\mathcal{B}}(\varepsilon,\delta) \leq C_2\frac{2+\ln(1/\delta)}{\varepsilon^2}\,.$$

Sada iz definicije agnostičkog PAC učenja slijedi da $\forall \varepsilon, \delta \in (0,1)$, $\forall f \in \mathcal{H}$ te za svaku distribuciju $\mathcal{D}$ nad $\mathcal{X}$ slijedi da korištenjem algoritma na skupu za učenje koji se sastoji od m nezavisnih primjera generiranih distribucijom $\mathcal{D}$, gdje je $m > m_{\mathcal{B}}(\varepsilon,\delta)$, algoritam vraća hipotezu h za koju vrijedi:

$$\mathbb{P}\left[L_{\mathcal{D},f}(h) \leq \min_{h'\in\mathcal{B}} L_{\mathcal{D},f}(h') + \varepsilon\right] \geq 1-\delta\,.$$

Kako (7.1) vrijedi za svaku distribuciju $\mathcal{D}$ i za svaki $f \in \mathcal{H}$, ako za veličinu m skupa za učenje S vrijedi

$$m \geq C_2\frac{2+\log(1/\delta)}{\varepsilon^2}\,,$$

slijedi

$$\mathbb{P}\left[L_{\mathcal{D}}(ERM_{\mathcal{B}}(S)) \leq 1/3 + \varepsilon\right] \geq 1-\delta\,.$$

Ako uzmemo da je $\varepsilon = 1/12$, onda s vjerojatnosti od barem $1-\delta$ vrijedi

$$L_{\mathcal{D}}(ERM_{\mathcal{B}}(S)) \leq \frac{1}{3} + \frac{1}{12} = \frac{1}{2} - \frac{1}{12}\,,$$

što znači da klasu $\mathcal{H}$ možemo naučiti γ-slabo za $\gamma = 1/12$.

Napomena 7.1.3. *Hipoteze iz klase $\mathcal{B}$ iz prošlog primjera nazivamo panjevima odluke (engl. decision stumps).*

7.2 AdaBoost algoritam

AdaBoost ili Adaptive Boosting je algoritam koji radi na principu iterativnog mijenjanja distribucije uzorkovanja u ovisnosti o pogrešci. Naime, neka je $S = \{(x_i, y_i) : i = 1, ..., m, x_i \in \mathcal{X}, y_i \in \{\pm 1\}\}$ fiksan skup od m elemenata, a $\mathcal{B}$ neka klasa baznih hipoteza (npr. panjevi odluke). Pretpostavimo da je WL algoritam koji za svaku distribuciju $\mathcal{D}$ nad $\{1, ..., m\}$ vraća hipotezu $h = WL(\mathcal{D}, S) \in \mathcal{B}$ tako da vrijedi

$$L_{\mathcal{D}}(h) = \sum_{i=1}^{m} D_i^{(t)} \mathbb{1}_{\{h_t(x_i) \neq y_i\}} \leq \frac{1}{2} - \gamma \,,$$

za neki $\gamma > 0$.

Kroz T iteracija ponavlja se postupak koji će nas po završetku dovesti do hipoteze s malom empirijskom greškom. Za svaki $t = 1, ..., T$, korištenjem algoritma WL na distribuciji $D^{(t)}$ i skupu S, dolazimo do hipoteze h_t definirane kao

$$h_t := WL(D^{(t)}, S) \,,$$

čija je prava greška po distribuciji $D^{(t)}$ dana sa

$$\epsilon_t := L_{D^{(t)}}(h_t) = \sum_{i=1}^{m} D_i^{(t)} \mathbb{1}_{\{y_i \neq h_t(x_i)\}} \,,$$

gdje po pretpostavci vrijedi

$$\epsilon_t \leq \frac{1}{2} - \gamma \,.$$

Koristeći definiciju težine w_t danu sa

$$w_t = \frac{1}{2} \log\left(\frac{1}{\epsilon_t} - 1\right) \,,$$

ažuriramo distribuciju $D^{(t)}$ tako da za svaki $i = 1, ..., m$ definiramo

$$D_i^{(t+1)} = \frac{D_i^{(t)} e^{-w_t y_i h_t(x_i)}}{\sum_{j=1}^{m} D_j^{(t)} e^{-w_t y_j h_t(x_j)}} \,.$$

Distribucija $D^{(t+1)}$ je dobro definirana jer vrijedi $D^{(t+1)} \in \mathbb{R}_+^m$ i

$$\sum_{i=1}^{m} D_i^{(t+1)} = 1 \,.$$

Uočimo da je w_t obrnuto proporcionalna grešci ϵ_t, stoga se na kraju iteracije t distribucija ažurira na način da će elementi od S na kojima je h_t točna imati manju težinu od onih elemenata na kojima daje pogrešnu predikciju. U konačnici će klasifikator kojeg nam daje AdaBoost biti baziran na težinskoj sumi slabih hipoteza dobivenih u svakoj iteraciji (vidi Slike 7.2 i 7.3 niže za ilustraciju).

AdaBoost algoritam

Ulazni podaci:
$S = (x_1, y_1), ..., (x_m, y_m)$
algoritam WL
broj ponavljanja T
Inicijalizacija: $D^{(1)} = \left(\frac{1}{m}, ..., \frac{1}{m}\right)$
Za $t = 1, 2, ..., T$:
definiraj klasifikator $h_t := WL(D^{(t)}, S)$
izračunaj $\epsilon_t = \sum_{i=1}^{m} \mathbb{1}_{\{y_i \neq h_t(x_i)\}}$
definiraj $w_t = \frac{1}{2} \log\left(\frac{1}{\epsilon_t} - 1\right)$
ažuriraj $D_i^{(t+1)} = \frac{D_i^{(t)} e^{-w_t y_i h_t(x_i)}}{\sum_{j=1}^{m} D_j^{(t)} e^{-w_t y_j h_t(x_j)}}$ za $i = 1, ..., m$
Rezultat: hipoteza $h_s(x) = \text{sign}\left(\sum_{t=1}^{T} w_t h_t(x)\right)$.

Sljedeći rezultat pokazuje da za fiksan skup za učenje S empirijska greška hipoteze koju dobijemo AdaBoost algoritmom opada eksponencijalno kako povećavamo T.

Teorem 7.2.1. *Neka je $S = \{(x_1, y_1), ..., (x_m, y_m)\}$ skup za učenje od m elemenata i neka u svakoj iteraciji WL vraća hipotezu za koju vrijedi*

$$\epsilon_t \leq \frac{1}{2} - \gamma .$$

Tada za empirijsku grešku hipoteze h_s dobivene AdaBoost algoritmom vrijedi

$$L_S(h_s) = \frac{1}{m} \sum_{i=1}^{m} \mathbb{1}_{\{h_s(x_i) \neq y_i\}} \leq e^{-2\gamma^2 T} .$$

Dokaz. Za svaki $t = 1, ..., T$ neka je

$$f_t = \sum_{p \leq t} w_p h_p ,$$

pa je hipoteza dobivena AdaBoost algoritmom dana sa $\text{sign} \circ f_T$. Neka je dodatno $f_0 = 0$. Za $t = 0, ..., T$ definirajmo

$$Z_t = \frac{1}{m} \sum_{i=1}^{m} e^{-y_i f_t(x_i)} .$$

Korištenjem činjenice da za fiksan $t = 0, ..., T$ i za svaki $i = 1, ..., m$ vrijedi

$$\mathbb{1}_{\{\text{sign}(f_t(x_i)) \neq y_i\}} \leq e^{-y_i f_t(x_i)} ,$$

iz definicije empirijske greške dobijemo da za svaki $t = 0, ..., T$ vrijedi

$$L_S(\text{sign}(f_t)) = \frac{1}{m} \sum_{i=1}^{m} \mathbb{1}_{\{\text{sign}(f_t(x_i)) \neq y_i\}} \leq \frac{1}{m} \sum_{i=1}^{m} e^{-y_i f_t(x_i)} = Z_t .$$

Stoga, za dokaz teorema dovoljno je pokazati da vrijedi $Z_T \leq e^{-2\gamma^2 T}$.

Iz $f_0 = 0$ slijedi $Z_0 = 1$, pa Z_T možemo napisati kao

$$Z_T = \frac{Z_T}{Z_{T-1}} \cdot \frac{Z_{T-1}}{Z_{T-2}} \cdots \frac{Z_2}{Z_1} \cdot \frac{Z_1}{Z_0} .$$

Slijedi da je dovoljno pokazati da za svaki $t = 1, ..., T$ vrijedi

$$\frac{Z_t}{Z_{t-1}} \leq e^{-2\gamma^2} . \tag{7.2}$$

Pokažimo prvo indukcijom da za svaki $t = 1, ..., m$ vrijedi

$$D_i^{(t)} = \frac{e^{-y_i f_{t-1}(x_i)}}{\sum_{j=1}^{m} e^{-y_j f_{t-1}(x_j)}} . \tag{7.3}$$

Ažuriranje distribucija korišteno u AdaBoost algoritmu dano je sa

$$D_i^{(t+1)} = \frac{D_i^{(t)} e^{-w_t y_i h_t(x_i)}}{\sum_{j=1}^{m} D_j^{(t)} e^{-w_t y_j h_t(x_j)}} .$$

Kako za $t = 1$ vrijedi $D_1 = (1/m, ..., 1/m)$, a iz definicije od f_t slijedi $f_1 = w_1 h_1$, dobijemo

$$\begin{aligned} D_i^{(2)} &= \frac{D_i^{(1)} e^{-w_1 y_i h_1(x_i)}}{\sum_{j=1}^{m} D_j^{(1)} e^{-w_1 y_j h_1(x_j)}} \\ &= \frac{\frac{1}{m} e^{-y_i f_1(x_i)}}{\sum_{j=1}^{m} \frac{1}{m} e^{-y_j f_1(x_j)}} \\ &= \frac{e^{-y_i f_1(x_i)}}{\sum_{j=1}^{m} e^{-y_j f_1(x_j)}} , \end{aligned}$$

tj. vrijedi (7.3) za $t = 2$.

Pretpostavimo da tvrdnja (7.3) vrijedi za neki $t \in \mathbb{N}$. Tada za $t + 1$ dobijemo

$$\begin{aligned} D_i^{(t+1)} &= \frac{\frac{e^{-y_i f_{t-1}(x_i)}}{\sum_{j=1}^m e^{-y_j f_{t-1}(x_j)}} e^{-w_t y_i h_t(x_i)}}{\sum_{j=1}^m \frac{e^{-y_j f_{t-1}(x_j)}}{\sum_{k=1}^m e^{-y_k f_{t-1}(x_k)}} e^{-w_t y_j h_t(x_j)}} \\ &= \frac{e^{-y_i f_{t-1}(x_i)} e^{-w_t y_i h_t(x_i)}}{\sum_{j=1}^m e^{-y_j f_{t-1}(x_j)} e^{-w_t y_j h_t(x_j)}} \\ &= \frac{e^{-\sum_{p\leq(t-1)} y_i w_p h_p(x_i)} e^{-w_t y_i h_t(x_i)}}{\sum_{j=1}^m e^{-\sum_{p\leq(t-1)} y_j w_p h_p(x_j)} e^{-w_t y_j h_t(x_j)}} \\ &= \frac{e^{-\sum_{p\leq t} y_i w_p h_p(x_i)}}{\sum_{j=1}^m e^{-\sum_{p\leq t} y_j w_p h_p(x_j)}} \\ &= \frac{e^{-y_i f_t(x_i)}}{\sum_{j=1}^m e^{-y_j f_t(x_j)}}, \end{aligned}$$

tj. vrijedi (7.3) za $t + 1$. Sada možemo zaključiti da (7.3) vrijedi za svaki $t \in \mathbb{N}$.

Korištenjem definicije od ϵ_t dobijemo

$$\epsilon_t = \sum_{i=1}^m D_i^{(t)} \mathbb{1}_{\{y_i \neq h_t(x_i)\}} = \sum_{i: y_i h_t(x_i) = -1} D_i^{(t)},$$

a iz $w_t = \frac{1}{2} \log\left(\frac{1}{\epsilon_t} - 1\right)$ dobijemo

$$e^{w_t} = \sqrt{1/\epsilon_t - 1}.$$

Sada za svaki $t = 1, ..., T$ vrijedi

$$\begin{aligned} \frac{Z_t}{Z_{t-1}} &= \frac{\sum_{i=1}^m e^{-y_i f_t(x_i)}}{\sum_{i=1}^m e^{-y_i f_{t-1}(x_i)}} \\ &= \frac{\sum_{i=1}^m e^{-y_i f_{t-1}(x_i)} e^{-y_i w_t h_t(x_i)}}{\sum_{i=1}^m e^{-y_i f_{t-1}(x_i)}} \\ &= \sum_{i=1}^m \frac{e^{-y_i f_{t-1}(x_i)}}{\sum_{i=1}^m e^{-y_i f_{t-1}(x_i)}} e^{-y_i w_t h_t(x_i)} \\ &= \sum_{i=1}^m D_i^{(t)} e^{-y_i w_t h_t(x_i)} \\ &= \sum_{i: y_i h_t(x_i) = 1} D_i^{(t)} e^{-w_t} + \sum_{i: y_i h_t(x_i) = -1} D_i^{(t)} e^{w_t} \end{aligned}$$

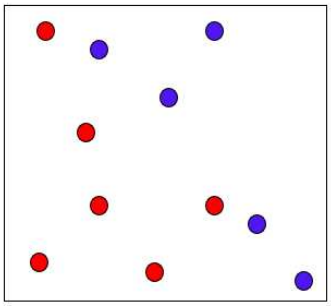
$$= (1-\epsilon_t)\frac{1}{\sqrt{1/\epsilon_t - 1}} + \epsilon_t \sqrt{1/\epsilon_t - 1}$$
$$= 2\sqrt{\epsilon_t(1-\epsilon_t)}\,.$$

Iz pretpostavke teorema da vrijedi $\epsilon_t \leq \frac{1}{2} - \gamma$ te korištenjem nejednakosti $1 - a \leq e^{-a}$ i činjenice da je $x(1-x)$ rastuća funkcija na $[0, 1/2]$ dobijemo

$$2\sqrt{\epsilon_t(1-\epsilon_t)} \leq 2\sqrt{(\frac{1}{2}-\gamma)(\frac{1}{2}+\gamma)} = \sqrt{1-4\gamma^2} \leq e^{-4\gamma^2} \leq e^{-4\gamma^2/2} = e^{-2\gamma^2}\,, \tag{7.4}$$

iz čega slijedi tvrdnja teorema. □

Jednostavna ilustracija AdaBoost algoritma može se vidjeti na sljedećem primjeru. Dan nam je skup točaka u ravnini $S = \{(x_{1,i}, x_{2,i}, y_i) : i \in \{1, ..., 11\}\}$, tako da točke označene plavom bojom imaju vrijednost $y_i = 1$, a točke označene crvenom bojom imaju vrijednost $y_i = -1$.



Slika 7.2: Skup za učenje S

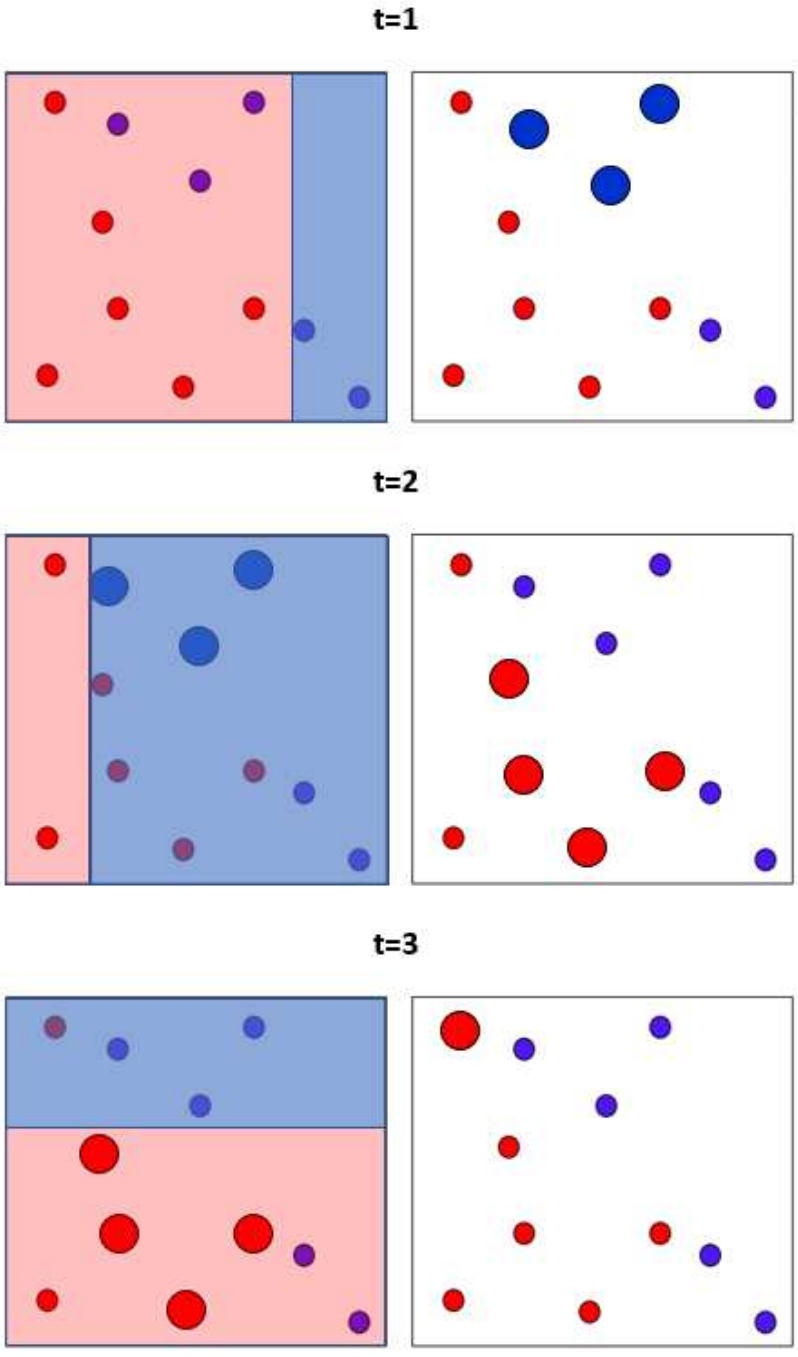
Neka je $T = 3$ i neka klasu baznih hipoteza $\mathcal{B}$ čine poluravnine usporedne s koordinatnim osima, tj.

$$\mathcal{B} = \{f_{i,\theta,b} : \theta \in \mathbb{R}, b \in \{\pm 1\}, i = 1, 2\},$$

gdje su funkcije $f_{i,\theta,b} : \mathbb{R}^2 \to \{\pm 1\}$ dane sa

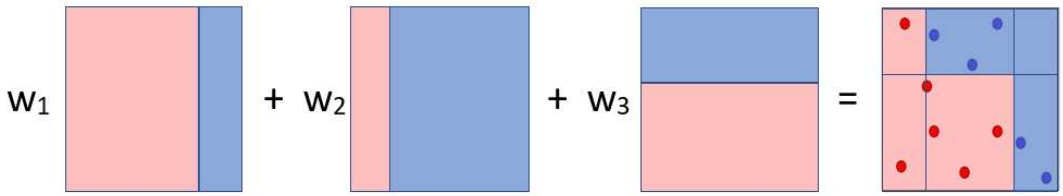
$$f_{i,\theta,b}(x_1, x_2) = \text{sign}(x_i - \theta) \cdot b\,.$$

Na sljedećoj slici prikazana je svaka iteracija algoritma, tj. u svakoj se iteraciji definira klasifikator iz $\mathcal{B}$ (lijevi stupac), a nakon toga se ažuririraju distribucije D^{t+1}, pri čemu veličina svake točke odgovara težini koju ta distribucija stavlja na tu točku (desni stupac).



Slika 7.3: Tri koraka AdaBoost algoritma (preuzeto iz [8])

Rezlultat AdaBoost algoritma je hipoteza dobivena kao predznak linearne kombinacije baznih hipoteza, što je ilustrirano na Slici 7.4.



Slika 7.4: Hipoteza dobivena AdaBoost algoritmom(preuzeto iz [8])

Napomena 7.2.2. *Jedna od čestih metoda za konstrukciju algoritma za slabo učenje je primjena ERM pravila na klasu $\mathcal{B}$, koja se naziva klasa baznih hipoteza. U Primjeru 7.1.2 u tu klasu spadaju panjevi odluke. Tada iz pretpostavki na početku potpoglavlja slijedi da je rezultat AdaBoost algoritma hipoteza iz klase $L(\mathcal{B}, T)$ dane sa*

$$L(\mathcal{B}, T) = \left\{ \text{sign}(\sum_{i=1}^{T} w_t h_t) : w \in \mathbb{R}^T, h_t \in \mathcal{B}, \forall t \in \{1, ..., T\} \right\} . \tag{7.5}$$

Može se pokazati (dokaz se može pronaći u [12]) da u slučaju kada je $T \geq 3$ i $VCdim(\mathcal{B}) \geq 3$ za VC dimenziju od $L(\mathcal{B}, T)$ vrijedi

$$VCdim(L(\mathcal{B}, T)) \leq T(VCdim(\mathcal{B}) + 1)(3 \log(T(VCdim(\mathcal{B}) + 1)) + 2) .$$

U Teoremu 7.2.1 pokazali smo da se empirijska greška hipoteze h_s dobivene AdaBoost algoritmom smanjuje kako se T povećava. Hipoteza h_s je dobivena kao linearna kombinacija T slabih hipoteza, pa u slučaju da te slabe hipoteze dolaze iz bazne klase $\mathcal{B}$ koja ima konačnu VC dimenziju, slijedi da je i VC dimenzija klase $L(\mathcal{B}, T)$, u koju spada i h_s, također konačna. Sada iz Teorema 5.2.4 slijedi da $L(\mathcal{B}, T)$ ima svojstvo uniformne konvergencije, pa će empirijska greška biti blizu pravoj grešci.

Bibliografija

[1] W.G. Macready D.H. Wolpert, *No Free Lunch Theorems for Optimization*, IEEE Transactions on Evolutionary Computation, https://ti.arc.nasa.gov/m/profile/dhw/papers/78.pdf/.

[2] D.Haussler, *Decision Theoretic Generalizations of the PAC Model for Neural Net and Other Learning Applications*, University of California.

[3] ———, *Overview of the Probably Approximately Correct Learning Framework*, University of California.

[4] M. Huzak, *Matematička statistika: skripta s predavanja*, https://web.math.pmf.unizg.hr/nastava/ms/index.php?sadrzaj=predavanja.php.

[5] J.Kun, *Probably Approximately Correct - a formal theory of Learning*, https://jeremykun.com/2014/01/02/probably-approximately-correct-a-formal-theory-of-learning/.

[6] J. Keshet, *Boosting*, Department of Computer Science Bar Ilan University, https://u.cs.biu.ac.il/~jkeshet/teaching/iml2015/iml2016_tirgul10.pdf.

[7] L.G.Valiant, *A theory of the learnable*, Communications of the ACM, 1984.

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

[9] U. Vazirani M.J. Kearns, *An Introduction to Computational Learning Theory*, The MIT Press, 1994.

[10] B. Barak S. Arora, *Computational Complexity: A Modern Approach*, Princeton University, 2007.

[11] A. Tewari S. Kakade, *Massart's Finite Class Lemma and Growth Function*, https://ttic.uchicago.edu/~tewari/lectures/lecture10.pdf.

[12] S.Ben David S.Shavel-Shwartz, *Understanding Machine Learning*, Cambridge University Press, 2014.

[13] Z. Vondraček, *Slučajni procesi: skripta s predavanja*, https://web.math.pmf.unizg.hr/~vondra/sp20-predavanja.html.

[14] X. Wu, *Lecture 7: Linear Predictors*, 2019, https://www.eecis.udel.edu/~xwu/class/ELEG867/Lecture7.pdf.

[15] ———, *Lecture 8: Boosting*, 2019, https://www.eecis.udel.edu/~xwu/class/ELEG867/Lecture8.pdf.

[16] ———, *Fundamental Theorem of Statistical Learning*, https://www.eecis.udel.edu/~xwu/class/ELEG867/Lecture6.pdf/.

[17] H.T.Lin Y.S.Abu-Mostafa, M.Magdon-Ismail, *Learning From Data*, 2012, http://amlbook.com.

Sažetak

Na početku rada definiramo najjednostavniju vrstu učenja koja se naziva minimizacija empirijskog rizika (ERM algoritam). Nakon toga definiramo glavni model za učenje, PAC model, koji se bazira na RA pretpostavci, te njegovu agnostičku varijantu koja ne zahtijeva tu istu pretpostavku. Uvođenjem pojma generalizirane funkcije gubitka, generalizirali smo model PAC učenja kako bismo ga mogli koristiti na široj skupini zadataka za učenje, a ne samo za binarnu klasifikaciju.

U središnjem dijelu rada dani su glavni teoremi. *No Free Lunch* teorem nam govori da ako nemamo nikakvih dodatnih pretpostavki na klasu hipoteza, ne postoji algoritam kojim možemo uspješno riješiti sve probleme učenja. To znači da svaki problem zahtijeva posebnu analizu, a pri biranju klase hipoteza veliki naglasak stavljamo na pronalazak ravnoteže između greške aproksimacije i greške procjene, što je u literaturi poznato kao *bias-complexity tradeoff*. Korištenjem Sauer-Shelah-Perles leme i Massartove leme dokazali smo glavni teorem ovog rada, *Fundamentalni teorem statističkog učenja*, kojim smo povezali pojmove uniformne konvergencije, (agnostičkog) PAC učenja, ERM algoritma i VC dimenzije za probleme binarne klasifikacije. Glavna poruka je da klasu hipoteza možemo naučiti u (agnostičkom) PAC smislu ako i samo ako ima konačnu VC dimenziju. Kvantitativna verzija tog teorema nam daje gornje i donje ograde na složenost učenja koje, između ostalog, ovise o VC dimenziji promatrane klase hipoteza.

U posljednjem dijelu rada usredotočili smo se na dva algoritma za probleme binarne klasifikacije. Prvi od njih koristi klasu poluprostora, a spada u jednu od najkorištenijih familija klasa hipoteza - linearne prediktore. Korištenjem Radonovog teorema dokazali smo konačnost VC dimenzije poluprostora, a nakon toga smo se usredotočili na Perceptron algoritam, iterativnu verziju ERM algoritma. Drugi algoritam koji smo opisali je AdaBoost algoritam koji spada u Boosting algoritme, a kao rezultat daje hipotezu koja ovisi o linearnoj kombinaciji nekih jednostavnih hipoteza.

Summary

In the beginning of this thesis, we define the simplest learning model called the Empirical Risk Minimisation (ERM) model. After that we define a formal learning model, the PAC model, which relies on the RA assumption, and its agnostic variant in which this assumption is omitted. By defining the generalised loss function we improved the PAC model so we can use it, not only in the context of binary classification, but also on a wider range of learning problems.

The main theorems are given in the central part of the thesis. *The No Free Lunch theorem* states that, if we don't have any additional assumptions on the hypothesis class, no learner can succeed on all learning tasks. This means that every learning task requires individual analysis and when choosing the hypothesis class, we have to pay attention on the balance between the approximation and estimation errors, which is known as *the bias-complexity tradeoff*. Using the Sauer-Shelah-Perles lemma i Massart lemma, we proved the key theorem of this thesis, *the Fundamental Theorem of Statistical Learning*, which reveals the connection between the notions of uniform convergence, (agnostic) PAC learning, ERM rule and VC dimension for binary classification problems. The theorem states that a hypothesis class is (agnostic) PAC learnable if and only if its VC dimension is finite. The quantitative version of the same theorem gives us the upper and lower bounds on the sample complexity of learning, which, among other things, depend on the VC dimension of the corresponding hypothesis class.

In the last part of the thesis, we focus on two algorithms for binary classification problems. First of them uses the hypothesis class of halfspaces which belongs to the family of linear predictors, one of the most often used family of hypothesis classes. Using the Radon theorem we showed that the VC dimension of halfspaces is finite and after that we focused on the Perceptron algorithm, an iterative version of the ERM model. The second algorithm is one of many Boosting algorithms called the AdaBoost algorithm. This algorithm outputs a hypothesis that depends on a linear combination of certain simple hypothesis.

Životopis

Rođena sam 4.8.1996. u Splitu, a odrasla u Runoviću pokraj Imotskog. Tijekom školovanja u Osnovnoj školi Runović raste moj interes za matematikom, stoga po završetku iste upisujem Prirodoslovno-matematičku gimnaziju dr. Mate Ujevića u Imotskom. Godine 2015. upisujem preddiplomski sveučilišni studij Matematika na Prirodoslovno-matematičkom fakultetu u Zagrebu, a nakon toga na istom fakultetu upisujem diplomski sveučilišni studij Matematička statistika.