Eötvös Loránd Tudományegyetem

Természettudományi Kar

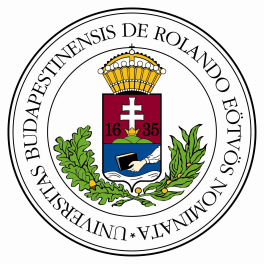
Ivkovic Iván

Formális módszerek a gépi tanulásban

Szakdolgozat

Matematika BSc, alkalmazott matematikus szakirány

Témavezető:
Lukács András
Számítástudomány Tanszék



Budapest, 2019

Köszönetnyilvánítás

Ezúton szeretném megköszönni témavezetőmnek, Lukács Andrásnak a segítséget a témaválasztásban és a szakirodalom kiválasztásában, illetve a szakdolgozat elkészítéséhez adott tanácsait.
Szeretném megköszönni a családomnak, hogy 15 éve támogatják tanulmányaimat és hogy lehetővé tették, hogy ezidő alatt csak az iskolával foglalkozhassak. Köszönöm, hogy elindítottak a gondolkodó ember útján.

Tartalomjegyzék

Bevezetés

Az elmúlt években a gépi tanulásban megfigyelt látványos fejlődés legfőképp gyakorlat orientált, a példák, feladatok, amelyek motiválják a terület fejlődését meghatározóan alkalmazás során merülnek fel. Hasonlóan a matematikai eszközökkel való modellezés más eszközeihez itt is kívánatos egy olyan elméleti megalapozás, amely formalizálja a gyakorlatban megoldandó problémák során csak intuitívan használt fogalmakat. Felépítve a gépi tanulás elméletét abban segítheti a gyakorlatot, hogy az egyes modellek elég pontos tanulhatóságára ad szükséges és elégséges feltételeket. Ilyenkor analitikus eszközökkel alátámasztva kaphatjuk meg, hogy egy adott feladat megoldása során mekkora valószínűséggel biztosan nem vétünk bizonyos korlátnál nagyobb hibát, elméleti határt szabva a modellek tanulhatóságára.

A dolgozat első felében fogalmakat bevezetve és ezen fogalmakat vizsgálva jutunk el olyan tételekhez, amelyek az előbb említett feltételeket megadják a tanuló modelljeinkhez.

Dolgozatom a Shai Shalev-Schwartz és Shai Ben-David - Understanding Machine Learning című monográfiájának első fejezetét dolgozza fel, a könyvben feladatként megfogalmazott állításokat a saját bizonyításaimmal kiegészítve. A könyv a fogalmak, levezetések során - a szerzők által kiemelten - nem várja el az olvasóktól a mértékelmélet ismeretét, így az e fogalomköröket érintő definíciókat, illetve a bizonyításokat kiegészítettem, és a be nem bizonyított részállításokat kiegészítettem a levezetéseimmel.

Az első fejezetben bevezetjük a dolgozat során használt alapfogalmakat, illetve jelöléseket. A továbbiakban a fő célunkhoz vezető definíció a hipotézis hibája, mellyel röviden ismerkedve és az első észrevételeket megtéve, ebben a fejezetben látjuk majd, hogy a tapasztalati hiba minimalizálása megefelelő feltételek mellett jó közelítése lehet a valódi hibának. A tapasztalati hibával egy ellenőrizhető karakterizációt adtunk az "elfogadható", valószínűleg kis hibával rendelkező hipotézisre. Ebben a fejezetben belátjuk még, hogy a hipotézisosztály méretének véges halmazra való korlátozásával nem léphet fel túlilleszkedés a tanuló algoritmus futása során, ha a kielégíthetőségi feltétel teljesül. Fontos eredmény még, hogy azon probléma esetén, hogy megfelelő valószínűséggel egy ε hibával rendelkező hipotézist kapjunk, ekkor ennek elérése érdekében egy alsó korlátot kapunk az adatgeneráló eloszlás szerint mintázott független minta méretére.

A második fejezetben vezetjük be a dolgozat során különböző szempontok szerint vizsgálni kívánt és a későbbiekben több lehetséges módon kiterjesztésre kerülő modellt, a PAC tanulást. A PAC tanulás két formáját fogjuk felhasználni a továbbiakban aszerint,

hogy feltehető-e, hogy létezik egy valószínűséggel pontos tanuló vagy sem, azaz hogy feltehető-e a kielégíthetőség vagy sem, így bevezetve az agnosztikus PAC tanulást. Ebben a fejezetben térünk ki a címkehalmazzal kapcsolatban néhány általánosításra, megmutatva, hogy az általunk vizsgált feladat ekvivalens vagy könnyen kiterjeszthető az általonasabbra, így a későbbiekben elég az egyszerűbb, könnyebben átlátható esetekre szorítkozni.

A harmadik fejezetben az egyenletes konvergencia tulajdonságot vezetjük be, a gépi tanulás kontextusában. Ezt a feltételt az a gyakorlatból eredeteztethető elvárás motiválta, hogy el szeretnénk érni, hogy a tanuló számára ismeretlen valódi kockázatnak a kiszámolható tapasztalati rizikó jó közelítése legyen minden hipotézisosztálybeli elemre. Ebben a pillanatban még nem világos ezen feltétel ilyen mértékű vizsgálata, de a dolgozat végén a statisztikai tanulás alptételében látjuk, hogy a tanulhatóságnak az egyenletes konvergencia szükséges és elégséges felétetele.

A negyedik fejezetben már tapasztalhatjuk az eddig elvégzett számolások, illetve megfigyelések gyümölcsét, bebizonyítjuk a No-Free-Lunch-tételt. Az előbbi tétel kimondja, hogy nincs univerzális tanuló, azaz nem létezik tanuló, amely az összes tanulási feladat megoldására alkalmas, tehát minden tanulóra létezik olyan feladat, amelyre elbukik, míg más tanuló ugyanerre a feladatra sikerrel alkalmazható. Ennek a tételnek egy nagyon fontos következménye, hogy a probléma sikeres megoldásához szükségünk van előismeretekre a hipotézisosztállyal kapcsolatban, melyet azon állítás bizonyításával igazolunk, hogy egy végtelen $\mathbf{X}$ alaphalmazból vesszük az összes $\mathbf{H} \to \{0, 1\}$ függvényből álló hipotézisosztályt és megmutatjuk, hogy ekkor $\mathbf{H}$ nem PAC tanulható. Ebben a fejezetben megvizsgáljuk még egy ERM_H hipotézis hibájának felbontását, mégpedig approximációs, illetve becslési hiba formájában, ahol ezen felosztás miatt egy kompromisszum szükségességével szembesülünk.

Az ötödik fejezetben bevezetjük a VC-dimezió fogalmát és néhány példán keresztül a VC-dimenzió kiszámolosávál intuitív képet is kapunk a VC-dimenzió definíciójáról, és hogy milyen mértékben tükrözi egy hipotézisosztállyal kapcsolatos elvárásainkat a véges VC-dimenzió. Ebben a fejezetben érkeztünk el arra a pontra, ahol az eddigi bizonyítások, fogalomvizsgálatok és megfigyelések után learathatjuk a babérokat, azaz kimondjuk és bebizonyítjuk a statisztikai tanulás alaptételét. Az alaptétel ekvivalens tulajdonságokat fogalmaz meg a tanulhatóságra vonatkozóan, azaz elértünk a beveztő elején említett, az alkalmazás által egyik motivációként felvetett célhoz: ekvivalens feltételeket adtunk a tanulhatóságra az általunk vizsgált modellek mellett.

1. fejezet

Alapfogalmak bevezetése

1.1. A statisztikai tanulás keretei

Egy általános statisztikai tanuló modellnek az alább definiált halmazokhoz, függvényosztályokhoz van hozzáférése:

Az *alaphalmaz* egy tetszőleges $\mathbf{X}$ halmaz, amely elemei a megfigyelni kívánt tulajdonságoknak megfelelő méretű vektorok, így vizsgálható reprezentációt adva a megoldanadó feladatoknak, problémáknak. A továbbiakban $\mathbf{X}$ bizonyos elemeit szeretnénk címkézni, az $\mathbf{X}$ elemeire, mint adatpontokra fogunk tekinteni, ahol általában $\mathbf{X} \subseteq \mathbb{R}^p$.

$\mathbf{Y}$-nal fogjuk jelölni a *címkehalmazt*, amely elemei a tanuló algoritmus által az adatpontokhoz hozzárendelt/hozzárendelendő következtetések. Általában feltesszük, hogy a címkehalmaz kételemű, $\{0, 1\}; \{+1, -1\}$.

A *tanító adathalmaz* egy, az $\mathbf{X} \times \mathbf{Y}$ elemeiből képzett sorozat:
$S := ((x_1, y_1), ..., (x_m, y_m))$. Azaz S címkézett adatpontok véges halmaza, amelyhez a tanuló hozzáfér és amelyből egy minél pontosabb predikciót szeretnénk felállítani a vizsgált probléma minden elemére. A tanító adathalmazra mindig sorozatként tekintve kezelhető a kétszer ugyanazon adatpont mintázása és a későbbiekben előnyt jelenthet, ha S sorba rendezhető.

Most definiáljuk a tanuló kimenetét. A fenti bemenetek segítségével a tanuló algoritmus feladata egy predikció felállítása az $\mathbf{X}$ és $\mathbf{Y}$ halmazok között, azaz megad egy $h : \mathbf{X} \to \mathbf{Y}$ függvényt, amit hipotézisnek vagy predikciónak nevezünk. Az outputként megadott hipozézissel címkézzük azon adatpontokat is, amelyekhez a tanulónak nem volt hozzáférése. Kételemű címkehalmaz esetén ez annyit jelent, hogy h megadja, hogy az adott adatpont

a két halmaz közül melyikbe esik. Vezessük be az $A(S)$ jelölést az A algoritmus által az S tanító adathalmaz bemenet esetén megadott predikcióra.

Vizsgáljuk meg a következő kérdéskört: Hogyan generáljuk az adatainkat? Először is tegyük fel, hogy az adatpontokat valamely valószínűségi eloszlás szerint generáltuk, jelöljük $\mathbf{D}$-vel, amelyet kézenfekvően értelmezzünk az $\mathbf{X}$ Borel-halmazain. Fontos megjegyezni, hogy a tanuló nem ismeri ezt az eloszlást, és az általunk később vizsgálandó általános tanulási modellek felépítéséből adódóan $\mathbf{D}$ legyen egy tetszőleges eloszlás. Tegyük fel, hogy létezik egy $f : \mathbf{X} \to \mathbf{Y}$ címkéző függvény, amely minden i-re $f(x_i) = y_i$ értéket vesz fel. Ezt a függvényt a tanuló nem ismeri, a cél ezen függvény minél pontosabb közelítése. Tehát az S tanító adathalmaz a $\mathbf{D}$ eloszlás által generált mintapontokból és ezekhez az f függvény által meghatározott címkékből képzett párok sorozata.

Miután definiáltuk azt a célfüggvényt, amit a tanulóval a legpontosabban szeretnénk közelíteni, felmerül a kérdés: Hogyan illetve milyen eszközökkel mérjük a pontosságot? Annak az eseménynek a valószínűségét, hogy egy tetszőleges, a fent ismertetett eloszlás által generált adatponthoz a hipotézis nem a megfelelő címkét rendeli, nevezzük a klasszifikáló modell hibájának. Formálisan, legyen adott az $\mathbf{X}$ Borel-halmazainak egy eleme $A \in \mathbf{B}(\mathbf{X})$, a $\mathbf{D}$ valószínűségi eloszlás az A eseményhez a $\mathbf{D}(A)$ számot rendeli, amely azt jelenti,hogy milyen eséllyel figyelünk meg egy A-beli pontot.

1.1.1. Definíció (A hipotézis hibája). *Legyen $h : \mathbf{X} \to \mathbf{Y}$ egy hipotézis, akkor a predikció hibája:*

$$L_{D,f}(h) \stackrel{def}{=} \mathbf{P}[h(x) \neq f(x)] \stackrel{def}{=} \mathbf{D}(\{x : h(x) \neq f(x)\})$$

A fent definiált hiba függ a $\mathbf{D}$ eloszlástól és az f címkézőfüggvénytől is, ezt indikálja az alsó indexben szereplő (D, f). A továbbiakban $L_{D,f}(h)$-t generelási hibának, valódi rizikónak vagy valódi hibának hívjuk.

1.1.2. Megjegyzés. *A tanuló algoritmusunk nem fér hozzá az előbbiekben definiált $\mathbf{D}$ eloszláshoz és az f címkézőfüggvényhez, ami azt jelenti, hogy nincs előzetes ismerete az adatpontok tulajdonságai esetén a következtetésekre. Az előbbi információk hiányában a tanuló csak a tanító adathalmaz vizsgálata során felismert mintázatokra, szabályszerűségekre támaszkodhat. Ezek segítségével létrehoz egy jóslást, amellyel az értelmezési tartomány tetszőleges pontjának címkéjére ad egy hipotézist.*

1.2. A tapasztalati kockázat minimalizálása

Ahogy már korábban definiáltuk, a tanuló algoritmus egy S tanító adathalmazt kap inputként, amely elemei egy ismeretlen $\mathbf{D}$ eloszlás szerint mintázottak és egy ismeretlen f függvény szerint címkézettek. A tanuló rendelkezésére álló információk alapján a pontos hiba meghatározására nincs lehetőség. Legyen az algoritmus által, egy S sorozat esetén adott hipotézis: $h_S : \mathbf{X} \to \mathbf{Y}$. A cél egy legjobban közelítő hipotézis megtalálása f-re és D-re nézve. Mivel a valódi rizikóhoz nem férünk hozzá, de a hipotézisek között fel kell állítanunk egy rendezést, amely az elérhető információk alapján lehetőséget ad egy optimális hipotézis választására, így definiálni fogunk egy hibaformulát a tanító halmazon. Természetes módon adódik a predikció sikerességének alábbi mértéke:

1.2.1. Definíció (A tapasztalati hiba). *Egy h hipotézis tapasztalati hibáját egy m méretű S tanuló adathalmazon a következő hányados definiálja:*

$$L_S(h) \stackrel{def}{=} \frac{|i \in [m] : h(x_i) \neq y_i|}{m},$$

ahol $[m] = \{1, ..., m\}$.

A tapasztalati hiba a vizsgálandó adathalmaz egy részén való pontosságról ad képet, nem foglalkozva a teljes adathalmazra vonatkozó hipotézis minőségével. Azt a tanulási paradigmát, amely során azt a h hipotézist választjuk, amely optimális az $L_S(h)$-ra nézve, tapasztalati hiba minimalizációnak, röviden ERM-szabálynak nevezzük.

Az ERM-szabály definiálása és alkalmazása természetesen adódik, de nem vigyázva hamis képet adhat a hipotézisünk tényleges minőségéről. Ha a tanuló algoritmusunk teljesen elsajátítja a tanító adathalmaz tulajdonságait, szabályait és képes minden esetben helyes predikciót adni a tanító adaton, arra engedne következtetni, hogy találtunk egy hipotézist, amely teljes mértékben megegyezik az f címkéző függvénnyel. Azonban végig gondolva a problémát, az alaphalmaz csak egy részét vizsgáltuk és az egy helytelen implikáció lenne, hogy akkor a halmaz összes elemére ugyanaz a hipotézis jó eredményt adhat. Ha az algoritmusunk által megadott hipotézis függvény túl pontosan illeszkedik a mintára, a tanító adathalmaz elemeinek tulajdonságait, következtetéseit nagy pontossággal követi, de a valóságban nagy hibával dolgozik, akkor túlilleszkedésről (overfitting) beszélünk. Általában egy minél simább függvényt szeretnénk az algoritmus kimenetének, amely emellett kis tapasztalati rizikóval rendelkezik.

1.3. Tapasztalati kockázat minimalizálása induktív torzítással

Az előző alfejezetben arra a következtetésre jutottunk, hogy az ERM-elv túlilleszkedéshez vezethet. Ahelyett hogy elvetnénk az ERM paradigmát, változtatásokat eszközölünk, javítva ezzel a szabályunk becslésén. Bevezetünk feltételeket, amelyek garantálják, hogy az ERM-elv mellett nem léphet fel túlilleszkedés, amelyek mellett az ERM hipotézis pontosan működik. Egy gyakori megoldás, hogy az ERM-elvet egy előre meghatározott téren alkalmazzuk, azaz még mielőtt a tanuló algoritmus megkapna bemenetként egy adathalmazt leszűkítjük a lehetséges predikciók osztályát. Ezt az osztályt a továbbiakban **H**-val fogjuk jelölni és hipotézis-osztálynak fogjuk nevezni, mely minden $h \in \mathbf{H}$ eleme **X**-ből **Y**-ba képez. Egy adott **H** hipotézisosztály és S tanító minta esetén az ERM_H tanuló algoritmus az ERM-elvet használja az optimális predikció megtalálására, a lehető legalacsonyabb S feletti hiba mellett, képlettel:

$$ERM_H(S) \in argmin_{h \in H} L_S(h).$$

Azzal, hogy megszorítottuk azt a halmazt, amely elemeiből az algortimus a predikciós szabályt választhatja, ezen részhalmaz irányában torzítottuk, ezt a leszűkítést induktív torzításnak nevezzük. Mivel a megszorításokat, előfeltételeket a hipotézissel kapcsolatban az algoritmus inputhoz való hozzáférése előtt meg kell adnunk, ezért szükségünk van bizonyos előismeretekre a megoldandó problémával kapcsolatban, mellyel már megvalósítható karakterizációt adhatunk a problémának, illetve a hipotézisnek. A tanuláselmélet egy alapkérdése, hogy mely hipotézisosztály felett nem vezet az ERM_H tanulás túlilleszkedéshez. Erre a kérdésre a későbbiekben választ kapunk.

1.4. Véges hipotézisosztály

Az előbbiekben arra a megállapításra jutottunk, hogy ahhoz, hogy az ERM_H ne vezessen túlilleszkedéshez, bizonyos megkötéseket kell bevezetni. Az egyik legegyszerűbb ilyen feltétel a hipotézisosztály méretének korlátozása. Ebben a részben belátjuk, hogy ha a **H** halmaz véges, akkor az ERM_H tanuló alkalmazása esetén nem léphet fel túlilleszkedés.

Kezdjük el vizsgálni az ERM_H tanulót véges $\mathbf{H}$ mellett. Jelöljük h_S-sel az S tanító minta és az $f : \mathbf{X} \to \mathbf{Y}$ címkézőfüggvény esetén ERM_H által eredményül adott hipotézist:

$$h_S \in argmin_{h \in H} L_S(h).$$

Ebben a fejezetben az alább definiált feltételek mellett végzünk számításokat.

1.4.1. Definíció (kielégíthetőségi feltétel). *Ezen feltétel szerint létezik egy $h^* \in \mathbf{H}$, amely 1 valószínűséggel megegyezik f-fel, azaz $\exists h^* \in \mathbf{H} : L_{D,f}(h^*) = 0$.*

1.4.2. Megjegyzés. *Az előbbi feltétel fennállása a következőt implikálja: Ha egy véletlenszerűen generált S adatpontjai D eloszlás alapján mintázottak és az f függvény által címkézettek, és ezen paraméterek mellett igaz a kielégíthetőségi feltétel, akkor $L_S(h^*) = 0$, így minden optimális ERM hipotézisre $L_S(h_S) = 0$.*

Azonban minket h_S tapasztalati rizikója helyett az $L_{D,f}(h_S)$ valódi kockázat érdekel. A statisztikai gépi tanulás egyik legalapvetőbb feltétele, hogy az S tanító adathalmaz adatpontjait a $\mathbf{D}$ eloszlás szerint egymástól függetlenül mintáztuk.

1.4.3. Definíció (iid feltétel). *A tanító adathalmazban szereplő mintapontok a $\mathbf{D}$ eloszlás szerint független, azonos eloszlásúak. Tehát az S tanító halmaz minden x_i újonnan, a korábban generáltaktól függetlenül mintázott pontjához az f függvény megad egy címkét. Ezt a feltételt a továbbiakban $S \sim \mathbf{D}^m$ jelöljük, ahol m az S-ben szereplő adatpontok száma.*

1.4.4. Megjegyzés. *Az eddig bevezetett formális definíciók alapján a problémát intuitívan a következő módon érdemes megfogni. Tekintsünk az S mintára, mint egy ablakra az alaphalmaz felett, amely részleges információt nyújt a vizsgálandó "világról". Nyilván ahogy nő a tanító adathalmaz mérete, úgy egyre pontosabban képes a kiakalakított hipotézis a $\mathbf{D}$ eloszlásra és f címkéző függvényre reflektálni.*

Az $L_{D,f}(h_s)$ tényleges rizikóra valószínűségi változóként tekintünk, hiszen függ S-től, amely elemeit véletlenszerűen generáltuk a $\mathbf{D}$ eloszlás szerint, illetve h_S-t is véletlenszerűen választottuk az optimális ERM hipotézisek közül. Tehát nem adna reális eredményt, ha elvárnánk, hogy minden S minta esetén képes az algoritmus használható klasszifikációt adni az egész adathalmazra. Ugyanis egy bizonyos valószínűséggel mindig fennáll az az esemény, hogy az S a $\mathbf{D}$ eloszlás szerint nem ad reprezentatív mintát. Például ha egy

bináris klasszifikáció esetén egy vizsgált minta minden eleme az egyik címkehalmazba esik, azonban az általunk nem ismert $\mathbf{D}$ eloszlás szerint az összes adatpont csupán 20%-a esik ebbe a címkehalmazba, akkor a hipotézisünk 80%-os valódi hibával rendelkezik. A továbbiakban jelöljük δ-val annak az eseménynek a valószínűségét, hogy nemreprezentatív mintát kapunk, és legyen $(1-\delta)$ ezek alapján a predikció megbízhatósági paramétere.

Ezenfelül, mivel nem tudjuk garantálni a tökéletes predikciót, bevezetünk még egy paramétert a hipotézis minőségének tükrözésére, amit pontossági paraméternek nevezünk, és ε-nal jelölünk. A következőképpen értelmezzük a tanuló kudarcát: $L_{D,f}(h_S) > \varepsilon$, azonban ha $L_{D,f}(h_S) \leq \varepsilon$, akkor közelítőleg helyes hipotézisről beszélünk. Lerögzítve egy $f : \mathbf{X} \to \mathbf{Y}$ címkézőfüggvényt, felülről szeretnénk korlátozni, m méretű S tanító adathalmazra a tanuló algoritmus kudarcának valószínűségét. Jelöljük $S|_X = (x_1, ..., x_m)$-mel a tanító halmaz adatpontjait. A következőkben az alábbi mennyiségre keresünk felső korlátot:

$$\mathbf{D}^m(\{S|_X : L_{D,f}(h_s) > \varepsilon\})$$

Jelöljük $\mathbf{H}_B$-vel azon hipotézisek halmazát, amelyek kudarchoz vezetnek az előbb leírt probléma esetén, azaz:

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

Azon $S|_X$ adatpontok halmazát, amelyekhez létezik egy $h \in \mathbf{H}_B$, amely minden x_i mintaponthoz $f(x_i)$-vel megegyező címkét rendel, azt a halmazt jelöljük:

$$\mathbf{M} = \{S|_X : \exists h \in \mathbf{H}_B, L_S(h) = 0\}$$

Annak az eseménynek a valószínűségét szeretnénk felülrő korlátozni, hogy $L_{D,f}(h_S) > \varepsilon$, de mivel feltettük, hogy létezik olyan hipotézis, amely megegyezik f-fel 1 valószínűséggel, és h_S egy optimális hipotézis volt az ERM-szabályra nézve, így $L_S(h_S) = 0$. Ezek alapján az $L_{D,f}(h_S) > \varepsilon$ esemény egy $h \in \mathbf{H}_B$ esetén csak úgy teljesühet, ha $L_S(h) = 0$. Másfelől pedig ez az esemény csak úgy állhat fenn, ha a vizsgált tanító adathalmaz adatpontjai $\mathbf{M}$-ben vannak, ezzel megmutattuk, hogy

$$\{S|_X : L_{D,f}(h_S) > \varepsilon\} \subseteq \mathbf{M}$$

Ahol $\mathbf{M}$-et a következő alakban is írhatjuk:

$$\bigcup_{h \in \mathbf{H}_B} \{S|_X : L_S(h) = 0\}$$

Így a következő egyenlőtlenséget kapjuk:

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

Mivel az eloszlás mint mérték σ-szubadditív, így

$$\mathbf{D}^m(\{S|_X : L_{D,f}(h_S) > \varepsilon\}) \leq \sum_{h \in \mathbf{H}_B} \mathbf{D}^m(\{S|_X : L_S(h) = 0\})$$

A jobb oldalt tagonként az alábbi módon tudjuk felülről becsülni, rögzítsünk le egy $h \in \mathbf{H}_B$ "rossz" hipotézist. Az $L_S(h) = 0$ esemény ekvivalens azzal, hogy $\forall i, h(x_i) = f(x_i)$. A tanító adathalmaz adatpontjait egymástól függetlenül, azonos eloszlás szerint mintáztuk, így az előbb leírt események egymástól függetlenek, azaz

$$\mathbf{D}^m(\{S|_X : L_S(h) = 0\}) = \prod_{i=1}^{m} \mathbf{D}(\{x_i : h(x_i) = f(x_i)\}) \tag{1.1}$$

Minden függetlenül generált tanító adathalmazbeli adatpontra az alábbi egyenlőtlenség a $h \in \mathbf{H}_B$, illetve a valódi rizikó definíciója miatt fennáll,

$$\mathbf{D}(\{x_i : h(x_i) = f(x_i)\}) = 1 - L_{D,f}(h) \leq 1 - \varepsilon \leq e^{-\varepsilon}$$

Az előző egyenlőtlenségekből kapjuk, hogy minden $h \in \mathbf{H}_B$-re

$$\mathbf{D}^m(\{S|_X : L_S(h) = 0\}) \leq e^{-m\varepsilon}$$

Végül az (1.1)-es egyenlőtlenség alapján

$$\mathbf{D}^m(\{S|_X : L_{D,f}(h_s) > \varepsilon\}) \leq |\mathbf{H}_B| e^{-m\varepsilon} \leq |\mathbf{H}| e^{-m\varepsilon}$$

Ezzel felső korlátot adva a tanuló kudarcára.

1.4.5. Következmény. *Legyen* $\mathbf{H}$ *véges hipotézisosztály,* $\delta \in (0,1), \varepsilon > 0$ *paraméterek, legyen* m *egy egész szám, amelyre*

$$m \geq \frac{ln(|\mathbf{H}|/\delta)}{\varepsilon}$$

Akkor minden f *címkéző függvényre és minden* $\mathbf{D}$ *valószínűségi eloszlásra, amelyre fenn áll a kielégíthetőségi feltétel,* $1-\delta$ *valószínűséggel egy* m *elemű független mintára minden* h_S *ERM hipotézis esetén a valódi rizikóra* ε *jó felső becslés,*

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

1.4.6. Megjegyzés. *A következményben szereplő,* m *alsó korlátját meghatározó hányados az előbbi levezetésből a következőképpen kapható:*

$$|\mathbf{H}|e^{-m\varepsilon} \leq \delta$$

$$ln(|\mathbf{H}|/\delta) \leq m\varepsilon$$

Ahol emlékeztetőül δ *annak a valószínűsége volt, hogy nemreprezentatív mintát kapunk, azaz így* $1-\delta$ *nem csak a reprezentatív minta generálásának a valószínűsége, hanem ezenfelül egy legalább* ε *pontossággal tanulható mintáé is.*

Az (1.4.5.) következményből tudjuk, hogy elég nagy m-re, az ERM_H egy véges hipotézis osztályon valószínűleg közelítőleg helyes. A következő fejezetben ezt a modellt, melyet a továbbiakban PAC-tanulásnak (Probably Approximately Correct) hívunk, fogjuk általánosabban vizsgálni.

2. fejezet

Egy formális tanuló modell

Ebben a fejezetben a fő tanuló modellünket fogjuk vizsgálni, a PAC tanulást, illetve ennek bizonyos kiterjesztéseit.

2.1. PAC tanulás

Az előző fejezet végén megmutattuk, hogy az ERM-szabály egy véges hipotézisosztály felett, megfelelően nagy méretű tanító adathalmaz mellett "nagy valószínűséggel" közelítőleg helyes predikciót ad. Még általánosabb tanulási modell a következő, melyet PAC (Probably Approximately Correct) tanulásnak nevezünk.

2.1.1. Definíció (PAC tanulhatóság). *Egy* $\mathbf{H}$ *hipotézisosztály PAC-tanulható, ha létezik egy* $m_H : (0,1)^2 \to \mathbb{N}$ *függvény és egy tanuló algoritmus a következő tulajdonságokkal: minden* $\mathbf{D}$*, az* $\mathbf{X}$ *értelmezési tartomány Borel-halmazain értelmezett valószínűségi eloszlásra, minden* $\varepsilon, \delta \in (0,1)$ *paraméterre és minden* $f : \mathbf{X} \to \{0;1\}$ *címkéző függvényre, ha a kielégíthetőségi feltétel teljesül, akkor az algoritmus egy* $m_H(\varepsilon, \delta) \leq m$ *méretű, egymástól függetlenül azonos* $\mathbf{D}$ *eloszlás szerint generált,* f *függvény által címkézett tanító minta esetén olyan* h *hipotézist ad vissza, amelyre* $L_{D,f}(h) \leq \varepsilon$ *igaz* $1 - \delta$ *valószínűséggel.*

Emlékeztetőül, az előbbi definícióban szereplő két paraméter (ε, δ) a pontossági, illetve a megbízhatósági paraméterek. Az ε paraméter azon h kimenetként kapott hipotézisek megtartását indukálja, amelyek esetén ε-nál nem tértünk el többel az optimálistól, míg a δ paraméter azt adja meg, hogy mekkora eséllyel lesz a klasszifikációnk ε pontosságú. A tanító adathalmaz adatpontjai egymástól függetlenül mintázottak, ezért előfordulhat az

is, hogy egy darab adatpontot generálunk újra és újra. Ilyen és ehhez hasonló, nem megfelelő információtartalommal rendelkező minták figyelmen kívül hagyására vezettük be a δ paramétert. Mindemellett, még ha szerencsések is vagyunk a generált minta információtartalmát illetően, lévén egy véges sorozat, nem képes visszaadni a **D** eloszlás minden tulajdonságát, részletét. A hipotézis ezen részletek megtanulhatatlanságából eredő hibáinak figyelmen kívül hagyására szolgál az ε paraméter. Ezekkel az általánosításokkal tehát egy, a valóságot sokkal pontosabban követni képes modellt kaptunk.

Legyen $m_H : (0,1)^2 \to \mathbb{N}$ azon függvény, amely megad egy alsó korlátot a generálandó adatpontok számára, melyre garantálhatjuk **H** PAC tanulhatóságát. Ezt a függvényt a minta komplexitásának nevezzük, és az előbb leírt két paraméter mellett a hipotézisosztály tulajdonságaitól is függ.

2.1.2. Megjegyzés. *Ha egy* **H** *osztály PAC tanulható, akkor több m_H függvény is létezik, amely mellett fennállnak a PAC tanulhatóság feltételei. Az egyértelműség kedvéért ezentúl legyen a minta komplexitása azon m_H függvény, amely tetszőleges ε, δ esetén $m_H(\varepsilon, \delta)$ azon minimális pozitív egész szám, amely kielégíti a PAC tanulhatóság definíciójában leírt feltételeket ε pontossággal és δ megbízhatósággal.*

2.1.3. Következmény. *Minden véges hipotézisosztály PAC tanulható*

$$m_H(\varepsilon, \delta) \leq \lceil ln(|\mathbf{H}|/\delta)/\varepsilon \rceil$$

komplexitással.

2.1.4. Állítás. *m_H monoton csökkenő ε-ban és δ-ban.*

Bizonyítás.

Legyen **H** *egy hipotézisosztály, amelyet binárisan szeretnénk klasszifikálni.* **H** *PAC tanulható és $m_H(.,.)$ a minta komplexitása.*

Legyen $0 < \epsilon_1 \leq \epsilon_2 < 1$. Definíció alapján:

ϵ_1-re: $\mathbf{P}(L_{D,f}(A(S_M)) \leq \epsilon_1) \geq 1 - \delta \iff m \geq m_H(\epsilon_1, \delta)$, *illetve*

ϵ_2-re: $\mathbf{P}(L_{D,f}(A(S_M)) \leq \epsilon_2) \geq 1 - \delta \iff m \geq m_H(\epsilon_2, \delta)$

minden $m \in \mathbb{N}$ esetén.

Ha $L_{D,f}(A(S_M)) \leq \epsilon_1$, akkor $L_{D,f}(A(S_M)) \leq \epsilon_2$, így ha $m = m_H(\epsilon_1, \delta)$, akkor teljesül az első egyenlőség, de ekkor a második egyenlőség is teljesül, tehát $m_H(\epsilon_1, \delta) \geq m_H(\epsilon_2, \delta)$. δ-ra hasonlóan bizonyítható.

□

Az előbb felépített modellt könnyen tovább általánosíthatjuk, így a megoldani kívánt feladatok egy jóval nagyobb halmaza is vizsgálhatóvá/modellezhetővé válhat. Ezen általánosításokat a következő két aspektusból fogjuk elemezni:

Az eddigiek során elvártuk, hogy egy $\mathbf{D}$ eloszlás szerint mintázott, f függvény által címkézett bemenetként kapott tanító adathalmaz mellett fennálljon a kielégíthetőségi feltétel, azaz f-hez képes az algoritmus a hipotézisosztályból egy olyan hipotézist visszaadni, hogy azok 1 valószínűséggel megegyeznek. Ám sok alapvető tanulási feladatnál ez a feltevés túl erősnek ígérkezik, vajon tényleg feltehető, hogy létezik olyan predikció, amely az összes adatpontra a megfelelő címkét rendeli? Előre reflektálva a következő alfejezetben vizsgált modellre - ahol feloldjuk a kielégíthetőségi feltételt -, mennyire tükrözi az a valóságot, hogy egy adatpont koordinátái által meghatározott tulajdonságok halmaza egyértelműen indukálja a címkét. Azaz valóban feltehető, hogy minden ponthoz csak egy címkét rendelünk? Például tekintsük egy hitelelbíráló feladatát, aki a havi jövedelem, a rendelkezésr álló fedezet és a kor alapján dönti el, hogy az ügyfél megkapja-e a kívánt hitelösszeget. Ebben a kontextusban nyílván naiv feltételezés lenne két, az előbb említett három paraméterben megegyező ügyfél esetén elutasítani azt az esetet, hogy különböző elfogadó státuszba kerüljenek a hitelezőnél.

Talán a legtermészetesebben adódó, az eddig leírtak általánosítására szolgáló feltevés az a címkehalmaz méretének bővítése. Eddig feltettük, hogy minden mintapontot kétféle osztályba sorolhatunk, azonban ezen osztályok számát egészen a megszámlálhatóan végtelenig bővítve különböző feladatok megoldhatóságai válnak elérhetővé számunkra. Például ha a péntek déli hőmérsékletre szeretnénk egy regressziós feladatot megoldani, akkor valós számokkal kell dolgoznunk, ahol még a címkehalmaz végessége sem tehető fel, míg véges címkehalmaz esetén történő klasszifikáció például a holnapi Nemzeti Sport címlapján szereplő cikk témája.

2.2. Agnosztikus PAC tanulás

A 2. fejezetben vizsgált modell során feltettük,hogy a kielégíthetőségi feltétel teljesül, ami nem volt más, mint $\exists h^* \in \mathbf{H} : \mathbf{P}[h^*(x) = f(x)] = 1$. Az előző alfejezetben leírtak alapján ezen feltétel elhagyásával egy általánosabb, szélesebb feladatkört megoldó tanuló modellt kapunk. Mivel azt a problémát, hogy azonos adatpont esetén különböző címkehalmazba eső változóink is lehetnek az f függvény mellett, nem lehet kezelni, ezért hagyjuk el a címkéző függvényt. Az f használata helyett vezessünk be egy adat-címkegeneráló eloszlást,

amelyre az $\mathbf{X} \times \mathbf{Y}$ Borel-halmazain értelmezett valószínűségi mértékként tekintünk, ahol továbbra is $\mathbf{X}$ az alaphalmaz és $\mathbf{Y}$ a címkehalmaz. Az eddig használt adatgeneráló eloszlás a továbbiakban az új eloszlás $\mathbf{D}_x$-szel jelölt marginálisa, míg a $\mathbf{D}((x,y) \mid x)$ annak a valószínűségét adja, hogy a megfelelően reprezentált x adatponthoz az y kijelölt címke tartozik.

A paradigmaváltás során újra kell definiálnunk a tapasztalati és a valódi rizikót. A $\mathbf{D}$ eloszlással mérhetjük egy h hipotézis hibáját egy véletlenszerűen generált címkézett adatponthalmazon. Egy h predikció valódi rizikóját a következőképpen definiáljuk:

$$L_D(h) \stackrel{def}{=} \mathbf{P}_{(x,y)\sim D}[h(x) \neq f(x)] \stackrel{def}{=} \mathbf{D}(\{(x,y) : h(x) \neq y\}).$$

Egy olyan hipotézist keresünk, amely minimalizálja a valódi kockázatot, de mivel a tanuló algoritmusnak nincs ismerete az eloszlásról, így ebben az esetben is a hibát a tapasztalati rizikóval mérhetjük, melynek az új modell melletti definíciója:

$$L_S(h) \stackrel{def}{=} \frac{|i \in [m] : h(x_i) \neq y_i|}{m}.$$

Adott S tanítóhalmaz esetén, $L_S(h)$ bármely $h : \mathbf{X} \to \{0,1\}$ függvény esetén az algoritmus számára kiszámolható. Megfigyelhető, hogy a tapasztalati kockázat nem más, mint az S-en egyenletes eloszlás valódi hibája. A cél ennél a modellnél sem más, mint megtalálni az optimális $h : \mathbf{X} \to \mathbf{Y}$ predikciót, amely "nagy valószínűséggel" közelítőleg minimalizálja a valódi kockázatot.

A kételemű $\{0,1\}$ címkehalmaz mellett, egy tetszőleges $\mathbf{D}$ eloszlás esetén a legpontosabban előrejelző függvény, a Bayes optimális jósló függvény:

$$f_D(x) = \begin{cases} 1 & \text{ha } \mathbf{P}[y = 1|x] \geq 1/2 \\ 0 & \text{egyébként} \end{cases}$$

2.2.1. Állítás. *Minden* $\mathbf{D}$ *valószínűségi eloszlás esetén a Bayes jósló,* f_D *optimális, azaz minden* $h : \mathbf{X} \to \{0,1\}$ *klasszifikáló esetén* $L_D(f_D) \leq L_D(h)$.

Bizonyítás.

Egy h hipotézis valódi rizikója ebben az esetben:

$$L_D(h) = \mathbf{D}(\{(x,y) : h(x) \neq y\}) = \mathbf{P}(h(x) \neq y) = \begin{cases} \mathbf{P}(y \neq 0|x) & ha\ h(x) = 0 \\ \mathbf{P}(y \neq 1|x) & ha\ h(x) = 1 \end{cases}$$

Az optimális klasszifikáció az, amely minimalizálja a következő veszteségfüggvényt:

$$\Phi(x) = \begin{cases} \mathbf{P}(y \neq 0|x) & ha\ h(x) = 0 \\ \mathbf{P}(y \neq 1|x) & ha\ h(x) = 1 \end{cases}$$

$\Longrightarrow$

$$\Phi(x) = \begin{cases} \mathbf{P}(y = 1|x) & ha\ h(x) = 0 \\ 1 - \mathbf{P}(y = 1|x) & ha\ h(x) = 1 \end{cases}$$

Így ha $\mathbf{P}(y = 1|x) < 1 - \mathbf{P}(y = 1|x)$*, akkor* $h(x) = 0$*-t kell választanunk, ha pedig* $\mathbf{P}(y = 1|x) > 1 - \mathbf{P}(y = 1|x)$*, akkor* $h(x) = 1$*-et kell választanunk, egyenlőség esetén tetszőlegesen választhatunk, nem befolyásolja a megoldást.*

Tehát $\mathbf{P}(y = 1|x) < 1 - \mathbf{P}(y = 1|x) \Longrightarrow \mathbf{P}(y = 1|x) < 1/2$*, így adódik a formula* f_D*-re, amely a levezetésből adódóan minimalizálja a* Φ *által mért veszteséget, de* Φ *választásából adódóan* f_D *a legkisebb valódi rizikóval rendelkező hipotézis az előbbiekben definiált problémára.*

□

Sajnos azonban mivel a $\mathbf{D}$ eloszlásról nincs semmi ismeretünk, így nem is tudjuk kihasználni f_D optimalitását. Mindemellett az előbbi mondat nem teljesen helytálló, ugyanis azzal, hogy beláttuk f_D valódi hibájának minimalitását, már fel tudjuk építeni a PAC tanulhatóságból természetesen adódó agnosztikus PAC tanulhatóságot. Az előbbi tétel alapján már tudjuk, hogy egy tanuló algoritmus kimenetként megadott hipotézise nem rendelkezhet kisebb valódi rizikóval, mint a Bayes predikció, így a feladat egy olyan hipotézis keresésére változik, amely valódi kockázata nem tér el túl nagy mértékben az optimális hipotézis valódi kockázatától. Ezek alapján már adódik az alábbi definíció:

2.2.2. Definíció (agnosztikus PAC tanulhatóság). *Egy* $\mathbf{H}$ *hipotézisosztály agnosztikusan PAC tanulható, ha létezik egy* $m_H : (0,1)^2 \to \mathbb{N}$ *függvény és egy tanuló algoritmus a következő tulajdonságokkal: Minden* $\varepsilon, \delta \in (0,1)$ *és minden* $\mathbf{D}$*, az alaphalmaz és címkehalmaz Descartes-szorzata által generált* σ*-algebrán értelmezett valószínűségi eloszlás esetén, ha az algoritmust egy egymástól függetlenül,* $\mathbf{D}$ *eloszlás szerint mintázott elemekkel rendelkező,* $m \geq m_H(\varepsilon, \delta)$ *méretű tanítóadathalmaz mellett futtatva megkapunk egy* h *hipotézist, amelyre legalább* $1 - \delta$ *valószínűséggel igaz, hogy*

$$L_D(h) \leq min_{h' \in H} L_D(h') + \varepsilon$$

A fent leírt definíció mellett, ha feltesszük a kielégíthetőségi feltételt, akkor visszakapjuk a PAC-tanulhatóság definícióját, tehát az előbbi definíció felépítése során helyesen jártunk el, hiszen egy olyan általánosítást kaptunk az agnosztikus PAC tanulhatóság fogalmával, amely generálja a PAC tanulhatóságot. A fő különbség a két modell között az, hogy míg a PAC tanulás során a hipotézisnek abszolút kicsi hibával kellett rendelkeznie, addig az agnosztikus verzió esetén egy relatív, az elérhető minimális legkisebb hiba mellett keresünk kis rizikójú hipotézist, azaz az agnosztikus modell sokkal jobban illeszkedik a probléma nehézségére, tanulhatósági minőségére.

2.3. A tanuló modell kiterjesztése

Az eddig vizsgált modelleinket ebben a fejezetben tovább általánosítjuk, mellyel egy jóval szélesebb, bonyolultabb problémakör megoldására alkalmas tanulási módszert kapunk.

Az első általánosítás, amely nem kíván nagyobb, mélyreható változtatásokat a modellünkön a többosztályos klasszifikáció. Ezzel a kiterjesztéssel azt a korábbi feltevésünket oldjuk fel, mely szerint a címkehalmaz legyen kételemű, azaz minden adatponthoz kétféle "tulajdonságot" jósolhatunk. A bináris klasszifikációval például a Nemzeti Sport cikkeinek sportágak szerinti klasszifikációja nem megoldható feladat. Tegyük fel, hogy a Nemzeti Sport egy minőségi újság, amely nem csak két sportágról közvetít eredményeket, ezáltal a címkehalmaznak a sportágakat tartalmazó véges halmaznak kell lennie. Itt az alaphalmaz az összes megjelent cikk a Nemzeti Sportban, melynek elemezhető reprezentációja: az egyes cikkek megfelelő tulajdonságaival elkódolt vektorokkal, mint adatpontokkal feleltetjük meg a cikkeket. (Ilyen tulajdonságok lehetnek például az egyes sportágakra jellemző kulcsszavak, az elért időeredmények és hasonló, a sportágak között a könnyű partícionálhatósághoz vezető tulajdonságok) Az előbb leírt példa megoldása nem különbözik sokban az eddig vizsgáltaktól, a tanító adathalmaz most is egy véges sorozat, melynek elemei címkézett adatpontok, a tanuló algoritmustól outputként ebben az esetben is egy predikciót várunk a pontokhoz rendelendő címkékre. Az egyetlen különbség, hogy az adatpontokhoz rendelhető címkék száma most nem feltétlenül csak kettő, hanem tetszőleges véges szám. Vegyük észre, hogy az előbbi engedmény mellett nincs szükség a hibaformulák megváltoztatására sem, tehát az eddigiekben vizsgált modellezéshez képest nem kapunk lényegében más feladatot, ha egy többosztályos klasszifikációs feladattal állunk szemben.

A következő általánosítással kapott feladatot regressziós feladatnak nevezzük. Regressziós problémák esetén egyfajta mintázatot keresünk az adatban, egy függvénnyel meg-

adható kapcsolatot szeretnénk felállítani az $\mathbf{X}$ és $\mathbf{Y}$ halmazok között. Például a tizennégy éves kosárlabdázó gyerekek felnőttkori magasságuk, illetve a csontsűrűségük, a nemük és a jelenlegi magasságuk között fennálló mintázatok alapján egy lineáris függvényt szeretnénk találni optimális predikció gyanánt a jósolt felnőttkori magasságra, tehát az alaphalmaz $\mathbb{R}^3$-nak részhalmaza. A példában szereplő probléma esetén az adatpontokhoz viszont valójában a címkézéshez hasonlóan valós számokat fogunk rendelni, így regresszió esetén $\mathbf{Y}$-t címkehalmaz helyett célhalmaznak fogjuk nevezni, azonban a tanító adathalmaz továbbra is az eddig használt véges sorozat marad. A feladat felépítése során nyilvánvalóan adódik, hogy a predikció sikerének mértékét az eddig látottaktól eltérően kell formalizálni. A továbbiakban regressziós probléma esetén egy $h : \mathbf{X} \to \mathbf{Y}$ hipotézis minőségét a valódi "címkék" és a jósolt "címkék" közötti várható négyzetes eltéréssel is mérhetjük, tehát

$$L_D(h) \stackrel{def}{=} \mathbf{E}_{(x,y)\sim D}(h(x) - y)^2$$

A fenti formula láttán természetesen merül fel az a gondolat, hogy a hipotézis minőségének mértékére esetleg egy általánosabb képletet bevezetve nincs-e lehetőségünk a problémák formálisabb felírására. A fejezet hátralévő részében ezt a kérdést fogjuk elemezni.

Legyen adott egy tetszőleges $\mathbf{H}$ halmaz, amely megadja az adott problémára használható hipotézisek rendszerét és egy $\mathbf{Z}$ alaphalmaz, legyen l egy $\mathbf{H} \times \mathbf{Z}$-ből a pozitív valós számokra képező függvény, ekkor l-t veszteségfüggvénynek nevezzük. A predikciós problémák esetén $\mathbf{Z} = \mathbf{X} \times \mathbf{Y}$ alakban áll elő. Figyeljük meg, hogy a most vizsgált, általánosított veszteségfüggvény bevezetésénél nem kötöttük ki, hogy egy klasszifikációs problémára adott hipotézis minőségét szeretnénk vele vizsgálni, tehát egy fent leírt l függvénnyel formálisabban vizsgálhatjuk tetszőleges problémák hibáját is.

Most definiáljuk a rizikófüggvényt, mint egy hipotézis várható vesztesége, azaz egy $h \in \mathbf{H}$ esetén a $\mathbf{Z}$ által generált σ-algebrán értelmezett $\mathbf{D}$ eloszlásra nézve a rizikófüggvény a következő:

$$L_D(h) \stackrel{def}{=} \mathbf{E}_{z\sim D}[l(h, z)]$$

Minden $h \in \mathbf{H}$ esetén az $l(h, .) : \sigma(\mathbf{Z}) \to \mathbb{R}_+$ függvényre valószínűségi változóként tekintünk és a valódi rizikó, $L_D(h)$ a az előbb leírt valószínűségi változó várható értéke. Tehát az előbbi definíció a következő kikötéssel válik teljessé, $l(h, .)$ függvény legyen mérhető a generált σ-algebrára nézve.

Tekintsük h várható veszteségét egy $\mathbf{D}$ eloszlás szerint vett z véletlen mintaelemnek. Hasonlóan ahhoz, ahogy a tapasztalati rizikót definiáltuk, a várható veszteség egy adott $S = (z_1, ...z_m) \in \mathbf{Z}^m$ minta esetén,

$$L_S(h) \stackrel{def}{=} \frac{1}{m} \sum_{i=1}^{m} l(h, z_i).$$

A veszteségfüggvényt tehát alkalmazhatjuk klasszifikációs és regressziós feladat megoldása során egyaránt. Ezen két típus veszteségmérésére mutatunk egy-egy példát:

A 0-1 veszteséget bináris vagy többosztályos klasszifikációs problémák esetén használjuk, és a leíró formula alapján megfigyelhető, hogy ebben az esetben a rizikófüggvény megegyezik a korábbiakban definiált valódi rizikóval. A valószínűségi változóink a $\sigma(\mathbf{X} \times \mathbf{Y})$-ból képeznek és a veszteségfüggvény az alábbi alakú:

$$l_{0-1}(h, (x, y)) \stackrel{def}{=} \begin{cases} 0 & \text{ha } h(x) = y \\ 1 & \text{ha } h(x) \neq y \end{cases}$$

A négyzetes veszteség regressziós problémák esetén használható, ahol az előző példához hasonlóan a z valószínűségi változónk szintén a $\sigma(\mathbf{X} \times \mathbf{Y})$-ból képez, míg a veszteség függvény a következő:

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

A fejezet során végzett számítások és észrevételek után végre eljutottunk oda, hogy formálisan kiáltalánosítottuk az agnosztikus PAC tanulási modellt és definiálhatjuk az agnosztikus PAC tanulhatóságot általános veszteségfüggvény mellett.

2.3.1. Definíció. *Egy* $\mathbf{H}$ *hipotézisosztály agnosztikusan PAC tanulható egy* $\mathbf{Z}$ *halmazra és egy* $l : \mathbf{H} \times \mathbf{Z} \to \mathbb{R}_+$ *függvényre nézve, ha létezik egy* $m_H : (0,1)^2 \to \mathbb{N}$ *függvény és egy tanuló algoritmus a következő feltételekkel: Minden* $\varepsilon, \delta \in (0,1)$ *és minden a* $\mathbf{Z}$ *által generált* σ*-algebrán értelmezett* $\mathbf{D}$ *eloszlás esetén,* $m \geq m_H(\varepsilon, \delta)$ *méretű, függetlenül,* $\mathbf{D}$ *eloszlás szerint mintázott tanító adathalmazra az algoritmus egy olyan* $h \in \mathbf{H}$ *eloszlást ad kimenetként, amelyre legalább* $1 - \delta$ *valószínűséggel teljesül, hogy*

$$L_D(h) \leq min_{h' \in H} L_D(h') + \varepsilon$$

,ahol $L_D(h) = \mathbf{E}_{z \sim D}[l(h, z)]$.

3. fejezet

Tanulás az egyenletes konvergencia fennállása esetén

Az első fejezetben vizsgált tanuló modell, a PAC modell esetén megmutattuk, hogy a kielégíthetőségi feltétel mellett minden véges hipotézisosztály PAC tanulható. Ebben a fejezetben pedig egy olyan eszközt keresünk, amely bizonyítékot ad egy véges hipotézisosztály agnosztikus PAC tanulhatóságára.

Mielőtt bevezetnénk formálisan az egyenletes konvergenciát, vizsgáljuk meg , hogy mi ezen feltétel alapötlete, mit fejez ki intuitívan. Emlékeztetőül egy adott $\mathbf{H}$ hipotézisosztály mellett az ERM-szabály a következőképpen működik: Egy bemenetként kapott S tanítóadathalmaz esetén a tanuló kiszámolja minden $h \in \mathbf{H}$ esetén a tapasztalati rizikót, majd pedig kimenetként megad egy, a tapasztalati rizikót minimalizáló hipotézist. Azt szeretnénk elvárni, hogy egy ilyen predikció optimális legyen a valódi rizikóra nézve is, azaz a legkisebb tapasztalati rizikóval rendelkező hipotézis maximálizálja az adatgeneráló eloszlásból kinyerhető információ-mennyiséget. Tehát a feltétel azt írja elő, hogy minden h hipotézis tapasztalati kockázata jó közelítése legyen a hipotézis valódi hibájának.

3.0.1. Definíció (ε-reprezentatív minta). *Egy S tanító adathalmazt ε-reprezentatívnak nevezünk a $\mathbf{Z}$ alaphalmazra, $\mathbf{D}$ eloszlásra, $\mathbf{H}$ hipotézisosztályra és l veszteségfüggvényre nézve, ha*

$$\forall h \in \mathbf{H}, |L_S(h) - L_D(h)| \leq \varepsilon.$$

A következő egyszerű állítással megmutatjuk, hogy ha egy minta $(\varepsilon/2)$-reprezentatív, akkor az ERM tanuló egy jó hipotézist fog adni.

3.0.2. Állítás. *Tegyük fel, hogy egy S tanító adathamaz $(\varepsilon/2)$-reprezentatív a megfelelő $\mathbf{Z}, \mathbf{D}, \mathbf{H}, l$ mellett. Ekkor az $ERM_H(S)$ minden $h_S \in argmin_{h\in H} L_S(h)$ kimenetként adott hipotézisére teljesül, hogy*

$$L_D(h_S) \leq min_{h\in H} L_D(h) + \varepsilon.$$

Bizonyítás. *h_S valódi rizikójára a minta előbb definiált tulajdonsága miatt a következő egyenlőtlenség teljesül:*

$$L_D(h_S) \leq L_S(h_S) + \frac{\varepsilon}{2},$$

h_S-t az ERM szabály alapján választottuk, így felülről lehet becsülni a tapasztatlati rizikóját bármely hipotézisosztálybeli predikció tapasztalati rizikójával:

$$L_S(h_S) + \frac{\varepsilon}{2} \leq L_S(h) + \frac{\varepsilon}{2}, \quad \forall h \in \mathbf{H}.$$

Ismét felhasználva, hogy S $(\varepsilon/2)$-reprezentatív, igaz a következő egyenlőtlenség és ezzel teljes a bizonyítás.

$$L_S(h) + \frac{\varepsilon}{2} \leq L_D(h) + \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = L_D(h) + \varepsilon$$

□

Az előbbi állítás alapján ahhoz, hogy az ERM szabály agnosztikus PAC tanuló, elegendő azt megmutatni, hogy legalább $1-\delta$ valószínűséggel egy véletlen tanító adathalmaz ε-reprezentatív, amely követelményt a következő definíció formalizálja.

3.0.3. Definíció (Egyenletes konveregncia). *Azt mondjuk, hogy egy $\mathbf{H}$ hipotézisosztály rendelkezik az egyenletes konvergencia tulajdonsággal a $\mathbf{Z}$ alaphalmaz és egy l veszteségfüggvény mellett, ha létezik egy $m_H^{UC} : (0,1)^2 \to \mathbb{N}$ függvény úgy, hogy minden $\varepsilon, \delta \in (0,1)$ és minden, $\mathbf{Z}$ által generált σ-algebrán éretlmezett $\mathbf{D}$ eloszlás esetén, ha S egy $m \geq m_H^{UC}(\varepsilon, \delta)$ méretű minta, mely elemei egymástól függetlenül, $\mathbf{D}$ eloszlás szerint generáltak, akkor S ε-reprezentatív legalább $1-\delta$ valószínűséggel.*

Hasonlóan a PAC tanulhatóság minta komplexitásának definíciójához, m_H^{UC} is egy alsó korlátot ad a minta azon méretére, amely mellett kielégíthető az egyenletes konvergencia, azaz hány darab mintapontra van szükségünk, hogy legalább $1-\delta$ valószínűséggel a minta ε-reprezentatív legyen. A bevezetett tulajdonságban szereplő egyenletesség arra utal, hogy egy rögzített mintaméret mellett a hipotézisosztály minden elemére és minden valószinűségi eloszlásra fennáll az előbb bevezetett feltétel.

3.0.4. Következmény. *Ha egy* $\mathbf{H}$ *hipotézisosztály rendelkezik az egyenletes konvergencia tulajdonsággal egy* m_H^{UC} *függvénnyel, akkor az osztály agnosztikusan PAC tanulható az* $m_H(\varepsilon, \delta) \leq m_H^{UC}(\varepsilon/2, \delta)$ *minta komplexitással.*

3.0.5. Tétel. *A véges osztályok agnosztikusan PAC tanulhatók az ERM algoritmust használva.*

Bizonyítás. Az előző következtetés alapján minden véges osztály agnosztikus PAC tanulhatóságához elég megmutatnunk, hogy az egyenletes konvergencia véges hipotézisosztály esetén fennáll. Ehhez rögzítsünk le egy ε-t és egy δ-t, mutatnunk kell egy m mintaméretet, amelyre minden $\mathbf{D}$ eloszlás esetén egy $S = (z_1, ..., z_m)$ független, azonos $\mathbf{D}$ eloszlás szerint mintázott tanító adathalmaz esetén $1 - \delta$ valószínűséggel igaz, hogy $\forall h \in \mathbf{H}, |L_S(h) - L_D(h)| \leq \varepsilon$, azaz

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

Ezzel ekvivalensen, a komplementer eseménnyel felírva

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

Vegyük észre, hogy

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

Alkalmazva, hogy a valószínűségi eloszlás mint mérték σ-szubadditív, a következő egyenlőtlenséget kapjuk

$$\mathbf{D}^m(\{S : \exists h \in \mathbf{H}, |L_S(h) - L_D(h)| > \varepsilon\}) \leq \sum_{h \in H} \mathbf{D}^m(\{S : |L_S(h) - L_D(h)| > \varepsilon\}). \quad (3.1)$$

Most megmutatjuk, hogy a szummában szereplő tényezők elég nagy m-re "elég kicsik", azaz megmutatjuk, hogy minden rögzített $h \in \mathbf{H}$-ra a valódi és tapasztalati rizikó közötti eltérés, $|L_D(h) - L_S(h)|$ kicsi. Emlékeztetőül a valódi rizikó $L_D(h) = \mathbf{E}_{z \sim D}[l(h, z)]$ és a tapasztalati kockázat $L_S(h) = \frac{1}{m} \sum_{i=1}^{m} l(h, z_i)$, de mivel minden z_i független, azonos $\mathbf{D}$ eloszlású, így az $l(h, z_i)$ valószínűségi változó várható értéke $\forall i$-re $L_D(h)$. Felhasználva, hogy az integrál lineáris operátor, $L_S(h)$-nak szintén $L_D(h)$ a várható értéke, tehát a becsülendő $|L_D(h) - L_S(h)|$ mennyiség nem más, mint az $L_S(h)$ valószínűségi változó várható értékétől való eltérése. A feladatunk így nem más, mint megmutatni, hogy $L_S(h)$

a várható értékére koncentrálódik, amely olvasatán kézenfekvőnek tűnhet a nagy számok erős törvényének alkalmazása. Ez azonban hibás lépés lenne, mert a nagy számok erős törvénye egy aszimptotikus eredményt ad $m \to \infty$ esetén, de a mi feladatunk a különbség valódi értékének becslése egy előre megadott véges m méretű mintára. Ebben az esetben a megfelelő megoldás egy valószínűségi változó és várható értéke közötti eltérés közelítésére, ha mértékelméleti átalakításokból nyerünk becslést a keresett különbségre.

Legyen $\xi_i := l(h, z_i) - L_D(h)$, illetve $\overline{\xi} := \frac{1}{m}\sum_{i=1}^{m} \xi_i$. Kihasználva az exponenciális függvény monotonitását és alkalmazva a Markov-egyenlőtlenséget tetszőleges $\lambda > 0$-ra,

$$\mathbf{P}(\overline{\xi} \geq \varepsilon) = \mathbf{P}(e^{\lambda\overline{\xi}} \geq e^{\lambda\varepsilon}) \leq e^{-\lambda\varepsilon}\mathbf{E}[e^{\lambda\overline{\xi}}].$$

Mivel független valószínűségi változóink vannak, így $\overline{\xi}$ definíciójából adódóan a következő átalakítások igazak

$$\mathbf{E}[e^{\lambda\overline{\xi}}] = \mathbf{E}[\prod_{i=1}^{m} e^{\frac{\lambda\xi_i}{m}}] = \prod_{i=1}^{m} \mathbf{E}[e^{\frac{\lambda\xi_i}{m}}] \tag{3.2}$$

Most becsüljük felülről tetszőleges i-re $\mathbf{E}[e^{\frac{\lambda\xi_i}{m}}]$-et. Feltehető, hogy l értelmezési tartománya a $[0,1]$ intervallum, ekkor figyeljük meg, hogy $\mathbf{E}[\xi_i] = 0$, illetve $\xi_i \in [0,1]$. Mivel $f(\xi_i) = e^{\frac{\lambda}{m}\xi_i}$ konvex függvény, így $\forall\alpha \in (0,1)$-re

$$f(\xi_i) \leq \alpha f(0) + (1-\alpha)f(1)$$

Válasszuk $\alpha = \dfrac{1-\xi_i}{1-0}$-t,

$$e^{\frac{\lambda}{m}\xi_i} \leq \frac{1-\xi_i}{1-0}e^{\frac{\lambda}{m}1} + \frac{\xi_i - 0}{1-0}e^{\frac{\lambda}{m}0} = (1-\xi_i)e^{\frac{\lambda}{m}} + \xi_i.$$

Várható értéket véve és kihasználva, hogy az integrál lineáris operátor azt kapjuk, hogy

$$\mathbf{E}[e^{\frac{\lambda}{m}\xi_i}] \leq (1 - \mathbf{E}[\xi_i])e^{\frac{\lambda}{m}} + \mathbf{E}[\xi_i] = e^{\frac{\lambda}{m}}$$

Így a következő egyenlőtlenség igaz minden $\lambda \geq \frac{m}{8}$-ra (a későbbiekben látjuk, hogy ez a megszorítás valójában nem zár ki vizsgálandó eseteket),

$$\mathbf{E}[e^{\frac{\lambda}{m}\xi_i}] \leq e^{\frac{\lambda^2}{8m^2}} \tag{3.3}$$

Így most már visszatérhetünk $\mathbf{P}(\overline{\xi} \geq \varepsilon)$ becslésére, a (3.2), illetve a (3.3) egyenlőtlenségekből azt kapjuk, hogy

$$\mathbf{P}(\overline{\xi} \geq \varepsilon) \leq e^{-\lambda\varepsilon}\prod_{i=1}^{m} e^{\frac{\lambda^2}{8m^2}} = e^{-\lambda\varepsilon + \frac{\lambda^2}{8m}}$$

Így a $\lambda = 4m\varepsilon$ választással

$$\mathbf{P}(\overline{\xi} \geq \varepsilon) \leq e^{-2m\varepsilon^2}.$$

Illetve $(-\overline{\xi})$-re a megfelelő változókkal kapjuk, hogy $\mathbf{P}(\overline{\xi} \leq -\varepsilon) \leq e^{-2m\varepsilon^2}$, amelyből $\mathbf{P}(|\overline{\xi}| \geq \varepsilon)$-ra kaptunk egy felső becslést:

$$\mathbf{P}(|\overline{\xi}| \geq \varepsilon) \leq 2e^{-2m\varepsilon^2}$$

Vegyük észre, hogy $|\overline{\xi}| = |\frac{1}{m}\sum_{i=1}^{m} \xi_i| = |\frac{1}{m}\sum_{i=1}^{m} l(h, z_i) - L_D(h)| = |L_S(h) - L_D(h)|$, azaz megkaptunk egy felső becslést a keresett $\mathbf{D}^m(\{S : |L_S(h) - L_D(h)| > \varepsilon\})$ mennyiségre, így a (3.1) egyenlőtlenséget folytatva:

$$\mathbf{D}^m(\{S : \exists h \in \mathbf{H}, |L_S(h) - L_D(h)| > \varepsilon\}) \leq \sum_{h \in H} 2e^{-2m\varepsilon^2} = 2|\mathbf{H}|e^{-2m\varepsilon^2}$$

Tehát a következő átalakítások után

$$2|\mathbf{H}|e^{-2m\varepsilon^2} \leq \delta \Longrightarrow \frac{2|\mathbf{H}|}{\delta} \leq e^{2m\varepsilon^2} \Longrightarrow m \geq \frac{log(2|\mathbf{H}|)/\delta}{2\varepsilon^2},$$

ha egy olyan m számot választunk, amelyre teljesül, hogy

$$m \geq \frac{log(2|\mathbf{H}|/\delta)}{2\varepsilon^2},$$

akkor

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

□

3.0.6. Következmény. *Legyen* $\mathbf{H}$ *egy véges hipotézisosztály, legyen* $\mathbf{Z}$ *alaphalmaz,* $l : \mathbf{Z} \to [0, 1]$*, akkor* $\mathbf{H}$*-re teljesül az egyenletes konvergencia*

$$m_H^{UC}(\varepsilon, \delta) \leq \lceil log(2|\mathbf{H}|/\delta)/2\varepsilon^2 \rceil$$

minta komplexitással.
Sőt, az osztály agnosztikus PAC tanulható az ERM algoritmust használva

$$m_H(\varepsilon, \delta) \leq m_H^{UC}(\varepsilon/2, \delta) \leq \lceil 2log(2|\mathbf{H}|/\delta)/\varepsilon^2 \rceil$$

minta komplexitással.

3.0.7. Megjegyzés. *Könnyen látható a bizonyításból és a* 3.0.6. *következményből, hogy ha nem tesszük fel, hogy a veszteségfüggvény a* $[0, 1]$*-be képez, hanem egy tetszőleges* $[a, b]$ *intervallumba, akkor a minta komplexitás a következőképpen változik:*

$$m_H(\varepsilon, \delta) \leq m_H^{UC}(\varepsilon/2, \delta) \leq \lceil 2log(2|\mathbf{H}|/\delta)(b-a)^2/\varepsilon^2 \rceil.$$

4. fejezet

A torzítás-komplexitás dilemma

A 2. fejezetben láttuk, hogy ha nem vagyunk figyelmesek, akkor a tanító adat félrevezetheti a tanulót, amely eredménye a túlilleszkedés. Ezen hiba kiküszöbölésére azt a megoldást választottuk, hogy a hipotézist egy megadott $\mathbf{H}$ osztályból vártuk. Az előre megadott hipotézisosztályra tekinthetünk úgy, mint a tanuló valamely, a feladattal kapcsolatos előismeretére való reflektálás. Tekintsük a következő példát: Tegyük fel, hogy életünkben most hagyjuk el a hazánkat először és valamely más éghajlati övezetbe utazunk, ahol a gyümölcsöket nem ismerjük, illetve tegyük fel, hogy a gyümölcsök az egyetlen táplálékforrás. Ekkor az ismeretlen gyümölcsökről szeretnénk eldönteni, hogy finomak-e vagy sem, azaz egy bináris klasszifikációt szeretnénk alkalmazni az előbb leírt problémára. Mivel hazánkban is teremnek gyümölcsök, ezáltal ismerjük a jellegzetességeiket, így egy egyszerű ötlet az, hogy a gyümölcsöket reprezentáljuk kételemű vektorokként a színük, illetve a keménységük alapján, tehát az adatpontjaink egy $[0,1] \times [0,1]$ négyzetbe esnek, ahol az egyik tengely a keménységet adja meg a puhától a kőkeményig, míg a másik tengely a színt zöldtől a barnáig. Ekkor az eddigi előismereteink alapján azt várjuk el, hogy egy négyzet jó predikciót adhat a számunkra ismeretlen gyümölcs finomságát illetően, azaz a klasszifikációs problémában megadtunk egy $\mathbf{H}$ hipotézisosztályt, amelyből egy jó predikciót várunk.

Vajon tényleg szükséges előismeret a tanuló sikeréhez? Esetleg létezhet univerzális tanuló, amely bármely probléma esetén felhasználható anélkül, hogy támaszkodnánk előzetes ismereteinkre a feladattal kapcsolatban? Pontosítsuk ezeket a kérdéseinket az eddig bevezetett definíciók, megfigyelések mellett. Egy ismeretlen, $\sigma(\mathbf{X} \times \mathbf{Y})$-n értelmezett $\mathbf{D}$ eloszlás által definiált speciális tanulási feladat, ahol a tanuló célja egy olyan $h : \mathbf{X} \to \mathbf{Y}$ predikció megtalálása, amely $L_D(h)$ kockázata elég kicsi. A kérdés tehát az, hogy vajon

létezik-e egy olyan A tanuló algoritmus és egy m mintaméret, hogy minden $\mathbf{D}$ eloszlás esetén, ha A egy m elemű, egymástól függetlenül, azonos $\mathbf{D}$ eloszlás szerint mintázott adatpontokból álló tanító adathalmazt kap bemenetként, akkor nagy valószínűséggel a kimenetként megadott hipotézis kis valódi rizikóval rendelkezik.

A fejezet első felében megadjuk a formális választ az előbb feltett kérdésekre, amelyet a No-Free-Lunch tétel karakterizál. A tétel azt állítja, hogy nem létezik univerzális tanuló, pontosabban a tétel bináris klasszifikáció probléma esetén kimondja, hogy minden tanulóra létezik egy valószínűségi eloszlás, amelyen kudarcot vall. Azt mondjuk, hogy a tanuló kudarcot vall, ha egy elemenként függetlenül, azonos eloszlás szerint mintázott adathalmaz mellett a tanuló valószínűleg nagy rizikóval rendelkezik, mondjuk $\geq 0,3$; amíg létezik egy másik tanuló, amely ugyanerre a feladatra kis rizikóval rendelkezik. Más szóval a No-Free-Lunch tétel szerint nem létezik tanuló, amely az összes tanulható feladatra alkalmas, minden tanuló számára léteznek olyan problémák, amelyekre elbukik, míg más tanulók sikeresen alkalmazhatók erre a problémára. Tehát egy bonyolult tanulási feladat megoldása során szükségünk van előismeretekre a $\mathbf{D}$ eloszlással kapcsolatban. Ilyen előzetes ismeret a $\mathbf{D}$ eloszlást illetően lehet például, amikor a PAC tanulás definiálásakor feltettük, hogy létezik olyan h hipotézis az előre meghatározott $\mathbf{H}$ hipotézisosztályból, amelyre a $L_D(h) = 0$. Egy gyengébb előismeret lehet a $\mathbf{D}$ eloszlásról az a feltétel, mely szerint $min_{h \in H} L_D(h)$ kicsi. Az eloszlásra vonatkozó előbbi gyengébb előfeltétel mellett, mint $\mathbf{D}$-re vonatkozó előismeret mellett vizsgáltuk az agnosztikus PAC tanulhatóságot, ahol biztosítottuk, hogy a megadott hipotézis hibája nem lehet sokkal nagyobb, mint a $min_{h \in H} L_D(h)$ érték.

A fejezet második felében pedig a hipotézisosztályok, mint az előismeretek formalizálására szolgáló eszköz választásának előnyeit, illetve hátrányait fogjuk vizsgálni, ehhez két osztályba fogjuk partícionálni az ERM algoritmus egy $\mathbf{H}$ osztály feletti hibáját. Az első komponens az előzetes ismeretek minőségét adja vissza, amelyet a hipotézisosztály elemeinek legkisebb hibájával mérünk, $min_{h \in H} L_D(h)$, ezt a komponenst approximációs hibának vagy torzításnak nevezzük. A második komponens pedig az a hiba, amely következtében túlilleszkedés léphet fel, ez függhet az osztály méretétől, illetve a komplexitásától és becslési hibának nevezzük. Ez a két hibatag kompromisszumot igényel a következőkkel kapcsolatban: még komplexebb $\mathbf{H}$ választása, amely csökkenti a torzítást, de nagyobb eséllyel vezet túlilleszkedéshez, illetve egy kevésbé komplex $\mathbf{H}$ között, amely mellett viszont nő a torzítás, de csökken a túlilleszkedés lehetősége.

4.0.1. Tétel (No-Free-Lunch). *Legyen A egy tetszőleges, bináris klasszifikáció feladat megoldására alkalmas algoritmus a $0-1$ veszteségfüggvényre nézve, egy $\mathbf{X}$ alaphalmaz felett. Legyen m egy tetszőleges $|\mathbf{X}|/2$-nél kisebb, a tanító adathalmaz méretét reprezentáló szám. Ekkor létezik egy $\mathbf{D}$ valószínűségi eloszlás $\sigma(\mathbf{X} \times \{0,1\})$ úgy, hogy*
a) Létezik egy $f : \mathbf{X} \to \{0,1\}$ függvény $L_D(f) = 0$ valódi rizikóval.
b) Legalább $1/7$ valószínűséggel az $S \sim \mathbf{D}^m$ választása mellett $L_D(A(S)) \geq 1/8$.

Bizonyítás. Legyen $\mathbf{C}$ egy $2m$ elemű részhalmaza $\mathbf{X}$-nek. Az intuíciónk a bizonyítással kapcsolatban az, hogy bármely tanuló algoritmus, amely $\mathbf{C}$ adatpontjainak csak a felét figyeli meg, nincs információja arról, hogy a többi $\mathbf{C}$-beli adatponthoz mely címkének kellene tartoznia. Tehát létezik egy olyan f függvény, amely ellentmond $A(S)$ algoritmus által, a $\mathbf{C}$ vizsgálatlan adatpotjaihoz rendelt címkéknek, ahol tekinthetünk f-re, mint a valóságot tükröző címkéző függvényre.

Figyeljük meg, hogy $T = 2^{2m}$ darab $\mathbf{C}$-ből $\{0,1\}$-be képező függvény létezik, jelöljük ezeket $f_1, ..., f_T$-vel. Minden ilyen függvényre legyen $\mathbf{D}_i$ a $\sigma(\mathbf{C} \times \{0,1\})$-en értelmezett valószínűségi eloszlás úgy, hogy

$$\mathbf{D}_i(\{(x,y)\}) = \begin{cases} 1/|\mathbf{C}| & \text{ha } y = f_i(x) \\ 0 & \text{különben} \end{cases}$$

Ekkor nyilván $L_{D_i}(f_i) = 0$, hiszen egy olyan (x,y) pár választásának valószínűsége, amelyre y megegyezik a valódi, f_i által meghatározott címkével $1/|\mathbf{C}|$, míg az $y \neq f_i(x)$ esetnek 0 a valószínűsége.

Most megmutatjuk, hogy minden A algoritmusra, amelynek egy m elemű $C \times \{0,1\}$-ből mintázott tanító adathalmazhoz van hozzáférése és egy $A(S) : \mathbf{C} \to \{0,1\}$ függvényt ad vissza, teljesül a következő

$$max_{i \in [T]} \mathbf{E}_{S \sim D_i^m}[L_{D_i}(A(S))] \geq 1/4. \tag{4.1}$$

Ez azt jelenti, hogy minden A' algoritmusra, amely inputként megkap egy m elemű $\mathbf{X} \times \{0,1\}$-ből képzett tanító adathalmazt, létezik egy $f : \mathbf{X} \to \{0,1\}$ függvény és egy $\mathbf{D}$ $\sigma(\mathbf{C} \times \{0,1\})$-en értelmezett eloszlás, amelyre $L_D(f) = 0$ és

$$\mathbf{E}_{S \sim D^m}[L_D(A'(S))] \geq 1/4 \tag{4.2}$$

Most megmutatjuk, hogy $\mathbf{P}[L_D(A'(S)) \geq 1/8] \geq 1/7$. Jelöljük $\xi = L_D(A'(S))$, akkor ξ egy olyan valószínűségi változó, amely $[0,1]$-ből vesz fel értékeket, és amely várható értékére

igaz, hogy $\mathbf{E}[\xi] \geq 1/4$. Legyen $\eta = 1 - \xi$, η egy nemnegatív valószínűségi változó, melyre $\mathbf{E}[\eta] = 1 - \mathbf{E}[\xi] \leq 3/4$, alkalmazzuk a Markov-egyenlőtlenséget:

$$\mathbf{P}[\xi \leq 1/8] = \mathbf{P}[1 - \xi \leq 7/8] = \mathbf{P}[\eta \leq 7/8] \leq \frac{\mathbf{E}[\eta]}{7/8} \leq \frac{3/4}{7/8} = \frac{6}{7}$$

Így a következő becslést kapjuk, $\mathbf{P}[\xi \geq 1/8] \geq 1/7$, azaz $\mathbf{P}[L_D(A'(S)) \geq 1/8] \geq 1/7$, amit be akartunk látni.

Most bebizonyítjuk a (4.2) egyenlőtlenséget. $k = (2m)^m$ lehetséges különböző m hosszú sorozat képezhető $\mathbf{C}$-ből, jelöljük ezeket a sorozatokat $S_1, ..., S_k$-val, míg ha az $S_j = (x_1, ..., x_m)$ adatpontsorozatot az f_j függvénnyel címkézzük, akkor ezt a tanító adathalmazt jelöljük $S_i^j = ((x_1, f_j(x_1)), ..., (x_m, f_j(x_m)))$-mel. Ha a valószínűségi eloszlás $\mathbf{D}_i$, akkor az A algoritmus számára elérhető tanító adathalmazok $S_1^i, ..., S_k^i$, illetve minden adathalmaz mintázásának a valószínűsége megegyezik, ezért

$$\mathbf{E}_{S \sim \mathbf{D}_i^m}[L_{D_i}(A(S))] = \frac{1}{k} \sum_{j=1}^{k} L_{D_i}(A(S_j^i)). \tag{4.3}$$

Kihasználva, hogy a maximum nagyobb vagy egyenlő, mint az átlag, illetve a minimum nem nagyobb, mint az átlag, kapjuk a következőt:

$$max_{i \in [T]} \frac{1}{k} \sum_{j=1}^{k} L_{D_i}(A(S_j^i)) \geq \frac{1}{T} \sum_{i=1}^{T} \frac{1}{k} \sum_{j=1}^{k} L_{D_i}(A(S_j^i)) = \frac{1}{k} \sum_{j=1}^{k} \frac{1}{T} \sum_{i=1}^{T} L_{D_i}(A(S_j^i)) \geq$$

$$\geq min_{j \in [k]} \frac{1}{T} \sum_{i=1}^{T} L_{D_i}(A(S_j^i)) \tag{4.4}$$

Rögzítsünk le egy $j \in [k]$-t, legyen $S_j = (x_1, ..., x_m)$ és legyen $v_1, ..., v_p$ azon adatpontok $\mathbf{C}$-ből, amelyek nem szerepelnek S_j-ben, ahol nyilván $p \geq m$. Ezért minden $h : \mathbf{C} \to \{0, 1\}$ függvényre és minden i-re igaz, hogy

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

$$\geq \frac{1}{2} \sum_{r=1}^{p} \mathbb{1}_{[h(v_r) \neq f_i(v_r)]} \tag{4.5}$$

Tehát

$$\frac{1}{T} \sum_{i=1}^{T} L_{D_i}(A(S_j^i)) \geq \frac{1}{T} \sum_{i=1}^{T} \frac{1}{2p} \sum_{r=1}^{p} \mathbb{1}_{[A(S_j^i)(v_r) \neq f_i(v_r)]} = \frac{1}{2p} \sum_{r=1}^{p} \frac{1}{T} \sum_{i=1}^{T} \mathbb{1}_{[A(S_j^i)(v_r) \neq f_i(v_r)]} \geq$$

$$\geq \frac{1}{2} min_{r\in[p]} \frac{1}{T} \sum_{i=1}^{T} \mathbb{1}_{[A(S_j^i)(v_r)\neq f_i(v_r)]} \tag{4.6}$$

Most rögzítsünk le egy $r \in [p]$-t. Minden $f_1, ..., f_T$ függvényt $T/2$ diszjunkt párba partícionálhatunk, ahol egy $(f_i, f_{i'})$ párra minden $c \in \mathbf{C}$ esetén $f_i(c) \neq f_{i'}(c)$ akkor és csak akkor, ha $c = v_r$. Mivel minden párra $S_j^i = S_j^{i'}$-nek fenn kell állnia, így

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

amiből következik, hogy

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

Felhasználva a (4.6), a (4, 4), illetve a (4.3) egyenlőtlenségeket, megkapjuk, hogy a (4.1) egyenlőtlenség teljesül, amivel a bizonyítás teljes. □

A fejezet elején feltett kérdésre reflektálva, a No-Free-Lunch tétel, mint eredmény hogy viszonyul az előzetes ismeretek szükségességéhez? Tekintsünk egy ERM tanulót az összes lehetséges $\mathbf{X}$-ből a $\{0, 1\}$-be képező függvények $\mathbf{H}$ halmazán, mely osztály reprezentálja az előzetes ismeretek hiányát, hiszen minden az alappontokból a címkehalmazba képező függvényt egy lehetséges predikcióként figyelembe vesz az algoritmus. A No-Free-Lunch tétel szerint minden algoritmus, amely egy $\mathbf{H}$ hipotézisosztályból választja a kimeneti hipotézist, kudarcot vall valamilyen tanulási feladaton, tehát ez az osztály nem PAC tanulható. Ezt a gondolatmenetet formalizálja az alábbi következmény:

4.0.2. Következmény. *Legyen* $\mathbf{X}$ *egy végtelen alaphalmaz és legyen* $\mathbf{H}$ *az összes* $\mathbf{X}$*-ből* $\{0, 1\}$*-be képező függvények halmaza. Ekkor* $\mathbf{H}$ *nem PAC tanulható.*

Bizonyítás. Indirekt tegyük fel, hogy a fent leírt hipotézisosztály PAC tanulható, ehhez válasszunk egy $\varepsilon < 1/8$, illetve $\delta < 1/7$ paramétereket. A PAC tanulhatóság definíciójából léteznie kell egy $m = m(\varepsilon, \delta)$ számnak úgy, hogy minden $\sigma(\mathbf{X} \times \{0, 1\})$-en értelmezett $\mathbf{D}$ adatgeneráló eloszlás esetén, ha egy $f : \mathbf{X} \to \{0, 1\}$ függvényre $L_D(f) = 0$, akkor legalább $1 - \delta$ valószínűséggel teljesül, hogy az A algoritmust egy m elemű egymástól függetlenül, azonos $\mathbf{D}$ eloszlás szerint generált S tanító adathalmazon futtatva, $L_D(A(S)) \leq \varepsilon$. Azonban most akalmazva a No-Free-Lunch tételt, mivel $|\mathbf{X}| > 2m$ minden tanuló algoritmusra, így speciálisan az általunk megfigyelt A algoritmusra létezik egy $\mathbf{D}$ eloszlás, amire legalább $1/7 > \delta$ valószínűséggel teljesül, hogy $L_D(A(S)) > 1/8 > \varepsilon$, ez azonban ellentmondáshoz vezetett, tehát a következményben megfogalmazott állítást bebizonyítottuk. □

De hogy előzhetők meg az ehhez hasonló hibák? Elkerülhetjük ezen lehetséges veszélyeket a No-Free-Lunch tételt alkalmazva: A vizsgált problémával kapcsolatos előismereteink alapján egy speciális hipotézisosztály felállításával elkerülhetjük a kudarchoz vezető eloszlásokat a feladat tanulása során, például a hipotézisosztály leszűkítéséhez is vezethet az előzetes ismereteink felhasználása. Nyilván ekkor felmerül a kérdés, hogyan választható ki egy jó hipotézisosztály? Az egyik szempontból szeretnénk garantálni, hogy ez az osztály tartalmazza a 0 valódi rizkóval rendelkező jóslást (lásd PAC tanulhatóság), vagy legalább azt, hogy a hipotézisosztályból választható olyan függvény, amely hibája az elérhető legkisebb valódi rizikótól csak kevéssel tér el (lásd agnosztikus PAC tanulhatóság). A másik szempontból viszont láttuk, hogy nem választhatjuk a legbővebb, az alaphalmazon értelmezett összes függvényt tartalmazó osztályt. Ezen kompromisszum vizsgálatát a következő alfejezetben tesszük meg.

4.1. A hiba felbontása

Az előbb feltett kérdés megválaszolásához osszuk fel egy ERM_H jósló hibáját két részre a követező módon, legyen h_S egy ERM_H hipotézis, ekkor

$$L_D(h_S) = \epsilon_{app} + \epsilon_{est}, \text{ ahol } \epsilon_{app} = min_{h \in H} L_D(h), \ \epsilon_{est} = L_D(h_S) - \epsilon_{app}. \tag{4.7}$$

Az approximációs hiba vizsgálata során tegyük fel, hogy a minimális kockázat elérhető egy, a hipotézisosztályban szereplő predikció által. A hibának ez a tagja megadja annak a mértékét, hogy mekkora rizikót vállalunk a szóba jöhető predikciók halmazának egy speciális **H** hipotézisosztályra való leszűkítésével, azaz mekkora az induktív torzítás. Az approximációs hiba nem függ a tanító adathalmaz méretétől, csupán az általunk előre megadott hipotézisosztálytól, azaz **H** méretének növelésével csökkenthető a hibának ezen része.

4.1.1. Megjegyzés. *A kielégíthetőségi feltétel mellett az approximációs hiba* 0*, míg ezen feltétel mellett ez a tag meg is nőhet. Figyeljük meg, hogy ebben az esetben* ϵ_{app} *mindig tartalmazza a Bayes optimális predikció hibáját, az elkerülhetetlen minimális hibát, amely a modell által vizsgált világ nemdeterminisztikusságát karakterizálja. Bizonyos esetekben célszerű az approximációs hibát a* $min_{h \in H} L_D(h)$ *helyett* $min_{h \in H} L_D(h) - \epsilon_{Bayes}$*-vel, a többlet hibával definiálni.*

Az ϵ_{est} becslési hibát az approximációs hiba és az ERM hipotézis által elért hiba közti különbség adja meg. Ezen hiba annak okán lép fel, hogy a tapasztalati kockázat a valódi kockázat egy becslése, tehát a predikció, amely minimalizálja a tapasztalati rizikót, csupán közelítése annak a predikciónak, amely minimalizálja a valódi rizikót. A becslés minősége függ a tanító adathalmaz méretétől, illetve a hipotézisosztály méretétől vagy komplexitásától. Ahogy korábban megmutattuk, véges hipotézisosztály esetén az ϵ_{est} hiba logaritmikusan nő $|\mathbf{H}|$ méretével arányosan, míg a tanító adathalmaz m méretével arányosan csökken.

Mivel a célunk minimalizálni a valódi kockázatot, így szembekerülünk egy optimalizálási feladattal, amelyet torzítás-komplexitás optimalizálásnak nevezünk. Egyfelől egy gazdag $\mathbf{H}$ osztály választása csökkenti az approximációs hibát, de ezzel együtt talán növeli a becslési hibát, mivel egy túl gazdag hipotézisosztály túlilleszkedéshez vezethet. Másfelől viszont $\mathbf{H}$-t egy kisméretű halmaznak választva növeljük az approximációs hibát, de ezzel együtt csökkentjük a becslési hibát, amely alulilleszkedéshez vezethet. Természetesen egy jó választás lehet $\mathbf{H}$-nak egy egyelemű halmaz, amely csak a Bayes optimális jóslót tartalmazza, de ebben az esetben a klasszifikáló az ismeretlen $\mathbf{D}$ eloszlástól függ.

A tanuláselmélet foglalkozik azzal, hogy elfogadható becslési hiba mellett milyen gazdagnak választhatjuk a $\mathbf{H}$ hipotéziosztályt. Sok esetben az empirikus kutatás egy jó hipotézisosztály megtalálására fokuszál egy bizonyos alaphalmaz esetén, itt a "jó" azt jelenti, hogy az approximációs hiba nem szélsőségesen nagy. Az ötlet az, hogy nem vagyunk tökéletesek és nem tudjuk, hogy hogy érdemes az optimális predikciót megalkotni, de rendelkezünk előzetes ismeretekkel a feladattal kapcsolatban, amely lehetővé teszi olyan hipotézisosztályok felállítását, melyekre mind az approximációs, mind a becslési hiba kicsi. Visszatérve a gyümölcsös példánkhoz, mivel nem ismerjük azt a predikciót, amely pontosan megmondja a számunkra ismeretlen gyümölcsről, hogy finom-e, viszont a számunkra eddig megismert gyümölcsök alapján tudjuk, hogy a szín-keménység által reprezentált térben a téglalap egy jó predikció lehet.

5. fejezet

A VC-dimenzió

Ebben a fejezetben a fő célunk a PAC tanulható hipotézisosztályok karakterizálása, azaz egy **H** hipotézisosztályról hogy dönthető el a PAC tanulhatósága. A korábbiakban láttuk, hogy a véges osztályok PAC tanulhatók, míg egy végtelen alaphalmaz felett az összes lehetséges hipotézis-függvényből álló halmaz nem PAC tanulható. Akkor vajon a tanulhatóság a halmaz végességével van kapcsolatban, létezhet végtelen méretű hipotéziosztály? Mi alapján állapítható meg az egyik halmazról, hogy tanulható, míg a másikról, hogy nem? Ezekre a kérdésekre keressük a választ a következőkben.

5.1. Nem-véges osztályok tanulhatósága

A 3. fejeteben bebizonyítottuk, hogy a véges osztályok tanulhatók és hogy egy hipotézisosztály mintakomplexitása felülről becsülhető az osztály méretének logaritmusával. Egy példát mutatva elvetjük a tanulhatóság azon lehetséges feltételét, hogy a hipotézisosztály mérete megfelelő karakterizációt ad a minta komplexitására, azaz mutatunk egy egyszerű végtelen osztályt, ami tanulható.

5.1.1. Állítás. *Legyen* $\mathbf{H} := \{h_a : a \in \mathbb{R}\}$*, ahol* $h_a : \mathbb{R} \to \{0,1\}$ *függvény a következő* $h_a(x) = \mathbb{1}_{[x<a]}$*, azaz* **H** *végtelen méretű. Ekkor* **H** *PAC tanulható az ERM-szabályt használva* $m_H(\varepsilon, \delta) \leq \lceil log(2/\delta)/\varepsilon \rceil$ *mintakomplexitással.*

Bizonyítás. Legyen a^* küszöb úgy, hogy $h^*(x) = \mathbb{1}_{[x<a^*]}$ valódi kockázata $L_D(h^*) = 0$. Legyen $\mathbf{D}_X$ a **D** eloszlás **X**-re vett marginálisa és legyen $a_0 < a^* < a_1$ úgy, hogy

$$\mathbf{P}_{x\sim D_X}[x \in (a_0, a^*)] = \mathbf{P}_{x\sim D_X}[x \in (a^*, a_1)] = \varepsilon.$$

Ha $\mathbf{D}_X(-\infty, a^*) \leq \varepsilon$, akkor legyen $a_0 := -\infty$, hasonlóan a_1 esetében. Legyen adott egy S tanító adathalmaz, legyen $b_0 := max\{x : (x, 1) \in S\}$ és $b_1 := min\{x : (x, 0) \in S\}$, ahol szintén ha nincs negatív elem S-ben, akkor legyen $b_1 = \infty$, hasonlóan b_0 esetében. Legyen b_S egy megfelelő küszöb egy h_S ERM hipotézishez, ahol ezek alapján $b_S \in (b_0, b_1)$, így $L_D(h_S) \leq \varepsilon$-ra egy jó bizonyíték, hogy $b_0 \geq a_0, b_1 \leq a_1$:

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

Kihasználva, hogy a $\mathbf{D}$ eloszlás mérték lévén σ-szubadditív kapjuk, hogy

$$\mathbf{P}_{S\sim D^m}[L_D(h_S) > \varepsilon] \leq \mathbf{P}_{S\sim D^m}[b_0 < a_0] + \mathbf{P}_{S\sim D^m}[b_1 > a_1]. \tag{5.1}$$

A $b_0 < a_0$ esemény akkor és csak akkor következik be, ha nem létezik S-beli elem az (a_0, a^*) intervallumból, melynek a valószínűsége ε, illetve abból, hogy S iid minta a következő átalakítások alkalmazhatók,

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

Mivel feltettük, hogy $m > log(2/\delta)/\varepsilon$, így az előbb kapott mennyiség legfeljebb $\delta/2$, az előbbiekhez hasonlóan $\mathbf{P}_{S\sim D^m}[b_1 > a_1] \leq \delta/2$, ezt felhasználva a (6.1) egyenlőtlenségből: $\mathbf{P}_{S\sim D^m}[L_D(h_S) > \varepsilon] \leq \delta$, illetve $\mathbf{P}_{S\sim D^m}[L_D(h_S) \leq \varepsilon] \leq 1 - \delta$.

□

5.2. A VC-dimenzió fogalmának bevezetése

Láttuk, hogy a hipotézisosztály végessége szükséges, de nem elégséges feltétele a tanulhatóságnak. Meg fogjuk mutatni, hogy az osztály VC-dimenzójának nevezett tulajdonság megfelelő karakterizációt ad a tanulhatóságnak. A VC-dimenzió bevezetéséhez elevenítsük fel a No-Free-Lunch-tételt, illetve annak bizonyítását. Megmutattuk, hogy a hipotézisosztály valamely előzetes ismeret segítségével történő leszűkítés nélkül konstruálható olyan valószínűségi eloszlás, amelyre a tanuló algoritmus gyengén teljesít, míg létezik egy másik algoritmus, amely sikeres ugyanarra az eloszlásra. Ezen eloszlás megalkotására a tétel bizonyításában vettünk egy véges $\mathbf{C} \subset \mathbf{X}$ halmazt és tekintettük azt az eloszláscsaládot, amely $\mathbf{C}$ elemeire koncentrálódott. Minden eloszlás egy "valódi" $\mathbf{C} \to \{0, 1\}$ célfüggvényből származtatott. Az algoritmus kudarcához felhasználtuk az összes $\mathbf{C} \to \{0, 1\}$ függvényt, mint lehetséges választást. Amikor a PAC tanulhatóságot vizsgáljuk, akkor a

lehetséges $h \in \mathbf{H}$ függvények halmazát leszűkítjük azokra, amelyek 0 valódi kockázattal rendelkeznek. Mivel mi valószínűségi eloszlásokat vizsgálunk, amelyek a $\mathbf{C}$ halmaz elemeire koncentrálódnak, ezért célravezetőbb azt vizsgálni, hogy $\mathbf{H}$ hogy viselkedik $\mathbf{C}$-n, így kapjuk a következő definíciót:

5.2.1. Definíció. *Legyen* $\mathbf{H}$ *az* $\mathbf{X} \to \{0,1\}$ *függvények egy halmaza és legyen* $\mathbf{C} = \{c_1, ..., c_m\} \subset \mathbf{X}$. *A* $\mathbf{H}$ *osztály* $\mathbf{C}$*-re való leszűkítése az összes* $\mathbf{H}$*-ból származtatható* $\mathbf{C} \to \{0,1\}$ *függvény, azaz*

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

ahol minden függvényt egy $|\mathbf{C}|$ *hosszú vektor reprezentál.*

Ha a $\mathbf{H}$ hipotézisosztály $\mathbf{C}$-re való leszűkítése az összes $\mathbf{C} \to \{0,1\}$ függvény, akkor azt mondjuk, hogy $\mathbf{H}$ szétválasztja $\mathbf{C}$-t, formálisan:

5.2.2. Definíció. *Egy* $\mathbf{H}$ *hipotézisosztály szétválasztja* $\mathbf{C} \subset \mathbf{X}$*-et, ha* $\mathbf{H}$ $\mathbf{C}$*-re való leszűkítése az összes* $\mathbf{C} \to \{0,1\}$ *függvény, vagyis* $\mathbf{H}_C = 2^{|C|}$.

Vizsgáljunk meg ezzel kapcsolatban egy egyszerű szemléltető példát! Legyen $\mathbf{H}$ az 5.1.1. állításban definiált osztály, válasszuk a $\mathbf{C} = \{c_1\}$ halmazt. Ekkor ha $a = c_1 + 1$, akkor $h_a(c_1) = 1$ és ha $a = c_1 - 1$, akkor $h_a(c_1) = 0$, azaz megadtuk az összes $\mathbf{H}$-ból képezhető $\mathbf{C} = \{c_1\} \to \{0,1\}$ függvényt, tehát ez a halmaz $\mathbf{H}_C$ és $\mathbf{H}$ szétválasztja $\mathbf{C}$-t. Most legyen $\mathbf{C} = \{c_1, c_2\}$, ahol $c_1 \leq c_2$. Ekkor nem létezik olyan $h \in \mathbf{H}$, amely c_1-hez 0-t, míg c_2-höz 1-et rendel, mert ha egy a $\mathbf{H}$-ban szereplő függvény c_1-hez 0-t rendel, akkor ez a h c_2-höz csak 0-t tud rendelni, tehát $|\mathbf{H}_C| = 3$, így $\mathbf{H}$ nem választja szét $\mathbf{C}$-t.

Visszatérve a No-Free-Lunch-tétel bizonyításában használt eloszlásra, ha egy $\mathbf{C}$ halmazt szétválasztja $\mathbf{H}$, akkor tetszőleges $\mathbf{C} \to \{0,1\}$ célfüggvényhez konstuálható eloszlás fenntartva a kielégíthetőségi feltételt, amelyből adódik:

5.2.3. Következmény. *Legyen* $\mathbf{H}$ *a* $\mathbf{C}$*-ből* $\{0,1\}$*-be képező függvények egy osztálya, legyen* m *a tanító adathalmaz mérete. Tegyük fel, hogy létezik egy* $2m$ *méretű* $\mathbf{C} \subset \mathbf{X}$ *halmaz, amelyet a* $\mathbf{H}$ *szétválaszt. Ekkor minden* A *tanuló algoritmusra létezik egy* $\sigma(\mathbf{X} \times \{0,1\})$*-en értelmezett* $\mathbf{D}$ *eloszlás és egy* $h \in \mathbf{H}$ *hipotézis úgy, hogy* $L_D(h) = 0$, *de legalább* $1/7$ *valószínűséggel az* $S \sim \mathbf{D}^m$ *választás felett* $L_D(A(S)) \geq 1/8$.

Az 5.2.3. következmény alapján, ha egy $\mathbf{H}$ szétválaszt egy $2m$ méretű $\mathbf{C}$-t, akkor nem tudunk m méretű mintával tanulni. Intuitívan, ha $\mathbf{H}$ szétválasztja $\mathbf{C}$-t és az algoritmus $\mathbf{C}$ adatpontjaiból képzett mintának csak a feléhez fér hozzá, akkor ezen adatpontokhoz rendelt címkék alapján azonban nem rendelkezik információval a többi $\mathbf{C}$-beli adatponthoz rendelendő címkéről, ez vezet minket a VC-dimenzió fogalmához.

5.2.4. Definíció (VC-dimenzió). *Egy* $\mathbf{H}$ *osztály VC-dimenziója, amelyet* $VCdim(\mathbf{H})$*-val jelölünk, a maximális mérete azon* $\mathbf{C} \subset \mathbf{X}$ *halmazoknak, amelyeket* $\mathbf{H}$ *szétválaszt. Ha* $\mathbf{H}$ *tetszőleges nagy méretű részhalmazokat szét tud választani, akkor* $\mathbf{H}$ *végtlen VC-dimenzóval rendelkezik.*

5.2.5. Tétel. *Legyen* $\mathbf{H}$ *egy végtelen VC-dimenzióval rendelkező osztály, akkor* $\mathbf{H}$ *nem PAC tanulható.*

Bizonyítás. Mivel $\mathbf{H}$ végtelen VC-dimenzióval rendelkezik, így minden m méretű tanító adathalmazra létezik $2m$ méretű szétválasztható halmaz, amiből az 5.2.3. következmény alapján a bizonyítás teljes. □

A későbbiekben látjuk, hogy a megfordítás igaz, azaz a véges VC-dimenzió garantálja a PAC tanulhatóságot, tehát a VC-dimenzó karakterizálja a tanulhatóságot, nézzünk néhány példát a bevezetett fogalmak értelmezésére.

5.3. Példák

Ebben az alfejezetben kiszámoljuk néhány hipotézisosztály VC-dimenzióját. Ahhoz, hogy egy $\mathbf{H}$-ról megmutassuk, hogy $VCdim(\mathbf{H}) = d$ a, következőket kell belátni:

1. Létezik egy d méretű $\mathbf{C}$ halmaz, amelyet $\mathbf{H}$ szétválaszt.
2. Egyetlen $(d+1)$ méretű $\mathbf{C}$ halmaz sem szétválasztható $\mathbf{H}$ által.

Az első példa vizsgálatát már részben elvégeztük, tekintsük az 5.1.1. állításban bevezetett hipotézisosztályt, amelyről megmutattuk, hogy tetszőleges $\mathbf{C} = \{c_1\}$ halmaz esetén $\mathbf{H}$ szétválasztja $\mathbf{C}$-t, tehát $VCdim(\mathbf{H}) \geq 1$. Azt is megmutattuk, hogy tetszőleges $\mathbf{C} = \{c_1, c_2\}$ halmaz esetén, ahol $c_1 \leq c_2$, $\mathbf{H}$ nem választja szét $\mathbf{C}$-t, azaz $VCdim(\mathbf{H}) = 1$.

Most legyen $\mathbf{H}$ az $\mathbb{R}$ intervallumainak indikátoraiból képzett osztály, azaz $\mathbf{H} = \{h_{a,b} : a, b \in \mathbb{R}, a < b\}$, ahol $h_{a,b} : \mathbb{R} \to \{0, 1\}$ és $h_{a,b}(x) = \mathbb{1}_{[x \in (a,b)]}$. Vegyük a $\mathbf{C} = \{1, 2\}$ halmazt, ekkor $|\mathbf{H}_C| = 4$, tehát $\mathbf{H}$ szétválasztja $\mathbf{C}$-t, azaz $VCdim(\mathbf{H}) \geq 2$. Most tekintsünk egy

tetszőleges $\mathbf{C} = \{c_1, c_2, c_3\}$ halmazt és tegyük fel, hogy $c_1 \leq c_2 \leq c_3$. Ekkor azonban az $(1, 0, 1)$ címkézés nem származtatható a hipotézisosztályból, hiszen ha c_1, c_3-t tartalmazza egy intervallum, akkor c_2-t is, azaz $\mathbf{H}$ nem választja szét $\mathbf{H}$-t, tehát $VCdim(\mathbf{H}) = 2$.

Tekintsük azt a példát, hogy $\mathbf{H}$ a tengelyekkel párhuzamos téglalapok, formálisan

$$\mathbf{H} = \{h_{(a_1,a_2,b_1,b_2):a_1 \leq a_2, b_1 \leq b_2}\},$$

ahol

$$h_{(a_1,a_2,b_1,b_2)} = \mathbb{1}_{[x_1 \in [a_1,a_2]]} \mathbb{1}_{[x_2 \in [b_1,b_2]]}.$$

Megmutatjuk, hogy $VCdim(\mathbf{H}) = 4$, ehhez mutatnunk kell egy négyelemű halmazt, amelyet $\mathbf{H}$ szétválaszt és meg kell mutatnunk, hogy nem létezhet 5 pont úgy, hogy ezt a halmazt $\mathbf{H}$ szétválasztja. 4 pontot, amely szétválasztható könnyű találni - például $(1, 0), (1, 2), (0, 1), (2, 1)$ - a síkon. Most tekintsünk egy tetszőleges öt pontból álló $\mathbf{C} \subset \mathbb{R}$ halmazt. $\mathbf{C}$-ben vegyük a legkisebb és a legnagyobb első, illetve második koordinátájú pontot, ezeket jelöljük rendre c_1, c_2, c_3, c_4-gyel és a fel nem használt csúcsot c_5-tel. Vizsgáljuk a $(c_1, c_2, c_3, c_4, c_5) \to (1, 1, 1, 1, 0)$ címkézést, de ezt lehetetlen az általunk vizsgált hipotézisosztályból származtatni, mert c_5 mindkét koordinátája kisebb valamely másik pont megfelelő koordinátájánál, illetve nagyobb is valamely másik pont megfelelő koordinátájánál. Ezért ha az első négy ponthoz egyet rendelünk, akkor az ötödikhez is csak egy rendelhető, tehát $\mathbf{C} = \{c_1, c_2, c_3, c_4, c_5\}$-t nem választja szét $\mathbf{H}$, azaz $VCdim(\mathbf{H}) = 4$.

Mit mondhatunk a véges hipotézisosztályokról a most bevezetett karakterizáció alapján? Legyen $\mathbf{H}$ egy véges osztály, ekkor nyilván minden $\mathbf{C}$ halmazra $|\mathbf{H}_C| \leq |\mathbf{H}|$ és emiatt $\mathbf{C}$ nem választható szét $\mathbf{H}$ által, ha $|\mathbf{H}| < 2^{|\mathbf{C}|}$, amiből kapjuk, hogy $VCdim(\mathbf{H}) \leq log_2(|\mathbf{H}|)$. Ebből látható, hogy véges halmazok PAC tanulhatósága következik egy még általánosabb megállapításból, a véges VC-dimenzióval rendelkező osztályok PAC tanulhatóságából, amelyet a következő alfejezetben vizsgálunk.

5.3.1. Megjegyzés. *Egy véges* $\mathbf{H}$ *osztály VC-dimenzója szignifikánsan kisebb lehet* $log_2(|\mathbf{H}|)$*-nál. Például legyen* $\mathbf{X} = \{1, ..., k\}$ *és tekintsük az* 5.1.1. *állításban definiált osztályt, ekkor* $|\mathbf{H}| = k$*, de* $VCdim(\mathbf{H}) = 1$*, ahol* k *tetszőlegesen nagy lehet, tehát a* $log_2(|\mathbf{H}|)$ *és a* $VCdim(\mathbf{H})$ *közti különbség tetszőlegesen nagy lehet.*

5.3.2. Megjegyzés. *Az eddigi példákban a VC-dimenzió megegyezett az osztályt definiáló paraméterek számával, de ez nem minden esetben igaz, konstruálható olyan példa, ahol egy paraméter mellett a VC-dimenzió végtelen.*

5.4. A PAC tanulás alaptétele

A korábbiakban megmutattuk, hogy végtelen VC-dimenzióval rendelkező osztály nem tanulható, ennek a fordítottja is igaz, amely a statisztikai tanulás alaptételéhez vezet:

5.4.1. Tétel (A statisztikai tanulás alaptétele). *Legyen* $\mathbf{H}$ *egy* $\mathbf{X} \to \{0,1\}$ *függvényekből álló hipotézisosztály* $\mathbf{X}$ *alaphalmaz felett és válasszuk a* $0-1$ *veszteségfüggvényt. Ekkor a következők ekvivalensek:*

1. $\mathbf{H}$ *rendelkezik az egyenletes konvergencia tulajdonsággal.*
2. *Minden ERM-szabály eredményes agnosztikus PAC tanuló* $\mathbf{H}$*-ra.*
3. $\mathbf{H}$ *agnosztikus PAC tanulható.*
4. $\mathbf{H}$ *PAC tanulható.*
5. *Minden ERM-szabály eredményes PAC tanuló* $\mathbf{H}$*-ra.*
6. $\mathbf{H}$ *véges VC-dimezióval rendelkezik.*

Bizonyítás. A 3. fejezetben beláttuk az $1 \to 2$-t, illetve a $2 \to 3$, $3 \to 4$ és $2 \to 5$ irányok triviálisak. A $4 \to 6$ és $5 \to 6$ a No-Free-Lunch-tételből következik. Az egyetlen nehéz implikáció, amelyet bizonyítanunk kell: a $6 \to 1$, ennek bizonyításához két fő lépés vezet.

Legyen $VCdim(\mathbf{H}) = d$, ekkor annak ellenére, hogy $\mathbf{H}$ akár végtelen méretű is lehet, ha $\mathbf{H}$-t leszűkítjük $\mathbf{C}$-re, akkor $|\mathbf{H}_C| \in \mathcal{O}(|\mathbf{C}|^d)$, amely mennyiség $|\mathbf{C}|$ méretében polinomiálisan nő. Ezt felhasználva a Sauer-lemmát fogjuk alkalmazni, amihez először meg kell ismerkednünk egy fogalommal.

5.4.2. Definíció (Növekedés függvény). *Legyen* $\mathbf{H}$ *egy hipotézisosztály, ekkor* $\mathbf{H}$ *növekedés függvénye,* $\tau_H : \mathbb{N} \to \mathbb{N}$*:*

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

Azaz $\tau_H(m)$ azon különböző függvények száma, amelyek egy m méretű $\mathbf{C}$ halmazból $\{0,1\}$-be képeznek és $\mathbf{H}$ osztály $\mathbf{C}$-re való leszűkítéséből nyerhetők, nyilvánvalóan, ha $VCdim(\mathbf{H}) = d$, akkor minden $m \leq d$-re $\tau_H(m) = 2^m$. A következő lemma azt állítja, hogy ha m nagyobb, mint a VC-dimenzió, akkor a növekedésfüggvény mindössze polinomiálisam nő m méretében.

5.4.3. Állítás (Sauer-lemma). *Legyen* $\mathbf{H}$ *egy* $VCdim(\mathbf{H}) \leq d < \infty$ *VC-dimenzióval rendelkező hipotézisosztály. Ekkor minden* m*-re,* $\tau_H(m) \leq \sum_{i=0}^{d} \binom{m}{i}$*, különösen ha* $m > d+1$*, akkor* $\tau_H(m) \leq (em/d)^d$

Ezt az állítást nem bizonyítjuk. A 3. fejezetben beláttuk, hogy a véges osztályok rendelkeznek az egyenletes konvergencia tulajdonsággal. Megmutatjuk, hogy az egyenletes konvergencia fennállása mellett a hipotézisosztály "tényleges mérete" kicsi, azaz $|\mathbf{H}_C|$ polinomiálisan nő $|\mathbf{C}|$-ben, formálisan:

5.4.4. Tétel. *Legyen* $\mathbf{H}$ *egy hipotézisosztály és legyen* τ_H *a növekedés függvénye. Ekkor minden* $\mathbf{D}$ *eloszlásra és minden* $\delta \in (0,1)$*-re legalább* $1-\delta$ *valószínűséggel* $S \sim \mathbf{D}^m$ *választás mellett teljesül, hogy*

$$|L_D(h) - L_S(h)| \leq \frac{4 + \sqrt{log(\tau_H(2m))}}{\delta\sqrt{2m}}$$

Ezek alapján már könnyen beláthatjuk a bizonyítás teljességéhez hiányzó $6 \to 1$ irányt, ehhez tehát elég belátni, hogy ha véges a VC-dimenzió, akkor fennáll az egyenletes konvergencia. Megmutatjuk, hogy

$$m_H^{UC}(\varepsilon,\delta) \leq 4\frac{16d}{(\delta\varepsilon)^2}log(\frac{16d}{(\delta\varepsilon)^2}) + \frac{16dlog(2e/d)}{(\delta\varepsilon)^2}$$

A Sauer-lemmából tudjuk, hogy $m > d$-re, $\tau_H(2m) \leq (2em/d)^d$, illetve erre az 5.4.4. tételt alkalmazva kapjuk, hogy legalább $1-\delta$ valószínűséggel igaz a következő:

$$|L_D(h) - L_S(h)| \leq \frac{4 + \sqrt{dlog(2em/d)}}{\delta\sqrt{2m}}.$$

Az egyszerűség kedvéért tegyük fel, hogy $\sqrt{dlog(2em/d)} \geq 4$, ami alapján,

$$|L_D(h) - L_S(h)| \leq \frac{1}{\delta}\sqrt{\frac{2dlog(2em/d)}{m}}$$

Ahhoz hogy megmutassuk, hogy az előző mennyiség ε-nál kisebb kell:

$$m \geq \frac{2dlog(m)}{(\delta\varepsilon)^2} + \frac{2dlog(2\varepsilon/\delta)}{(\delta\varepsilon)^2}.$$

Algebrai átalakítások után kapjuk, hogy

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

□

5.4.5. Megjegyzés. *A VC-dimenzió nem csak a PAC tanulhatóságot karakterizálja, hanem meghatározza a minta komplexitást, amelyet a következő tétel formalizál.*

5.4.6. Tétel (A statisztikai tanulás alaptételének kvantitív verziója). *Legyen* $\mathbf{H}$ *az* $\mathbf{X} \to \{0,1\}$ *függvények egy osztálya, ahol* $\mathbf{X}$ *az alaphalmaz, illetve használjuk a 0-1 veszteségfüggvényt. Tegyük fel, hogy* $VCdim(\mathbf{H}) = d < \infty$. *Ekkor létezik* C_1, C_2 *konstans úgy, hogy*

1. $\mathbf{H}$ *rendelkezik az egyenletes konvergencia tulajdonsággal a következő minta komplexitással*

$$C_1 \frac{d + log(1/\delta)}{\varepsilon^2} \leq m_H^{UC}(\varepsilon, \delta) \leq C_2 \frac{d + log(1/\delta)}{\varepsilon^2}.$$

2. $\mathbf{H}$ *agnosztikus PAC tanulható a következő minta komplexitással*

$$C_1 \frac{d + log(1/\delta)}{\varepsilon^2} \leq m_H(\varepsilon, \delta) \leq C_2 \frac{d + log(1/\delta)}{\varepsilon^2}.$$

3. $\mathbf{H}$ *PAC tanulható a következő minta komplexitással*

$$C_1 \frac{d + log(1/\delta)}{\varepsilon} \leq m_H(\varepsilon, \delta) \leq C_2 \frac{d + log(1/\delta)}{\varepsilon}.$$

A fejezet során megadtuk a statisztikai tanulás alaptételének segítségével a PAC tanulhatóság pontos karakterizációját a VC-dimenziókat használva, amely azon maximális méret, amely esetén az alaphalmaz egy részhalmazát a hipotézisosztály szétválasztja. Az alaptétel tehát azt mondja ki, hogy egy osztály akkor és csak akkor PAC tanulható, ha véges VC-dimenzióval rendelkezik és meghatározza az ehhez szükséges minta komplexitást.

Irodalomjegyzék

[1] Shai Shalev-Shwartz and Shai Ben-David , *Understanding Machine Learning*, Cambridge University Press, 2014