Eötvös Loránd Tudományegyetem
Természettudományi Kar

A statisztikai tanuláselmélet alapjai

Csillag Márton

Matematika BSc
Matematikai elemző szakirány

Témavezető:
Neogrády-Kiss Márton
Alkalmazott Analízis és Számításmatematikai Tanszék

Budapest, 2019

Tartalomjegyzék

Bevezetés

Szakdolgozatomban egy manapság igencsak népszerű tudományterület, a gépi tanulás matematikai elméletével foglalkozok. Napjainkban a világunk digitalizálódása miatt iszonyatos mennyiségű adatot termelünk minden egyes percben. Emellett a tárolási kapacitás, az adattárházak mérete is nagymértékben növekszik, ez a két tényező pedig azt eredményezi, hogy a rendelkezésre álló adatmennyiség soha nem látott méreteket ölt. Az emberek természetesen a rendelkezésre álló adatokat szeretnék hasznosítani, számukra hasznos információkat kinyerni belőle. Tulajdonképpen ez az igény hívta életre a gépi tanulás tudományát. Interdiszciplináris területként több ágról építkezik, ide sorolható többek között a valószínűségszámítás, a statisztika, az algoritmuselmélet, a számítástudomány, a komplex rendszerek elmélete és az információelmélet. Bizonyos ágai kapcsolatban állnak a neurobiológiával is.

Az első fejezet egy átfogó képet ad a gépi tanulásról: annak alapvető feladatáról, típusairól. Kitérünk az előzetes tudásunk alkalmazásának fontosságára, illetve egy szemléletes példán keresztül megismerkedünk a terület filozófiai alapvetésével, az Occam borotvája elvvel.

Ezután a statisztikai tanuláselmélet alapfogalmainak ismertetése következik, majd a hipotézisosztályok tanulhatóságának legalapvetőbb formájával, a PAC (valósznűleg közelítőleg helyes) tanulhatósággal foglalkozunk. A modellt később kiterjesztjük, az agnosztikus PAC-tanulhatóság fogalmához vezető lépéseket részletesen megvizsgáljuk. Megismerkedünk a Nincs ingyen ebéd tétellel, amely a terület egyik fontos eredménye. A negyedik fejezetben konkrét osztályok tanulhatóságát vizsgáljuk, majd bevezetjük a Vapnik-Chervonenkis dimenzió fogalmát, amely vizsgálódásunk egyik központi témája. Ehhez kapcsolódóan a hipotézisosztályok növekedési függvényeire is kitérünk. Végezetül kimondjuk és bebizonyítjuk a statisztikai tanuláselmélet alaptételét, amely az addig bevezetett fogalmakat fogja össze.

1. fejezet

A gépi tanulás

Ebben a fejezetben a gépi tanulás alapjaival foglalkozunk, arról beszélünk, hogy milyen feladatokra lehet használni, és mi a célja.

1.1. Mi a gépi tanulás?

A gépi tanulás kifejezést olvasva alapvető kérdések merülhetnek fel bennünk: *Mi az, hogy tanulás? Hogyan tanul egy gép? Mikor sikeres a tanulás?*

Az első kérdésre nincsen általánosan egzakt válaszunk, ám az intuíciónkra támaszkodva a következő megállapításra juthatunk: *A tanulás a tapasztalataink felhasználásával eddig ismeretlen tudás megszerzése.* A feladat megoldásához tehát szükségünk lesz „tapasztalatokra", ez általában egy adathalmaz formájában fog megjelenni. Tegyük fel, hogy van egy feladatunk és egy algoritmusunk, ami megkap bemenetként egy halmazt, és ennek felhasználásával egy hipotézist (modellt) állít fel, amit később az adott feladat megoldására fog használni. Az adathalmaz tanulmányozásával az algoritmus „tapasztalatot" gyűjt, vagyis a halmazban olyan alapvető mintázatokat, összefüggéseket keres és jegyez meg, amelyeket később fel tud használni a feladat megoldására. Erre mondjuk, hogy az algoritmus tanul. Ez matematikailag nehezen megfogható, és nagyon sokféle egzakt definíciót lehet megfogalmazni a konkrét problémára vonatkozóan, mindenesetre hasznos a következő általános megfogalmazás [4]:

1.1. Definíció. *Legyen adott egy T feladatosztály és egy P teljesítmény-mérték. Azt mondjuk, hogy az A algoritmus az E tapasztalatból tanul, ha a P szerint mért teljesítménye javul a T feladatain E felhasználásával.*

Alapvető célunk tehát olyan algoritmus írása, amely adathalmazokban keres általános érvényű törvényszerűségeket, mintázatokat, amelyek segítségével általánosíthatunk addig nem látott adatpontokra. Fontos megjegyezni, hogy az általánosításnak központi szerepe van. Egy hagyományos megoldást úgy képzelhetünk el, mint egy utasítássorozatot, ami minden konkrét helyzetre alkalmazza a megfelelő

választ. Ez esetünkben elég költséges és nem is mindig lehetséges. Ezzel szemben a gépi algoritmusnak nem mondjuk meg explicite hogy mit tegyen, hanem a tanulás által állapítja meg a megfelelő válaszokat. Tehát a gépi algoritmus kimenete szintén egy program vagy egy függvény lesz. Ezt általában elképzelhetjük egy paraméterezett modellként, mint például egy mesterséges neurális hálózat vagy egy lineáris regressziós modell.

Tekintsünk példaként egy jól ismert alkalmazást. Egy olyan algoritmust szeretnénk létrehozni, amely újságcikkek kategorizálását végzi. Ehhez először is megállapítunk véges sok kategóriát, amik szerepelhetnek az általunk tekintett cikkek témájaként (például sport, politika, tudomány, időjárás), ezek lesznek az osztálycímkék. Célunk az adatpontok különböző, diszjunkt osztályokba való sorolása. Az ilyen feladatot klasszifikációnak nevezik. Tanulóhalmazként az algoritmusnak adunk újságcikkeket, amelyek (emberek által) el vannak látva a megfelelő osztálycímkékkel. A tanulás során az algoritmus olyan jellemzőket keres, amelyek az egyes kategóriákba tartozó cikkek sajátosságai. Ez történhet például úgy, hogy bizonyos kulcsszavak együttes jelenléte alapján következtet a cikk témájára. Ekkor a tanulási folyamat végén jó esetben egy olyan algoritmust kapunk kézhez, amely hatékonyan tudja kategorizálni az addig nem látott újságcikkeket is. Az előző definícióra vonatkoztatva a T feladatosztály itt az újságcikkek kategorizálása, az E tapasztalatot az előre kategorizált cikkek jelentik. A teljesítményt több P mérték szerint mérhetjük, a legtermészetesebb például az, hogy az újonnan kapott, cikkek hány százalékán találja el az algoritmus a valós osztálycímkét.

Két olyan alapvető esetet különíthetünk el, amikor a gépi tanulás eszköztárának alkalmazásával jobb megoldást érünk el - ráadásul gyorsabban -, mint a hagyományos programozási módszerekkel:

Emberi képességeket meghaladó feladatok. Nyilvánvaló, hogy egy rendelkezésre álló hatalmas adathalmaz érdemi feldolgozása emberi erővel lehetetlen. Mivel az esetek többségében azt sem tudjuk meghatározni, hogy pontosan mi az a jellemző, amit keresünk, így a hagyományos programok sem használhatóak hatékonyan. Számos sikeres ipari alkalmazást hívtak életre az ilyen típusú feladatok, példáként hozhatjuk a már említett újságcikk-kategorizálást, vagy az interneten lévő célzott hirdetéseket, amelyeket a böngészési előzményeikhez igazítva, személyre szabva kapnak a felhasználók. A tudományos kutatások világából érdemes megemlíteni a csillagászati adatok elemzésére használt algoritmusokat, amelyek segítségével új égitesteket, galaxisokat azonosítanak a kutatók.

Ember által természetesen végzett feladatok. Itt elsősorban nem az adathalmaz hatalmas mérete okozza a problémát, hanem az, hogy nem tudunk olyan

egzakt szabályokat alkotni, amelyeket aztán programkódok formájában leírhatunk. Tekintsük példaként a szövegfelismerést, vagy objektumok azonosítását képeken. Hiába az egyik legalapvetőbb emberi készség, hogy felismerjük a beszédet, mégsem tudunk rámutatni egy-egy kimondott szó kvantifikálható jellemzőire egy rögzített hangfáljban. Bizonyos gépi tanulási modellcsaládok (különösképp a neurális hálózatok) rendkívül jó eredményeket érnek el ilyen jellegű problémák megoldásában.

A gépi tanulási algoritmusok fontos jellemzője a rugalmasság. Ha van egy gépi tanulási algoritmusunk, amit egy adott adathalmazon tanítottunk, ám később számos új adatpontot kapunk, akkor nem kell egy teljesen új modellt létrehozni, - ami rendkívül költséges lenne -, hanem elég a meglévőt továbbtanítani az újonnan kapott adatokkal. Ezáltal algoritmusunkat naprakészen tudjuk tartani, ami az alkalmazások jellegéből adódóan nagyon fontos.

Összefoglalva, a gépi tanulási algoritmusok olyan feladatokat oldanak meg, amelyeket hagyományos programokkal lehetetlen, vagy nagyon nehéz megoldani. Egy gépi algoritmus bemenete az S tanulóhalmaz (tapasztalat), kimenete pedig egy h hipotézis, amely általában egy paraméteres modell, amit a tanulási folyamat során megfigyelt szabályszerűségek alapján határoz meg.

1.2. A tanulás típusai

A gépi tanulási feladatokat több dimenzió mentén különíthetjük el.[7] Fontos megjegyezni azonban, hogy ezek a felbontások nem fednek le minden gépi tanulási problémát, előfordulhat, hogy egy-egy feladat megoldásához egy olyan új megközelítésre van szükség, amely nem sorolható be egyik alábbi kategóriába se. Illetve sokszor a különböző tanulási módszerek kombinálása vezet eredményre.

Felügyelt – felügyelet nélküli. A különbség itt elsősorban a bemenetként kapott adathalmazban van. Egy felügyelt tanuló olyan adatpontokkal dolgozik, amelyeknek címkéje, másnéven célváltozó-értéke adott egy tanító által. A tanulóalgoritmus feladata az újonnan kapott adatpontok célváltozó-értékének jóslása. Ide tartoznak a klasszifikációs feladatok (például a fent említett újságcikk-besorolás), illetve a regressziós problémák, ahol a célváltozó folytonos értékű. A felügyelet nélküli tanulás során az algoritmus kizárólag adatpontokat kap. A feladata lehet az, hogy az adathalmazban bármilyen jellemzőket keressen, amely jellemzők alapján az adatpontok elkülönülnek egymástól. Például rendelkezésünkre állnak vásárlási adatok. A célunk csoportok kialakítása olyan módon, hogy a „hasonló” fogyasztási szokásokkal rendelkező vásárlók egy csoportba tartozzanak. Az algoritmus tehát a talált mintázatok alapján határozza meg, hogy mikor hasonló két vásárló. Másik példaként említhetjük a kívülálló adatpontok (outlierek) detektálását, ezt többek között bankkártya-csalások kiszűrésére alkalmazzák.

Online – batch. Az online és batch tanulókat aszerint különböztetjük meg, hogy míg az előbbinek a környezeti változásokra folyamatosan reagálva kell döntéseket hoznia, utóbbinak elég egy jelentős méretű adathalmaz megvizsgálása után. Például ha a célzott hirdetéseket meghatározó algoritmust tekintjük, akkor annak egy új, kevés böngészési előzménnyel rendelkező felhasználóról ugyanúgy meg kell jósolnia, hogy milyen hirdetések érdekelhetik. Így előfordulhat, hogy az online tanuló kezdetben, a kevés rendelkezésre álló adat miatt gyenge teljesítményt nyújt.

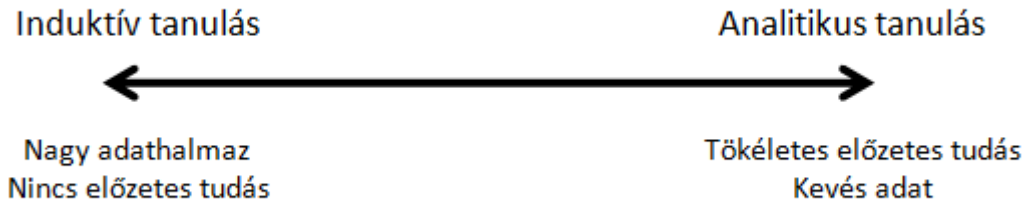
Aktív – passzív. A tanulókat aszerint is elkülöníthetjük, hogy a tanítóval kezdeményeznek-e interakciót. Míg a passzív tanuló csupán a tanítótól kapott adatpontokkal dolgozik, az aktív kéréseket, kérdéseket is intézhet a környezete, vagyis a tanító felé. Az újságcikk-osztályozó algoritmust elképzelhetjük aktív tanulóként is. Ekkor például egy-egy kérdéses, címkézetlen adatpontot kapva megkérdezheti a tanítót, hogy az alternatívák közül melyik a helyes. Olyan is előfordulhat, hogy az algoritmus maga állít össze egy cikket, majd megkéri a tanítót, hogy határozza meg a helyes osztálycímkét. Aktív tanulók alkalmazása bizonyos esetekben nagyon hasznos lehet, hiszen a megfelelő kérdések feltételével az algoritmus biztosabb szabályszerűségeket alkothat.

Megerősítéses tanulás. Egy különálló tanulási forma, amely nem sorolható be egyik említett kategóriába sem, bizonyos alkalmazásokban azonban nagyon elterjedt. Tekintsünk egy olyan algoritmust, amelyet szeretnénk megtanítani sakkozni. A rendelkezésre álló információ ekkor (a megengedett lépések halmazán túl) mindössze az éppen aktuális táblaállás. Az algoritmusnak ezek alapján kell döntést hoznia a következő lépéséről. A megerősítéses tanulás fő jellemzője, hogy az algoritmus egy-egy lépés után nem kap közvetlen visszajelzést annak helyességéről, csak a játszma végén kap jutalmat vagy büntetést annak megfelelően, hogy végül nyert-e [8].

1.3. Az előzetes tudás fontossága

Habár célunk az, hogy algoritmusunk minél inkább „önállóan”, külső beavatkozás nélkül keressen mintázatokat az adathalmazban, a feladatra vonatkozó tudásunkat általában megszorítások formájában be kell építeni az algoritmusba. Például legyen az a feladatunk, hogy általános összefüggéseket keressünk ingatlanok jellemzői és eladási áruk között. Ekkor előfordulhat, hogy az általunk vizsgált adathalmazban az ingatlan házszáma jelentősen korrelál az eladási árral, így a kimenetként kapott modellben erős magyarázó erejű változóként szerepel majd. Józanul belegondola viszont tudhatjuk, hogy ez a kapcsolat kizárólag a véletlennek köszönhető, valós összefüggés nincs a két adat között. Ezt úgy tudjuk megelőzni, ha bizonyos jellemzőket előre kiszűrünk, és csak a releváns jellemzők között kereshet az algoritmus. Ez az előzetes tudás alkalmazásának egyik formája, amikor olyan módon szűkítjük le a lehetséges hipotézisek halmazát (jelen esetben a magyarázó változók szelekciójával), hogy ne kaphassunk olyan hipotézist, ami ellentmond a valósággal, vagy nem felel meg az előzetes elvárásainknak. Szintén ide sorolható, amikor eldöntjük, hogy milyen modellcsaládot alkalmazunk egy adott feladat megoldására. Például ha képeket szeretnénk klasszifikálni, minden bizonnyal konvolúciós neurális hálózatokkal kezdünk el dolgozni, Egy konvolúciós hálózat egy ugyanolyan mély neurális hálózathoz képest sokkal kevesebb paraméterrel rendelkezik. Ennek oka az, hogy azt gondoljuk, egy kép címkéje ugyanaz marad, ha elforgatjuk a képet, vagy ha bizonyos részeit a képen belül arrébb tolunk. [3]

Tehát az előzetes tudás alkalmazása legtöbbször a szóba jöhető hipotézisek halmazának megszorításában nyilvánul meg. Az előzetes tudáshoz kapcsolódóan a tanulási módszereket elkülöníthetjük két véglet szerint az alábbi ábrának megfelelően.



1.1. ábra. *Az induktív és az analitikus tanulás legfőbb jellemzői*

Az induktív tanulási módszerek az adathalmazban lévő mintázatok, szabályszerűségek alapján állítanak fel hipotéziseket, amiket általános érvényűnek fogadnak el. Ezáltal általánosítanak egy sokaság minden elemére pusztán a tanulóhalmazban lévő információ alapján. Az analitikus tanulás ezzel szemben olyan hipotéziseket választ, ami kizárólag az előzetes tudásra támaszkodik, a megkapott adatokat deduktív módon meg tudja magyarázni a tanító által adott szabályok alapján. Megjegyezzük, hogy a legtöbb ma használatos algoritmus leginkább az induktív kategóriába sorolható. A következő táblázat összehasonlítja a két módszer előnyeit és hátrányait. [4]

	Induktív tanulás	**Analitikus tanulás**
Cél:	A hipotézis illeszkedjen az adatokra	A hipotézis feleljen meg a megadott szabályoknak
Módszer:	Statisztikai következtetés	Deduktív következtetés
Előnyök:	Nem kell előzetes tudás	Kevés adatból is képes tanulni
Buktató:	Kis adathalmaz	Hibás előzetes feltételezések

1.1. táblázat. *Az induktív és az analitikus tanulás összehasonlítása*

1.4. Occam borotvája

> *Az egyszerűbb hipotézisek általában jobban megfelelnek a valóságnak, mint a hosszú, összetett magyarázatok.*

Ezt a középkori filozófiai tételt William Ockham skolasztikus gondolkodó rögzítette a 14. századi Angliában. Az elvnek (amelyet még a takarékosság vagy tömörség elvének is szoktak nevezni) számos értelmezése van, azonban mindegyiknek ugyanaz a lényege: a jelenségeket próbáljuk meg a lehető legegyszerűbb módon magyarázni, „borotváljuk le” azokat a plusz feltételezéseket, amelyek nem szükségesek az értelmes magyarázathoz. Például a természettudományos kutatók sokszor azért is preferálják az egyszerűbb elméleteket, mert azok könnyebben tesztelhetők kisérleti úton. De az elv használata matmematikailag is indokolható. Nyilvánvalóan kevesebb egyszerűbb hipotézis van, mint összetett. Így ha egy egyszerűbb hipotézis konzisztens az adatokkal, akkor ennek oka jóval kisebb valószínűséggel lehet statisztikai hiba. Mivel rengeteg nagyon hosszú és összetett hipotézis létezik, ezért már-már törvényszerű, hogy azok közül lesz néhány, ami az adathalmazzal konzisztens. A gyakorlatban is pontosan ezt tapasztalhatjuk. Ha egy adathalmaz összefüggéseire több, egymástól eltérő, de hasonlóan összetett hipotézis is magyarázatot ad, akkor ezek nagy valószínűséggel nem fognak jól általánosítani az adathalmazon kívüli esetekre. [2]

Tekintsünk egy olyan példát, ami egyszerre mutatja meg az Occam borotvája elv és az előzetes tudás alkalmazását [5]. Itt ez utóbbi úgy jelenik meg, hogy a lehetséges hipotéziseknek a bekövetkezési valószínűségeit előre meghatározzuk (valamilyen tapasztalat vagy preferencia alapján).

A feladat a következő: játékpartnerünk kitalál egy egyszerű aritmetikai szabályt, majd véletlenszerűen választ ilyen tulajdonságú számokat az {1,2, ...,100} számhalmazból. Nekünk a kiválasztott számok alapján rá kell jönnünk a szabályra. Lehetséges hipotézisek lehetnek például a 6 többszörösei, 20-nál kisebb számok, prímek, stb.

Legyen $\mathcal{H} = \{h_1, h_2, ..., h_n\}$ az összes lehetséges hipotézis halmaza (megjegyezzük, hogy a jelenleg vizsgált feladatban $\mathcal{H}$ egy véges halmaz, ha a hipotéziseket az {1,2, ...,100} részhalmazaival reprezentáljuk), és jelölje S a rendelkezésünkre álló információt. Ekkor ha h egy hipotézist jelöl, akkor felírható a Bayes tétel:

$$\mathbb{P}(h|S) = \frac{\mathbb{P}(S|h)\mathbb{P}(h)}{\sum_{h_i \in \mathcal{H}} \mathbb{P}(h_i, S)}. \tag{1.1}$$

Azt a h^* hipotézist célszerű választani, amelynek a bekövetkezése a legvalószínűbb a rendelkezésre álló adatok mellett:

$$h^* = \arg\max_{h \in \mathcal{H}} \mathbb{P}(h|S).$$

Tegyük fel, hogy a következő számhalmazt kapjuk: $S = (16,8,2,64)$. Ekkor intuíciónk azt sugallja, hogy a keresett hipotézis a *kettő hatványai* lesz. De miért nem a *páros számok* vagy a *kettő hatványai, kivéve a 32-t* hipotézisekre gondolunk, amik úgyanúgy konzisztensek a rendelkezésre álló adatokkal?
Mivel azt tettük fel, hogy a játékpartnerünk az általa választott hipotézisből véletlenszerűen választ számokat az S sorozatba, így annak a valószínűsége, hogy az N elemű S mintát kapjuk a h hipotézisből:

$$\mathbb{P}(S|h) = \left[\frac{1}{|h|}\right]^N.$$

Ezt *likelihood*nak nevezik. A képletből látható, hogy a modell a kisebb hipotéziseket részesíti előnyben. A jelenlegi feladatban ez pontosan az Occam borotvája elv alkalmazását jelenti. Az $S = (16,8,2,64)$ sorozat esetén a két hipotézisre vonatkozó valószínűség: $\mathbb{P}(S|h_{kh}) = (1/6)^4$ és $\mathbb{P}(S|h_{ps}) = (1/50)^4$. Előbbi majdnem 5000-szerese az utóbbinak, így igazolva látjuk az intuíciónkat, miszerint a *kettő hatványai* hipotézis valószínűbb.
Arra a kérdésre, hogy miért nem a h'_{kh} = *„kettő hatványa, kivéve a 32-t"* hipotézist választjuk, az előzetes feltételezéseink adnak választ. Az előző okfejtés alapján ez ugyanis közel sem lenne egyértelmű: figyeljük meg, hogy $\mathbb{P}(S|h'_{kh}) > \mathbb{P}(S|h_{kh})$. Ám

mivel ez a hipotézis – szemben a másik kettővel – rettentően „természetellenesnek” tűnik, ezért előzetesen nagyon kis valószínűséget adtunk neki, vagyis a *prior* valószínűségnek nevezett $\mathbb{P}(h'_{kh})$ annyira kicsi, hogy 1.1 szerint mégiscsak h_{kh}-t választjuk.

Megjegyezzük, hogy egy hipotézis bonyolultsága függhet az adott nyelvtől amiben leírjuk, ezért az Occam borotvája elvnek a matematikailag pontosabb használata során rögzíteni szoktak egy nyelvet, amivel minden hipotézis le van írva.

2. fejezet

PAC tanulás

Az előző fejezetben általánosan beszéltünk a gépi tanulásról. Ebben a fejezetben bevezetjük a tanulás további vizsgálatához szükséges matematikai kereteket és a PAC (valószínűleg közelítőleg helyes) tanulás fogalmát [9] [7].

2.1. Matematikai keretrendszer

Az alapfeladatot egy szemléletes példán keresztül fogjuk ismertetni: tegyük fel, hogy egy trópusi szigeten kötünk ki. Hamar kiderül, hogy a helyiek kedvelt csemegéje a mangó, amely ráadásul rendkívül tápláló is. Mi azonban még soha nem láttunk ilyen gyümölcsöt, és az őslakosokkal sem tudunk hatékonyan kommunikálni, így csak a helyben gyűjtött tapasztalataink alapján tudunk képet formálni arról, hogy mi határozza meg egy mangó élvezeti értékét. Úgy döntünk, hogy a gyümölcsök színét és keménységét fogjuk a továbbiakban tanulmányozni, és e két jellemzővel próbáljuk magyarázni, hogy az adott mangó finom-e vagy sem. Vagyis összefüggéseket keresünk az adott gyümölcs tulajdonságai és íze között, azaz egy hipotézist állítunk fel. Célunk egy olyan hipotézis, amelynek segítségével később nagy valószínűséggel finom mangót tudunk választani.

Általánosan az **alaphalmaz** egy tetszőleges $\mathcal{X}$ halmaz, amelynek elemeit címkézni, avagy kategorizálni szeretnénk. Ezen elemeket általában bizonyos (általunk választott) tulajdonságokat mérő vektorokkal reprezentáljuk. A fenti példánkban az alaphalmaz a mangó színét és keménységét megadó két dimenziós vektorokból áll.

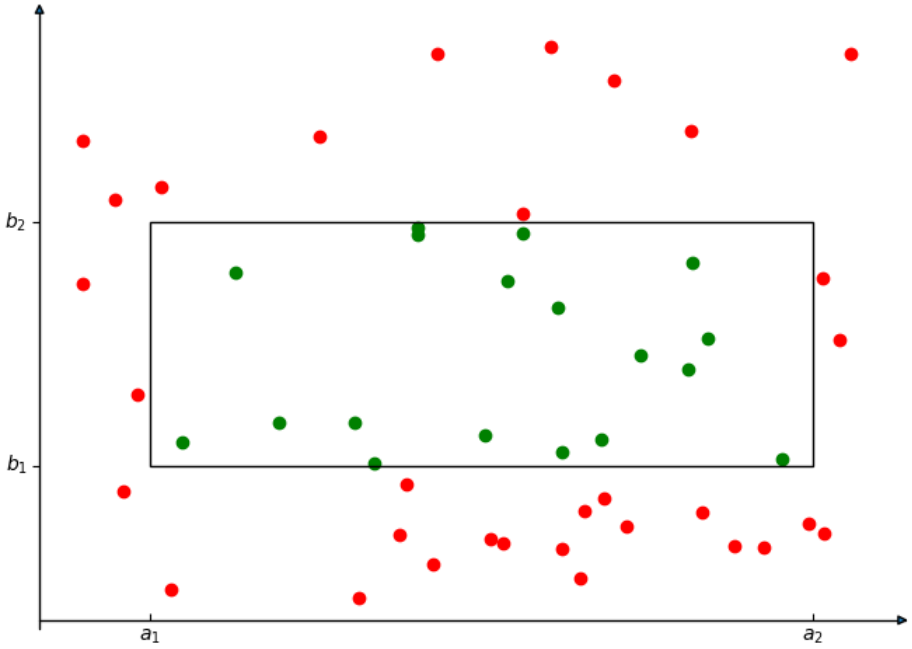
A statisztikai tanulás alapfeltevése szerint a rendelkezésünkre álló adathalmaz egy **független, azonos eloszlású minta**, melynek elemei az **általunk ismeretlen** $\mathcal{D}$ eloszlás szerint kerülnek a tanulóhalmazba. Ezt úgy képzelhetjük el, hogy nincs jelen se segítő, se ártó szándékú tanár, aki az algoritmus számára olyan tanulóhalmazt állít össze, amely a lehető legtöbb, illetve legkevesebb információt tartalmazza az ismeretlen eloszlásról, amivel jelentősen megkönnyítené, illetve megnehezítené a

tanulás folyamatát. Magyarul az adatpontokat véletlenszerűen kapjuk a $\mathcal{D}$ eloszlásból. Ezzel a feltevéssel tehát utat nyitunk a statisztikai okfejtéseknek.

Mivel klasszifikációs feladatról beszélünk, ezért szükséges megállapítanunk a **lehetséges címkék halmazát**. Ezt jelöljük $\mathcal{Y}$-nal. Példánkban a célváltozónak két lehetséges értéke van, ilyenkor rendszerint a következő megválasztással dolgozunk: $\mathcal{Y} = \{0,1\}$. Ez a legegyszerűbb eset, a későbbiek során is ezzel foglalkozunk, kivéve a **2.3** és **2.4** alfejezetekben. Itt 0 jelöli azt az esetet, amikor a mangó nem finom, 1 pedig azt, amikor finom. Az $S = \big((x_1, y_1), ..., (x_m, y_m)\big)$ véges, $\mathcal{X} \times \mathcal{Y}$-beli párokból álló sorozatot nevezzük **tanulóhalmaz**nak. Ezek azok a címkékkel ellátott pontok, amikhez a tanulóalgoritmus hozzáfér. Az A algorimtus bemenete tehát az S halmaz, kimenete pedig egy $h : \mathcal{X} \to \mathcal{Y}$ hipotézis (címkézőfüggvény), vagyis $A(S) = h$. A példa szerint ez egy olyan függvény, ami minden mangóról ránézésre meg tudja mondani, hogy finom-e. Ugyan a szigeten minden mangó vagy finom vagy nem, vagyis az osztályozás diszjunkt, ám arról nincs információnk, hogy az általunk vizsgált két jellemző alapján ez egyértelműen meghatározható-e, vagyis hogy létezik-e egy „helyes" $f : \mathcal{X} \to \mathcal{Y}$ címkézőfüggvény. Egyelőre az egyszerűség kedvéért tegyük fel, hogy létezik ilyen f. Ez az a függvény, amit szeretnénk minél jobban megközelíteni az algoritmusunk által talált hipotézissel. Fontos megemlíteni, hogy ez hogyan kapcsolódik az előzetes ismereteinkhez. Azt feltételezzük ugyanis, hogy az általunk megválasztott tulajdonságok egyértelműen meghatározzák a gyümölcs finomságát, tehát már az alaphalmaz reprezentálásába is beépítettük az előzetes tudásunkat. Másrészt beépíthetjük a tudásunkat a - továbbiakban központi szerepet játszó - hipotézisosztály meghatározásával is. A $\mathcal{H}$ **hipotézisosztály** azon függvények halmaza, amelyek között az algoritmusunk keres. Ha a mangó keménységét (x_1) és színét (x_2) egy derékszögű koordinátarendszerben ábrázoljuk, akkor választhatjuk például a $\mathcal{H}$-t a következő alakú függvények halmazának:

$$h_T(x_1, x_2) = \begin{cases} 1 & \text{ha} \quad (x_1, x_2) \in T \\ 0 & \text{ha} \quad (x_1, x_2) \notin T \end{cases}$$

Ekkor minden h_T hipotézist négy paraméter határoz meg, a tengelyekkel párhuzamos oldalú T téglalap oldalainak koordinátái (a 2.1 ábrán ezeket a_1, a_2, b_1, b_2 jelöli). h_T a téglalap belsejébe eső adatpontokhoz az 1 címkét rendeli (zöld pontok), a többihez a 0-t (piros pontok).



2.1. ábra. *Egy lehetséges h_T hipotézis*

2.1.1. A hiba mérése

Az algoritmus teljesítményét mérni szeretnénk, ezt az általa felállított hipotézisen keresztül tudjuk megtenni. A következő definícióban megadjuk a hipotézis valós hibáját, amit végül minimalizálni szeretnénk.

2.1. Definíció. *A $h: \mathcal{X} \to \mathcal{Y}$ hipotézis valós hibája a $\mathcal{D}$ eloszlás és f címkézőfüggvény mellett:*

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

A valós hiba tehát annak a valószínűsége, hogy a $\mathcal{D}$ eloszlásból véletlenszerűen mintavételezett x adatpont valós címkéjét a hipotézis nem találja el, azaz rossz osztályba sorolja. Ezt a mennyiséget közvetlenül nem tudjuk meghatározni, ugyanis nem ismerjük sem a $\mathcal{D}$ eloszlást, sem az f címkézőfüggvényt. Természetes gondolat, hogy a hibát a következőképp mérjük:

2.2. Definíció. *A $h : \mathcal{X} \to \mathcal{Y}$ hipotézis tapasztalati hibája az $S = \big((x_1, y_1), ..., (x_m, y_m)\big) \subset (\mathcal{X} \times \mathcal{Y})^m$ tanulóhalmazon:*

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

Figyeljük meg, hogy a tapasztalati hiba adott hipotézisre csak az S tanulóhalmaztól függ, és értékét az adja meg, hogy a hipotézis az S-beli címkézett adatpontoknak hányadrészén találta el a helyes értéket. Mivel az S halmaz tartalmazza az

összes információt, amit a $\mathcal{D}$ eloszlásról tudunk, ezért ésszerűnek tűnik olyan hipotézist választani, ami S elemeivel a leginkább kontisztens, tehát minimalizálja a tapasztalati hibát. Azt a tanulási paradigmát, ami olyan h hipotézist keres a $\mathcal{H}$-ból, amire $L_S(h)$ értéke minimális, **ERM-szabály**nak nevezzük (*Emipirical Risk Minimization*). Az ilyen hipotézisek halmazát $\mathrm{ERM}(S)$-sel (vagy $\mathrm{ERM}_{\mathcal{H}}(S)$-sel) fogjuk jelölni, azaz

$$\mathrm{ERM}_{\mathcal{H}}(S) = \arg\min_{h \in \mathcal{H}} L_S(h).$$

Az algoritmus által kiválasztott függvényt pedig h_S-el jelöljük, utalva az adathalmaztól való függésre.

Fontos itt megjegyezni, hogy nem minden $\mathcal{H}$ esetén határozható meg polinomidőben az a hipotézis, amelyre $L_S(h)$ minimális. Ez sok esetben komoly megvalósíthatósági korlátot jelent, és azt eredményezi, hogy vagy a $\mathcal{H}$-t szorítjuk meg olyan módon, hogy az arra alkalmazott feladat már polinomidőben megoldható legyen, vagy pedig megelégszünk egy olyan algoritmussal, ami a minimumot keresi, de nem feltétlenül találja meg. Megvalósíthatósági kérdéssekkel itt nem fogunk foglalkozni.

Az ERM egyik hátulütője lehet, hogy annyira szerencsétlen módon mintavételezünk, hogy egy teljesen hibás képünk lesz a $\mathcal{D}$ eloszlásról, és a tanulóalgoritmus olyan hipotézist talál, amelynek tapasztalati hibája kicsi, ám tényleges hibája elfogadhatatlanul nagy lesz. Tegyük fel, hogy kezdetben a szigeten csupa olyan mangót kóstolunk, amelynek rossz az íze. Ekkor arra a h_S hipotézisre, amely minden x_i ponthoz a 0 értéket rendeli, $L_S(h_S) = 0$. Mivel a tapasztalati hiba ennél kisebb nem lehet, így $h_S \in \mathrm{ERM}(S)$. Ám ha a $\mathcal{D}$ eloszlás olyan, hogy minden mangó azonos valószínűséggel kerül a tanulóhalmazba, és azoknak pontosan a fele rossz ízű, akkor $L_{\mathcal{D},f}(h_S) = 1/2$.
Előfordulhat olyan eset is, amikor (legtöbbször a túl hosszú tanulási folyamat következtében) a hipotézis az S halmaz pontjainak egyedi jellemzőit „jegyzi meg", így az $\mathcal{X}$ halmazon való általánosításra nem alkalmas. Ezt a jelenséget **túltanulás**nak nevezzük. A túltanulás elkerülése a gépi tanulás egyik központi feladata.

Az ERM paradigmán kívül vannak más elterjedt tanulási szabályok is. Például az MDL (Minimum Description Length) paradigma, ami lényegében az Occam borotvája-elv egyik lehetséges formalizálása. Az MDL-algoritmusok a tapasztalati hiba mellett figyelembe veszik a hipotézis reprezentálásának (egy fixált nyelv által meghatározott) nagyságát is, és az egyszerűbb hipotéziseket preferálják. A továbbiakban ERM elven működő algoritmusokkal foglalkozunk, amelyeket röviden **ERM-algoritmusok**nak nevezünk.

2.2. Alapmodell

Az előzőekben feltettük hogy létezik egy „helyes” címkézőfüggvény, arról azonban nincs információnk, hogy ez benne van-e az általunk definiált $\mathcal{H}$ hipotézisosztályban. A következőkben sokszor használjuk majd alábbi feltevést, amit már az előzőekben is megfogalmaztunk.

2.3. Definíció. *Megvalósíthatósági feltevés*
Létezik olyan $h^* \in \mathcal{H}$ hipotézis, amelyre $L_{\mathcal{D},f}(h^*) = 0$.

Figyeljük meg, hogy ekkor egy valószínűséggel $L_S(h^*) = 0$. Hiszen $L_{\mathcal{D},f}(h^*) = = 0$ pontosan azt jelenti, hogy a $\mathcal{D}$ eloszlás szerint mintavételezve egy pontot, h^* egy valószínűséggel jól címkézi meg a pontot, tehát ha mintavételezünk véges sok pontot, akkor ez ugyanúgy áll az összes adatpontra egyszerre. Látható, hogy egy ERM algoritmus által választott h_S-re $L_S(h_S) = 0$, azonban azt nem tudjuk, hogy ez a h_S megegyezik-e a megvalósíthatósági feltevés szerint létező h^* hipotézissel. Számunkra az a fontos, hogy minél kisebb *valós hibájú* hipotézist találjunk, így az $L_{\mathcal{D},f}(h_S)$ értékről szeretnénk valamit mondani.

Az S tanuló adathalmaz generálását (és ezáltal a h_S hipotézis kiválasztását is) egy véletlen folyamat vezérli, ebből következik, hogy a hipotézis $L_{\mathcal{D},f}(h_S)$ valós hibája egy véletlen szám, vagyis *valószínűségi változó*. Nem várhatjuk el, hogy csupán az S tanulóhalmaz ismeretében következtetni tudjunk a valós hibára, hiszen előfordulhat, hogy S elemeinek eloszlása nem reprezentálja jól az ismeretlen $\mathcal{D}$ eloszlást. Az ilyen szerencsétlen esetek miatt bevezetünk egy δ paramétert. Ekkor $1-\delta$ lesz annak a valószínűsége, hogy reprezentatív mintánk van, ezt *konfidencia paraméter*nek nevezzük. Hogy ez matematikailag pontosan mit jelent, a következőkben válik világossá. Egy algoritmustól ugyanakkor az sem várható el, hogy megtalálja a megvalósíthatósági feltevés szerint létező „tökéletes” hipotézist, amelynek valós hibája 0. Rözgítsünk egy $\varepsilon \geq 0$ *pontossági paramétert* is, ami egyfajta teljesítményi korlátként szolgál. Azt várjuk el az algoritmus által visszaadott h_S hipotézistől, hogy valós hibája ε-nál kisebb legyen, ezáltal egy közelítőleg helyes megoldást kapjunk. Az általános definíció előtt nézzük meg, hogy mit mondhatunk egy véges sok elemből álló hipotézisosztályról, ha az algoritmusunk ERM-algoritmus.

2.4. Állítás. *Legyen $\mathcal{H}$ egy véges hipotézisosztály, $\delta \in (0,1), \varepsilon > 0$ tetszőlegesen adott számok. Legyen $m \in \mathbb{Z}$ olyan szám, amelyre teljesül, hogy*

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

Ekkor tetszőleges $\mathcal{D}$ eloszlás mellett, és olyan f címkézőfüggvény esetén, ahol $\mathcal{H}$ teljesíti a megvalósíthatósági feltevést, ha az m elemszámú S halmaz elemeit egymástól függetlenül mintavételezzük a $\mathcal{D}$ eloszlásból, akkor tetszőleges $h_S \in \mathrm{ERM}_{\mathcal{H}}(S)$ hipotézisre teljesül, hogy

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

Bizonyítás. Jelölje $S|_x{=}(x_1, ...x_m)$ a tanulóhalmazban szereplő adatpontokat. Annak a valószínűsége, hogy a tanulás sikertelen lesz

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

Célunk az, hogy belássuk, hogy ez a valószínűség legfeljebb δ lehet. Jelölje $\mathcal{H}_R$ a „rossz" hipotézisek halmazát, amelyeknek valós hibája ε-nál nagyobb:

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

Továbbá legyen a félrevezető minták halmaza

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

Ez azt jelenti, hogy egy M-beli $S|_x$ mintához létezik olyan $h{\in}\mathcal{H}_R$ hipotézis, amely a tanulóhalmazon hiba nélkül teljesít, ezáltal egy $\mathrm{ERM}_{\mathcal{H}}$ algoritmus által kiválasztásra kerülhet, valós hibája azonban ε-nál nagyobb.

Mivel a megvalósíthatósági feltevés fennáll, ezért $L_S(h_S){=}0$, vagyis $L_{\mathcal{D},f}(h_S){>}\varepsilon$ csak akkor következhet be, ha $L_S(h) = 0$ valamely $h \in \mathcal{H}_R$ hipotézisre, vagyis ha a minta félrevezető. Ezáltal

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

Ezek alapján M felírható a következő alakban:

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

Az eddigieket felhasználva

$$\begin{aligned}\mathcal{D}^m\big(\{S|_x : L_{\mathcal{D},f}(h_S) > \varepsilon\}\big) &\le \mathcal{D}^m(M)\\ &= \mathcal{D}^m\big(\cup_{h\in\mathcal{H}_R}\{S|_x : L_S(h) = 0\}\big)\\ &\le \sum_{h\in\mathcal{H}_R} \mathcal{D}^m\big(\{S|_x : L_S(h) = 0\}\big). \qquad (2.1)\end{aligned}$$

Rögzített $h \in \mathcal{H}_R$ hipotézis esetén az összeg tagjait felülről tudjuk becsülni, felhasználva azt, hogy a minta elemei egymástól függetlenek, illetve az $1-\varepsilon \le e^{-\varepsilon}$ azonosságot. Tehát

$$\begin{aligned}\mathcal{D}^m\big(\{S|_x : L_S(h) = 0\}\big) &= \mathcal{D}^m\big(\{S|_x : \forall i, h(x_i) = f(x_i)\}\big)\\ &= \prod_{i=1}^{m} \mathcal{D}\big(\{x_i : h(x_i) = f(x_i)\}\big)\\ &= \prod_{i=1}^{m}(1 - L_{\mathcal{D},f}(h)) \le \prod_{i=1}^{m}(1-\varepsilon) = (1-\varepsilon)^m \le e^{-\varepsilon m}. \qquad (2.2)\end{aligned}$$

A 2.1 és 2.2 egyenlőtlenségeket felhasználva

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

A feltételnek megfelelő m választásával megkapjuk, hogy a keresett valószínűség értéke legfeljebb δ lehet. □

Tehát véges hipotézisosztály és meghatározott méretű tanulóhalmaz esetén az $\mathrm{ERM}_{\mathcal{H}}$ elven működő algoritmus **valószínűleg** ($1-\delta$ valószínűséggel) **közelítőleg helyes** (legfeljebb ε valós hibával) megoldást eredményez, függetlenül az ismeretlen $\mathcal{D}$ eloszlástól és f címkézőfüggvénytől.

Így adódik a valószínűleg közelítőleg helyes tanulási modell, amelynek formális definíciója a következő:

2.5. Definíció. *PAC-tanulhatóság*
A $\mathcal{H}$ hipotézisosztály PAC-megtanulható, ha létezik egy $m_{\mathcal{H}}^{PAC} : (0,1)^2 \to \mathbb{N}$ függvény és A tanulóalgoritmus, amelyre teljesülnek a következők:
Tetszőleges $\varepsilon, \delta \in (0,1)$ valós számok, $\mathcal{X}$ feletti $\mathcal{D}$ eloszlás, és $f : \mathcal{X} \to \{0,1\}$ címkézőfüggvény esetén, ha az $m \geq m_{\mathcal{H}}^{PAC}(\varepsilon, \delta)$ számosságú S halmaz elemeit egymástól függetlenül mintavételezzük a $\mathcal{D}$ eloszlásból és f szerint címkézzük, illetve a megvalósíthatósági feltevés fennáll, akkor az $A(S) = h \in \mathcal{H}$ hipotézisre

$$\mathop{\mathbb{P}}_{S\sim\mathcal{D}^m}\big(L_{\mathcal{D},f}(h) \leq \varepsilon\big) \geq 1-\delta.$$

Vagyis általunk választott tetszőleges ε pontossági és $1-\delta$ konfidencia paraméterek esetén létezik olyan m mintaelemszám, hogy az m elemű S tanulóhalmaz ismeretében az A algoritmus olyan h hipotézist választ a $\mathcal{H}$ hipotézisosztályból, amelynek valós hibája $1-\delta$ valószínűséggel kisebb lesz, mint ε.
Ezáltal rögzíteni tudjuk, hogy mekkora bizonyossággal, és milyen pontos megoldást szeretnénk kapni. Fontos kiemelni, hogy egy hipotézisosztály PAC-megtanulhatósága **nem függ az ismeretlen eloszlástól** olyan értelemben, ahol az eloszlásra és a címkézőfüggvényre fennáll a megvalósíthatósági feltevés.

Az $m_{\mathcal{H}}^{PAC}$ függvény a szükséges minta elemszámát határozza meg, hogy az elvárásoknak megfelelő hipotézist tudjunk találni. Ha $\mathcal{H}$ PAC-megtanulható, akkor több megfelelő $m_{\mathcal{H}}^{PAC}$ függvény is létezhet, így megállapodás szerint adott ε és δ esetén $m_{\mathcal{H}}^{PAC}(\varepsilon, \delta)$ a legkisebb olyan egész számot jelöli, amely teljesíti a definícióban megadott feltételeket. A 2.4 állítást átírva kapjuk az alábbi következményt.

2.6. Következmény. *Minden véges $\mathcal{H}$ hipotézisosztály PAC-megtanulható, ahol*

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

2.3. Agnosztikus PAC-tanulhatóság

A 2.5 definíció komoly megszorításokkal él:

- a gépi tanulási feladatoknak mindössze egy osztályára, a bináris klasszifikációs feladatokra vonatkozik (hiszen a címkék lehetséges értékeinek halmaza {0,1})
- a megvalósíthatósági feltevés teljesülését általában lehetetlen ellenőrizni
- nem mindig realisztikus az a feltevés, hogy a tulajdonságok amiket mérünk, teljesen meghatározzák az adott objektum címkéjét. Ugye előfordulhat, hogy két, külsőre teljesen ugyanolyan mangó közül az egyiknek rohadt a belseje.

Ezen okok miatt közlünk egy általánosabb modellt is.

A továbbiakban elhagyjuk a megvalósíthatósági feltevést, azaz nem feltétlenül létezik $h^* \in \mathcal{H}$, hogy $L_{\mathcal{D},f}(h^*) = 0$. Ekkor azonban irreális elvárás lenne, hogy az algoritmus által talált h hipotézis valós hibája tetszőlegesen kicsi legyen, hiszen előfordulhat például, hogy mindegyik $\mathcal{H}$-beli hipotézis valós hibája egy konstans számnál nagyobb. Ezért olyan hipotézist fogunk elfogadni közelítőleg helyesnek, amelynek valós hibája a hipotézisosztályon belül elérhető minimális hibánál legfeljebb ε-nal nagyobb, vagyis teljesül rá, hogy

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

Az S tanulóhalmaz elemeire eddig úgy tekintettünk, mint olyan (x_i, y_i) elemek, amelyekre teljesül, hogy $x_i \sim \mathcal{D}$ és $f(x_i) = y_i$. Vagyis az ismeretlen $\mathcal{D}$ eloszlást követő x (ami általában jellemzők egy vektora) és a hozzá tartozó osztálycímke által meghatározott rendezett párok. Ezentúl a $\mathcal{D}$ eloszlást az egész $Z = \mathcal{X} \times \mathcal{Y}$ halmazon értelmezzük. Ez egy együttes eloszlás az adatpontok és a lehetséges címkék halmaza fölött. A feltételes eloszlás definíciója miatt

$$\mathbb{P}(x = x_i, y = y_i) = \mathbb{P}(x = x_i)\mathbb{P}(y = y_i | x = x_i),$$

tehát érdemes úgy gondolni az eloszlásra, hogy egy marginális eloszlás szerint megkapjuk az adatpontot, és utána egy feltételes valószínűség mutatja azt, hogy az adott jellemzők mellett mekkora valószínűséggel tartozik az adott osztályba. Vegyük észre, hogy ilyenkor nem feltétlenül létezik olyan $\mathcal{X} \to \mathcal{Y}$ függvény, amely tökéletesen leírja az adathalmazt. Ha létezik is, akkor azt innentől beépítettük a $\mathcal{D}$ eloszlásba (ekkor a fenti feltételes valószínűségi tag értéke egy osztályra 1 lesz, a többire 0). Az alaphalmaz ilyen módú definiálásával könnyen értelmezhetővé válik az az eset is, amikor $\mathcal{Y}$ nem véges halmaz, hanem például $\mathbb{R}$. Ebben az esetben **regressziós feladat**ról beszélünk, és a hibát is máshogyan kell mérni mint eddig.

2.3.1. Veszteségfüggvények

2.7. Definíció. *Veszteségfüggvény*
Adott $\mathcal{H}$ hipotézisosztály és Z alaphalmaz esetén az $\ell : \mathcal{H} \times Z \to \mathbb{R}_+^0$ függvényt veszteségfüggvénynek nevezzük.

Rögzített h hipotézis esetén tehát a Z alaphalmaz minden pontjához hozzárendelünk egy veszteséget, ami attól függ, hogy h milyen teljesítményt nyújt az adott pontra vonatkozóan. Ezt úgy értelmezhetjük, hogy a hipotézist „büntetjük" olyan esetekben, amikor tévesen határozza meg egy-egy adatpont osztálycímkéjét (regressziós feladat esetén a célváltozó valódi és becsült értékének eltérésének mértékét vizsgáljuk). Így általánosíthatjuk egy hipotézis valós és tapasztalati hibáját.

2.8. Definíció. *Valós hiba*
A $h \in \mathcal{H}$ hipotézis valós hibája adott Z feletti $\mathcal{D}$ eloszlás esetén:

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

2.9. Definíció. *Tapasztalati hiba*
A $h \in \mathcal{H}$ hipotézis tapasztalati hibája az $S = (z_1, ..., z_m) \in Z^m$ tanulóhalmazon:

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

Vagyis a h hipotézis valós hibája definíció szerint a veszteségfüggvény várható értéke a z adatponton, amelyet a $\mathcal{D}$ eloszlásból véletlenszerűen választunk. A hipotézis tapasztalati hibája pedig h átlagos vesztesége a tanulóhalmaz pontjaira nézve.

A felügyelt tanulásos feladatokban leggyakrabban a következő két veszteségfüggvényt használjuk (mindkét esetben $Z = \mathcal{X} \times \mathcal{Y}$):

0–1:. Ezt a veszteségfüggvényt klasszifikációs feladatokra használjuk, vagyis diszkrét értékű $\mathcal{Y}$ esetén

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

Vagyis a büntetés 1, ha a hipotézis nem találja el az adatpont címkéjét, és 0, ha eltalálja.

négyzetes:. A regressziós feladatok esetében, a büntetésfüggvény értéke a hipotézis által jósolt érték és a valódi érték különbségének négyzete, ezáltal egységnyi eltérést a valódi értéktől távol jobban büntetünk, mint hozzá közel.

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

Mindent összevetve már definiálhatjuk a PAC-tanulhatóságot általános esetben.

2.10. Definíció. *Agnosztikus PAC-tanulhatóság*
A $\mathcal{H}$ hipotézisosztály adott Z alaphalmaz és $\ell : \mathcal{H} \times Z \to \mathbb{R}_+^0$ veszteségfüggvény esetén agnosztikusan PAC-megtanulható, ha létezik olyan $m_{\mathcal{H}}^{AgPAC} : (0,1)^2 \to \mathbb{N}$ függvény, és A tanulóalgoritmus, amelyre teljesülnek a következők:
Tetszőleges $\varepsilon, \delta \in (0,1)$ valós számok és Z feletti $\mathcal{D}$ eloszlás esetén, ha az $m \geq$ $\geq m_{\mathcal{H}}^{AgPAC}(\varepsilon, \delta)$ elemű S halmaz elemeit egymástól függetlenül mintavételezzük a $\mathcal{D}$ eloszlásból, akkor az $A(S) = h \in \mathcal{H}$ hipotézisre teljesül

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

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

2.4. Egyenletes konvergencia

A következőkben a $\mathcal{H}$ hipotézisosztálynak egy olyan tulajdonságát fogjuk vizsgálni, amely teljesülése esetén az osztály agnosztikusan PAC-megtanulható.

Az ERM elven működő algoritmusok azt a hipotézist választják, amelyiknek a tanulóhalmazon elért hibája minimális. Mi azonban továbbra is csak egy *ablakon* (az S tanulóhalmazon) keresztül pillanthatunk rá a *világra* (az ismeretlen $\mathcal{D}$ eloszlásra), és csak ez alapján tudunk döntést hozni a hipotézisekről. A kérdés, hogy az ERM-elvet alkalmazva kapott h_S hipotézis *általánosan* is kellően jól teljesít-e. Gondoljuk meg, hogy a kapott hipotézisünk biztosan megfelel az elvárásoknak, ha garantálni tudjuk, hogy a tapasztalati hiba minden $h \in \mathcal{H}$ hipotézis esetén közel van a valós hibához. Ezt a következőképpen formalizáljuk:

2.11. Definíció. *ε-reprezentatív minta*
Az S tanulóhalmaz ε-reprezentatív rögzített Z alaphalmaz, $\mathcal{H}$ hipotézisosztály, ℓ veszteségfüggvény és $\mathcal{D}$ eloszlás mellett, ha

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

A következő lemma azt mondja ki, hogy $(\varepsilon/2)$-reprezentatív minta esetén a tapasztalati hibát minimalizáló ERM tanulási szabályt alkalmazva garantáltan ε-közel kerülünk a $\mathcal{H}$ hipotézisosztályon elérhető legkisebb valós hibához. Ezt úgy is megfogalmazhatjuk, hogy bármilyen ERM-algoritmus sikeres tanuló.

2.12. Lemma. *Legyen S egy $\frac{\varepsilon}{2}$-reprezentatív minta. Ekkor minden* ERM*-szabállyal kapott $h_S \in \arg\min_{h \in \mathcal{H}} L_S(h)$ hipotézisre:*

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

Bizonyítás. Minden $h \in \mathcal{H}$ esetén:

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

ahol az első és a harmadik egyenlőtlenség abból következik, hogy S $\frac{\varepsilon}{2}$-reprezentatív, a második pedig azért igaz, mert a h_S hipotézist egy ERM-algoritmus eredményeként kaptuk, ami miatt a tapasztalati hibájánál semelyik $\mathcal{H}$-beli hipotézis tapasztalati hibája sem lehet kisebb. □

A véletlen mintavételezési folyamat miatt azt nyilván nem tudjuk garantálni, hogy S egy valószínűséggel ε-reprezentatív legyen, de bizonyos esetekben tetszőleges $\delta \in (0,1)$ paraméterhez meghatározható olyan mintaelemszám, hogy $1-\delta$ valószínűséggel ε-reprezentatív mintához jussunk. Ilyenkor azt mondjuk, hogy a hipotézisosztályunk egyenletesen konvergens.

2.13. Definíció. *Egyenletes konvergencia*
A $\mathcal{H}$ hipotézisosztály adott Z alaphalmaz és $\ell : \mathcal{H} \times Z \to \mathbb{R}_+^0$ veszteségfüggvény esetén egyenletesen konvergens, ha létezik olyan $m_{\mathcal{H}}^{UC} : (0,1)^2 \to \mathbb{N}$ függvény, hogy tetszőleges $\varepsilon, \delta \in (0,1)$ és Z feletti $\mathcal{D}$ eloszlás esetén, ha az $m \geq m_{\mathcal{H}}^{UC}(\varepsilon, \delta)$ elemszámú S halmaz elemeit egymástól függetlenül mintavételezzük a $\mathcal{D}$ eloszlásból, akkor S legalább $1-\delta$ valószínűséggel ε-reprezentatív.

A definícióban szereplő $m_{\mathcal{H}}^{UC}$ függvény, hasonlóan, mint a PAC-tanulhatóság esetében, azt határozza meg, hogy rögzített ε és δ paraméterek mellett mekkora az a minimális mintaelemszám, hogy legalább $1-\delta$ valószínűséggel ε-reprezentatív mintánk legyen. Az egyenletes konvergencia tulajdonság jelentőségét is az adja, hogy független az ismeretlen $\mathcal{D}$ eloszlástól.

Figyeljük meg, hogy az előbbi lemma és az egyenletes konvergencia definíciója (ha ε helyett $\frac{\varepsilon}{2}$-re alkalmazzuk) közvetlenül azt implikálja, hogy az egyenletes konvergencia elégséges feltétele az agnosztikus PAC-megtanulhatóságnak.

2.14. Következmény. *Ha a $\mathcal{H}$ hipotézisosztály egyenletesen konvergens az $m_{\mathcal{H}}^{UC}$ függvény mellett, akkor a hipotézisosztály agnosztikusan PAC-megtanulható, és $m_{\mathcal{H}}^{PAC}(\varepsilon, \delta) \leq m_{\mathcal{H}}^{UC}(\varepsilon/2, \delta)$. Továbbá bármilyen* ERM-*algoritmus sikeres agnosztikus PAC-tanuló.*

3. fejezet

Nincs ingyen ebéd tétel

Az eddigiekben láttuk, hogy véges hipotézisosztályok minden esetben PAC-megtanulhatóak. A következő tétel fontos következménye, hogy az összes függvényt tartalmazó hipotézisosztály nem PAC-megtanulható.

3.1. Tétel. *Nincs ingyen ebéd*
Legyen A tetszőleges tanulóalgoritmus bináris klasszifikációs feladatra adott $\mathcal{X}$ alaphalmaz és ℓ_{0-1} veszteségfüggvény mellett. Jelölje $m < \frac{|\mathcal{X}|}{2}$ a tanulóhalmaz méretét. Ekkor létezik olyan $\mathcal{X} \times \{0,1\}$ feletti $\mathcal{D}$ eloszlás, amelyre:

1. *Létezik $f : \mathcal{X} \to \{0,1\}$ függvény, hogy $L_{\mathcal{D}}(f) = 0$.*
2. *Legalább $\frac{1}{7}$ valószínűséggel az $S \sim \mathcal{D}^m$ tanulóhalmazon $L_{\mathcal{D}}\big(A(S)\big) \geq \frac{1}{8}$.*

A tétel állítása szerint tehát minden tanulóalgoritmushoz található olyan feladat, amin elbukik, miközben ugyanerre a feladatra létezik sikeres tanuló. Ez utóbbira triviális példa az az ERM algoritmus, ahol $\mathcal{H} = \{f\}$.

Bizonyítás.
Tekintsünk egy $2m$ elemszámú $C \subset \mathcal{X}$ halmazt, és legyen $\mathcal{F} = \{f_1, ...f_T\}$ az összes $C \to \{0,1\}$ függvény halmaza. Ekkor nyilván $|\mathcal{F}| = 2^{2m}$. Minden f_i függvényhez konstruáljunk egy $C \times \{0,1\}$ feletti $\mathcal{D}_i$ eloszlást a következő módon:

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

Ekkor nyilván $L_{\mathcal{D}_i}(f_i)$=0. Belátjuk, hogy ekkor minden A tanulóalgoritmusra fennáll, hogy

$$\max_{i \in [T]} \mathop{\mathbb{E}}_{S \sim \mathcal{D}_i} \Big(L_{\mathcal{D}_i}\big(A(S)\big) \Big) \geq \frac{1}{4}. \tag{3.1}$$

Végül megmutatjuk, hogy ebből már következik a tétel.

A bizonyítást tehát 3.1 igazolásával kezdjük. C elemeiből összesen $k = (2m)^m$ sorozatot készíthetünk. Jelölje ezeket $S_1, ..., S_k$. Minden S_j=$(x_1, ..., x_m)$ esetén jelölje S_j^i azt a sorozatot, amely az $(x_l, f_i(x_l))$ (l=1, . . . , m) címkézett elemeket tartalmazza. Így tehát, ha az eloszlás $\mathcal{D}_i$, akkor A az $S_1^i, ..., S_k^i$ tanulóhalmazok egyikét kaphatja bemenetként. A független mintavételezésre vonatkozó feltevés alapján ezek közül az összeset azonos valószínűséggel kaphatjuk meg.

$$\begin{aligned}
\max_{i\in[T]} \mathop{\mathbb{E}}_{S\sim\mathcal{D}_i}\Big(L_{\mathcal{D}_i}\big(A(S)\big)\Big) &= \max_{i\in[T]} \frac{1}{k}\sum_{j=1}^{k} L_{\mathcal{D}_i}\big(A(S_j^i)\big) \\
&\geq \frac{1}{T}\sum_{i=1}^{T}\frac{1}{k}\sum_{j=1}^{k} L_{\mathcal{D}_i}\big(A(S_j^i)\big) \\
&= \frac{1}{k}\sum_{j=1}^{k}\frac{1}{T}\sum_{i=1}^{T} L_{\mathcal{D}_i}\big(A(S_j^i)\big) \\
&\geq \min_{k\in[j]} \frac{1}{T}\sum_{i=1}^{T} L_{\mathcal{D}_i}\big(A(S_j^i)\big). \qquad (3.2)
\end{aligned}$$

Az utolsó egyenlőtlenség jobb oldalán álló kifejezésre fogunk mondani egy olyan alsó korlátot, amely minden k-ra fennáll, így a minimálisra is. Rögzítsünk tehát egy tetszőleges $k \in [j]$ számot. Legyen $S_j = (x_1, ..., x_m)$ és legyenek $\{v_1, ...v_p\}$ azok a C-beli elemek, amik nem jelennek meg az S_j-ben. Ekkor nyilván $p \geq m$, és minden $h : C \to \{0,1\}$ függvényre fennáll a következő:

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

ahol $\mathbb{1}_{[B]}$ a B esemény indikátorát jelöli, aminek értéke pontosan akkor 1, ha B fennáll, különben 0. Mivel ez minden h-ra, így $A(S_j^i)$-re is teljesül. Ezért

$$\begin{aligned}
\frac{1}{T}\sum_{i=1}^{T} L_{\mathcal{D}_i}\big(A(S_j^i)\big) &\geq \frac{1}{T}\sum_{i=1}^{T}\frac{1}{2p}\sum_{r=1}^{p} \mathbb{1}_{[h(v_r)\neq f_i(v_r)]} \\
&= \frac{1}{2p}\sum_{r=1}^{p}\frac{1}{T}\sum_{i=1}^{T} \mathbb{1}_{[h(v_r)\neq f_i(v_r)]} \\
&\geq \frac{1}{2}\min_{r\in[p]} \frac{1}{T}\sum_{i=1}^{T} \mathbb{1}_{[h(v_r)\neq f_i(v_r)]}. \qquad (3.3)
\end{aligned}$$

Most rögzítsünk egy $r \in [p]$ számot. Képezhetünk az $f_1, ..., f_T$ függvényekből egyértelműen diszjunkt párokat olyan módon, hogy minden $(f_i, f_{i'})$ párra teljesül, hogy

$f_i(c) \neq f_{i'}(c)$ pontosan akkor, ha $c = v_r$. Ebből kifolyólag $S_j^i = S_j^{i'}$, és $A(S_j^i)(v_r) = A(S_j^{i'})(v_r)$, ami vagy az $f_i(v_r)$ vagy pedig az $f_{i'}(v_r)$ értékkel lesz egyenlő. Tehát

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

Ez mind a $\frac{T}{2}$ párra teljesül, ezért

$$\frac{1}{T} \sum_{i=1}^{T} \mathbb{1}_{[A(S_j^i)(v_r) \neq f_i(v_r)]} = \frac{1}{2}. \tag{3.4}$$

Az 3.2, 3.3 és 3.4 eredményeket felhasználva:

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

Ezzel tehát a 3.1 azonosságot beláttuk. Ez nyilvánvalóan azt implikálja, hogy minden A tanulóalgoritmusoz létezik olyan $f : \mathcal{X} \to \{0,1\}$ címkézőfüggvény és $\mathcal{X} \times \{0,1\}$ feletti $\mathcal{D}$ eloszlás, hogy $L_{\mathcal{D}}(f) = 0$, és

$$\mathop{\mathbb{E}}_{S \sim \mathcal{D}^m} \Big(L_{\mathcal{D}}\big(A(S)\big) \Big) \geq \frac{1}{4}. \tag{3.5}$$

Innen pedig egyszerűen következik a tétel állítása. Az egyszerűség kedvéért jelöljük a [0,1]-be képező $L_{\mathcal{D}}\big(A(S)\big)$ valószínűségi változót θ-val, és legyen $\mathbb{E}(\theta) = \mu$. Ekkor $Z = 1 - \theta$ nemnegatív valószínűségi változóra alkalmazva a Markov-egyenlőtlenséget (5.3) kapjuk, hogy

$$\mathbb{P}\left(\theta < \frac{1}{8}\right) = \mathbb{P}\left(1 - \theta > \frac{7}{8}\right) = \mathbb{P}\left(Z > \frac{7}{8}\right) \leq \frac{\mathbb{E}(Z)}{\frac{7}{8}} = \frac{1-\mu}{\frac{7}{8}}.$$

Ekkor

$$\mathbb{P}\left(\theta \geq \frac{1}{8}\right) \geq 1 - \frac{1-\mu}{\frac{7}{8}} = \frac{\mu - \frac{1}{8}}{\frac{7}{8}} \geq \frac{\frac{1}{4} - \frac{1}{8}}{\frac{7}{8}} = \frac{1}{7}.$$

Vagyis beláttuk, hogy $\mathop{\mathbb{P}}_{S \sim \mathcal{D}^m} \Big(L_{\mathcal{D}}\big(A(S)\big) \geq \frac{1}{8} \Big) \geq \frac{1}{7}$. □

3.2. Következmény. *Legyen $\mathcal{X}$ egy végtelen alaphalmaz, és legyen $\mathcal{H}$ az a hipotézisosztály, amely az összes $\mathcal{X} \to \{0,1\}$ függvényt tartalmazza. Ekkor $\mathcal{H}$ nem PAC-megtanulható.*

Bizonyítás. Indirekt tegyük fel, hogy a hipotézisosztály megtanulható.
Válasszunk két tetszőleges $\varepsilon < 1/8$ és $\delta < 1/7$ pozitív értéket. Ekkor a PAC-megtanulhatóság definíciója szerint léteznie kell egy olyan A tanulóalgoritmusnak és $m = m^{PAC}(\varepsilon, \delta)$ számnak, hogy minden $\mathcal{X} \times \{0,1\}$ feletti $\mathcal{D}$ eloszlás esetén, ha létezik $f : \mathcal{X} \to \{0,1\}$, amelyre $L_{\mathcal{D}}(f) = 0$, akkor az $S \sim \mathcal{D}^m$ mintára:

$$\mathop{\mathbb{P}}_{S \sim \mathcal{D}^m} \big(L_{\mathcal{D},f}(A(S)) \leq \varepsilon \big) \geq 1 - \delta.$$

Alkalmazhatjuk a Nincs ingyen ebéd tételt (hiszen $|\mathcal{X}| > 2m$), ami szerint minden tanulóalgoritmusra, így A-ra is létezik olyan $\mathcal{D}'$ eloszlás és hozzá egy f függvény ami ebben az esetben eleme $\mathcal{H}$-nak, hogy

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

Ez a megválasztott ε és δ értékek miatt ellentmond az előzővel. □

4. fejezet

A Vapnik-Chervonenkis dimenzió

A fejezetben a Vapnik-Chervonenkis dimenzió fogalmát vezetjük be, és járjuk körül. A hipotézisosztályok gazdagságát kifejező mérőszámot eredetileg Vladimir Vapnik és Alexey Chervonenkis orosz statisztikusok fogalmazták meg 1968-ban. A továbbiakban kizárólag bináris klasszifikációs problémákat tekintünk az ℓ_{0-1} veszteségfüggvénnyel.

4.1. Küszöbfüggvények

Tudjuk tehát, hogy a véges hipotézisosztályok PAC-megtanulhatóak, és láttunk egy példát nem megtanulható végtelen hipotézisosztályra is. Ezek alapján jogosan gondolhatnánk, hogy tetszőleges hipotéziosztály PAC-megtanulhatóságát annak számossága határozza meg. A feltételezés azonban nem igaz, ahogyan azt a következő példa is mutatja.

Tekintsük azt a hipotézisosztályt, ami a küszöbfüggvényeket tartalmazza, ahol a h_a küszöbfüggvény:

$$h_a : \mathbb{R} \to \{0,1\} \quad \text{és} \quad h_a(x) = \begin{cases} 0 & \text{ha} \quad x < a \\ 1 & \text{ha} \quad x \geq a \end{cases}$$

Az így meghatározott $\mathcal{H}_{küszöb} = \{h_a : a \in \mathbb{R}\}$ nyilván végtelen.

4.1. Állítás. *$\mathcal{H}_{küszöb}$ PAC-megtanulható bármilyen* ERM*-algoritmussal.*

Bizonyítás. Legyen $a^* \in \mathbb{R}$ olyan, amelyre $L_{\mathcal{D}}(h_{a^*}) = 0$. Jelölje $\mathcal{D}_{\mathcal{X}}$ az $\mathcal{X}$ alaphalmaz feletti peremeloszlást. Válasszuk meg az $a_0 < a^* < a_1$ értékeket az alábbi módon:

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

Ha $\mathcal{D}_x(-\infty, a^*) \leq \varepsilon$, akkor $a_0 = -\infty$, és a_1-re hasonlóan. Adott S tanulóhalmaz esetén legyen $b_0 = \max\{x : (x,0) \in S\}$ (ha nem létezik ilyen, akkor $b_0 = -\infty$), és legyen $b_1 = \min\{x : (x,1) \in S\}$ (vagy ha S minden eleme 0 címkéjű, akkor $b_1 = \infty$).

Jelölje b_S egy $h_S \in \mathrm{ERM}(S)$ hipotézishez tartozó küszöbértéket. Ekkor értelemszerűen $b_S \in (b_0, b_1)$ (ez, mint máshol is, pontosabban 1 valószínűséggel igaz). Ezért, $L_{\mathcal{D}}(h_S) \leq \leq \varepsilon$ igazolásához elég belátni, hogy $b_0 \geq a_0$ és $b_1 \leq a_1$, ezért

$$\begin{aligned} \mathop{\mathbb{P}}_{S\sim\mathcal{D}^m}\big(L_{\mathcal{D}}(h_S) > \varepsilon\big) &\leq \mathop{\mathbb{P}}_{S\sim\mathcal{D}^m}\big(b_0 < a_0 \text{ vagy } b_1 > a_1\big) \\ &\leq \mathop{\mathbb{P}}_{S\sim\mathcal{D}^m}\big(b_0 < a_0\big) + \mathop{\mathbb{P}}_{S\sim\mathcal{D}^m}\big(b_1 > a_1\big). \end{aligned}$$

Feltettük, hogy S elemei a h_{a^*} függvény által vannak címkézve, ezért minden $(x, y) \in \in S$-re, ha $x < a^*$ akkor $y = 0$. Emiatt $b_0 < a^*$ és így a $b_0 < a_0$ esemény pontosan akkor következik be, ha S egyik pontja sem esik az (a_0, a^*) intervallumba. Ennek valószínűsége a független mintavételezés miatt:

$$\begin{aligned} \mathop{\mathbb{P}}_{S\sim\mathcal{D}^m}\big(b_0 < a_0\big) = \mathop{\mathbb{P}}_{S\sim\mathcal{D}^m}\big(\forall (x,y) \in S : x \notin (a_0, a^*)\big) = \prod_{i=1}^{m} \mathop{\mathbb{P}}_{(x_i,y_i)\sim\mathcal{D}}\big(x_i \notin (a_0, a^*)\big) \\ = (1-\varepsilon)^m \leq e^{-\varepsilon m} \leq \frac{\delta}{2}, \qquad \text{ha } m > \frac{1}{\varepsilon}\ln\Big(\frac{2}{\delta}\Big). \end{aligned}$$

Hasonló m választással adódik, hogy $\mathop{\mathbb{P}}_{S\sim\mathcal{D}^m}\big(b_1 > a_1\big) \leq \frac{\delta}{2}$. Találtunk tehát egy olyan, ε-tól és δ-tól függő m mintaelemszámot, hogy ha legalább ennyi adatpontot kapunk, akkor $1-\delta$ valószínűséggel legfeljebb ε a hibája az algoritmusnak.

4.2. A VC-dimenzió

Adódik az igény, hogy hipotézisosztályok egy olyan jellemzőjét keressük, amely a PAC-tanulhatóságot meghatározza. Mint azt később látni fogjuk, a most bevezetendő Vapnik-Chervonenkis dimenzió pontosan alkalmas lesz erre.

4.2. Definíció. *Legyen $\mathcal{H}$ hipotézisosztály, amelynek elemei $\mathcal{X} \to \{0,1\}$ függvények, és legyen $C = \{c_1, ..., c_m\} \subset \mathcal{X}$. Ekkor $\mathcal{H}$ megszorítását a C halmazra jelöljük a következővel:*

$$\mathcal{H}_C = \big\{g : C \to \{0,1\} : \exists h \in \mathcal{H} \text{ hogy } h(c_i) = g(c_i)\ \forall c_i \in C\big\}.$$

Mivel $\mathcal{H}_C$ elemei $\{0,1\}$-be képző függvények, ezért reprezentálhatók $\{0,1\}^m$-beli vektorokként is:

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

Amennyiben $\mathcal{H}$-nak a C-re való megszorítása az összes $C \to \{0,1\}$ függvény, akkor azt mondjuk, hogy $\mathcal{H}$ telíti C-t, vagy hogy C teljes.

4.3. Definíció. *A $\mathcal{H}$ hipotézisosztály telíti a véges $C \subset \mathcal{X}$ halmazt, ha $\mathcal{H}_C$ az összes $C \to \{0,1\}$ függvény halmaza, vagyis $|\mathcal{H}_C| = 2^{|C|}$*

Figyeljük meg, hogy ha van egy $\mathcal{H}$ által telített C halmazunk, akkor tetszőleges algoritmusra tudunk mutatni egy olyan eloszlást, amin a tanulóalgoritmus sikertelen lesz, ha $m \leq \frac{|C|}{2}$ és ε és δ megfelelően kicsik, hiszen a 3.1 tétel bizonyításában az eloszlást a C-re megszorított összes függvényből generáltuk, ami ebben az esetben megegyezik $\mathcal{H}_C$-val.

4.4. Következmény. *Legyen $\mathcal{H}$ hipotézisosztály, amelynek elemei $\mathcal{X} \to \{0,1\}$ függvények, és legyen m a tanulóhalmaz mérete. Tegyük fel, hogy létezik olyan teljes $C \subset \mathcal{X}$ halmaz, amelyre $|C| = 2m$. Ekkor minden A tanulóalgoritmushoz létezik olyan $\mathcal{X} \times \{0,1\}$ feletti $\mathcal{D}$ eloszlás és $h \in \mathcal{H}$ hipotézis, hogy $L_{\mathcal{D}}(h) = 0$, de*

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

A fenti fogalmak segítségével definiáljuk egy hipotézisosztály Vapnik-Chervonenkis dimezióját.

4.5. Definíció. *VC-dimenzió*
A $\mathcal{H}$ hipotézisosztály VC-dimenziója a legnagyobb $C \subset \mathcal{X}$ halmaz mérete, amelyet $\mathcal{H}$ telít. Ha $\mathcal{H}$ tetszőlegesen nagy halmazt telít, akkor azt mondjuk, hogy VC-dimenziója végtelen.

4.3. Példák

Ha igazolni akarjuk, hogy $\mathrm{VCdim}(\mathcal{H}) = d$, az alábbiakat kell belátni:

1. Létezik $\{C \subset \mathcal{X} : |C| = d\}$ úgy, hogy $\mathcal{H}$ telíti C-t.
2. Nem létezik olyan $\{C' \subset \mathcal{X} : |C'| = d+1\}$, amit $\mathcal{H}$ telít.

A következőkben néhány egyszerű hipotézisosztály VC-dimenzióját határozzuk meg. [7]

Küszöbfüggvények
Tekintsük az előzőekben definiált $\mathcal{H}_{küszöb}$ hipotézisosztályt, és legyen $C = \{c_1\}$ tetszőleges egyelemű halmaz. Be kell látnunk, hogy $\mathcal{H} = \mathcal{H}_{küszöb}$-nek erre a C-re való megszorítása lehet 0 és 1 is. Ez nyilvánvalóan teljesül, hiszen ha $a > c_1$ akkor a h_a hipotézisre $h_a(c_1) = 0$, míg tetszőleges $b < c_1$-re $h_b(c_1) = 1$. Most tekintsük a $C = \{c_1, c_2\}$ halmazt, ahol egyszerűség kedvéért legyen $c_1 < c_2$. Könnyen találhatunk olyan hipotéziseket, amelyek megszorítása rendre a $(0,0), (0,1)$ és $(1,1)$ értékek. Viszont $(1,0) \notin \mathcal{H}_C$, ugyanis ha egy hipotézis a c_1 ponthoz 1-et rendel, akkor a c_2-höz is. Ez tetszőleges kételemű C-re teljesül, ezért $\mathcal{H}$ nem telít egyetlen kételemű halmazt sem, így $\mathrm{VCdim}(\mathcal{H}_{küszöb}) = 1$.

Intervallumok

Legyen $\mathcal{H} = \{h_{a,b} : a, b \in \mathbb{R}, a < b\}$ az a hipotézisosztály, melynek elemei az (a, b) nyílt intervallum pontjaihoz 1-et rendelnek:

$$h_{a,b}(x) = \begin{cases} 1 & \text{ha} \quad x \in (a, b) \\ 0 & \text{ha} \quad x \notin (a, b) \end{cases}$$

Egyszerűen látható, hogy a kételemű C halmazokat $\mathcal{H}$ telíti, így $\text{VCdim}(\mathcal{H}) \geq 2$. Viszont tetszőleges $C = \{c_1, c_2, c_3 : c_1 < c_2 < c_3\}$ halmazt $\mathcal{H}$ nem telíthet, ugyanis az (1,0,1) címkézés nem állítható elő, mivel ez azt jelentené, hogy $c_1, c_3 \in (a, b)$ és $c_2 \notin (a, b)$, ami ellentmond annak a feltevésnek, hogy $c_2 \in (c_1, c_3)$. Így tehát $\text{VCdim}(\mathcal{H}) = 2$.

Tengelyekkel párhuzamos oldalú téglalapok

Nézzük meg **2.** fejezet elején bevezetett példának a hipotézisosztályát:

$$\mathcal{H} = \left\{h_{(a_1,a_2,b_1,b_2)} : a_1 \leq a_2 \text{ és } b_1 \leq b_2\right\}, \quad \text{ahol} \quad h_{(a_1,a_2,b_1,b_2)} : \mathbb{R}^2 \to \{0,1\}$$

$$\text{és} \quad h_{(a_1,a_2,b_1,b_2)}(x_1, x_2) = \begin{cases} 1 & \text{ha } a_1 \leq x_1 \leq a_2 \text{ és } b_1 \leq x_2 \leq b_2 \\ 0 & \text{különben} \end{cases}$$

4.1. ábra. *Négy és ötelemű ponthalmazok*

Ekkor könnyedén találhatunk 4 olyan pontot, amelyet $\mathcal{H}$ telít (1.ábra). Tekintsünk egy tetszőleges $C \subset \mathbb{R}^2$ ötelemű halmazt. Rögzítsük C pontjai közül azokat, amelyeknek az első koordinátája a legkisebb, illetve a legnagyobb, majd amelyeknek a második koordinátája a legkisebb illetve a legnagyobb. Jelölje c_5 azt a pontot, amelyet nem választottunk ki ilyen módon. Ekkor az (1,1,1,1,0) címkézést egyetlen $h_{(a_1,a_2,b_1,b_2)}$ hipotézis sem tudja előállítani, ugyanis minden ilyen esetben a téglalap tartalmazza a c_1, c_2, c_3, c_4 pontokat, de ekkor a c_5 pontot is tartalmaznia kell. Vagyis $\mathcal{H}$ nem telíthet 5 elemű halmazt, ezért $\text{VCdim}(\mathcal{H}) = 4$.

Végtelen VC-dimenzió

Fontos megjegyezni, hogy ugyan az eddig vizsgált példákban a hipotézisosztály VC-dimenziója megegyezett a paramétereinek számával, ez azonban nem szükségszerű. Tekintsük például a $\mathcal{H} = \{h_{\omega,\theta} : \omega, \theta \in \mathbb{R}\}$ osztályt, ahol

$$h_{\omega,\theta} : \mathbb{R} \to \{0,1\} \text{ és } h_{\omega,\theta}(x) = \lceil 0.5 \sin(\omega x + \theta) \rceil.$$

Ekkor $\mathcal{H}$-nak mindössze két paramétere van, azonban ezeknek megfelelő beállításával tetszőleges d-re található d olyan pont a számegyenesen, amelyet $\mathcal{H}$ telít, ugyanis az ω paraméterrel a szinuszfüggvény frekvenciáját tetszőlegesen megnövelhetjük, így tetszőlegesen közeli pontok is kaphatnak különböző címkét, a θ paraméter pedig a függvény vízszintes eltolását teszi lehetővé. Ezáltal $\mathrm{VCdim}(\mathcal{H}) = \infty$.

4.3.1. Véges hipotézisosztályok

Legyen $\mathcal{H}$ egy véges hipotézisosztály. Ekkor nyilvánvalóan $|\mathcal{H}_C| \leq |\mathcal{H}|$, így $\mathcal{H}$ nem telíthet olyan C-t, amelyre $|\mathcal{H}| < 2^{|C|}$. Ebből adódik a véges hipotézisosztályok VC-dimenziójára vonatkozó becslés:

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

Ez sokszor nagyon gyenge korlátot határoz meg. Például legyen $\mathcal{X} = \{1, \ldots, k\}$, és a hipotézisosztályt definiáljuk úgy, hogy a $\mathcal{H}_{küszöb}$ hipotézisosztályt megszorítjuk k-ra. Ekkor ennek VC-dimenziója továbbra is 1, azonban a mérete k. Ezáltal a méretének a logaritmusa is tetszőlegesen nagy lehet.

4.4. Növekedési függvény

A $\mathcal{H}$ hipotéziszosztály növekedési függvényét interpretálhatjuk az osztály gazdagságát mérő függvényként. $\mathrm{VCdim}(\mathcal{H})$-nál kisebb vagy egyenlő elemszámú tanulóhalmazok esetén $\mathcal{H}$ nagyon gazdag, hiszen van olyan halmaz ezek között, amelyet telít, vagyis rajta az összes lehetséges címkézést elő tudja állítani. Érdekes azt vizsgálni, hogy $\mathrm{VCdim}(\mathcal{H})$-nál nagyobb értékek esetén a növekedési függvény milyen viselkedést mutat. Ehhez kapcsolódó fontos eredmény a Sauer-Shelah-Perles lemma, amelyet később a statisztikai tanuláselmélet alaptételének bizonyításában is fel fogunk használni.

4.6. Definíció. *Növekedési függvény*
Definiáljuk a $\mathcal{H}$ hipotézisosztály $\tau_{\mathcal{H}}: \mathbb{N} \to \mathbb{N}$ növekedési függvényét a következőképpen:

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

4.7. Lemma. *Sauer-Shelah-Perles*
Ha $\mathrm{VCdim}(\mathcal{H}) \leq d < \infty$ valamely $\mathcal{H}$ hipotézisosztályra, akkor a növekedési függvényt a következőképpen tudjuk becsülni tetszőleges m esetén:

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

Bizonyítás. A lemma bizonyításához elégséges a következő, erősebb állítást belátni: Tetszőleges $C = \{c_1, ..., c_m\} \subset \mathcal{X}$ halmazra fenáll:

$$\text{Minden } \mathcal{H} \text{ hipotézisosztály esetén } \left|\mathcal{H}_C\right| \leq \left|\{B \subseteq C : \mathcal{H} \text{ telíti } B\text{-t}\}\right|. \tag{4.1}$$

Ebből már következni fog a lemma állítása, ugyanis

$$\begin{aligned}\tau_{\mathcal{H}}(m) = \max_{C\subset\mathcal{X}:|C|=m} \left|\mathcal{H}_C\right| &\leq \max_{C\subset\mathcal{X}:|C|=m} \left|\{B \subseteq C : \mathcal{H} \text{ telíti } B\text{-t}\}\right| \\ &\leq \max_{C\subset\mathcal{X}:|C|=m} \left|\{B \subseteq C : |B| \leq d\}\right| = \sum_{i=0}^{d} \binom{m}{i}.\end{aligned}$$

Itt azt kell csak meggonndolni, hogy mivel $\mathcal{H}$ VC-dimenziója legfeljebb d, ezért ennél nagyobb B halmazt nem telíthet, és ekkor már a legfeljebb d elemű részhalmazok számára vonatkozó képletből adódik az összefüggés.

A 4.1 becslést m-szerinti indukcióval fogjuk belátni: $m=1$ esetén $C=\{c_1\}$, és $\mathcal{H}_C = \{h(c_1) : h \in \mathcal{H}\}$. Ha $|\mathcal{H}_C|=2$, akkor C teljes részhalmazai a C és az üres halmaz (az üres halmazt minden esetben teljesnek tekintjük), vagyis az egyenlőtlenség mindkét oldalán 2 áll. Ha $|\mathcal{H}_C| = 1$, akkor C nyilván nem lehet teljes, így az egyenlőtlenség jobb oldala is 1 lesz.
Tegyük fel, hogy az állítás minden $k \leq m$ esetén igaz, azt kell belátnunk, hogy ekkor tetszőleges $m+1$ elemű halmazra is teljesülni fog 4.1.
Ehhez rögzítsünk egy $\mathcal{H}$ hipotézisosztályt, és legyen

$$C = \{c_0, c_1, ..., c_m\}, \text{ illetve } C' = \{c_1, ..., c_m\}.$$

Definiáljuk a következő halmazokat:

$$\begin{aligned}Y_0 = \big\{g : C' \to \{0,1\} : \exists h \in \mathcal{H} \text{ amelyre vagy } h(c_i) = g(c_i)\ \forall c_i \in C' \text{ és } h(c_0) = 0 \\ \text{vagy } h(c_i) = g(c_i)\ \forall c_i \in C' \text{ és } h(c_0) = 1\big\},\end{aligned}$$

$$\begin{aligned}Y_1 = \big\{g : C' \to \{0,1\} : \exists h \in \mathcal{H} \text{ amelyre } h(c_i) = g(c_i)\ \forall c_i \in C' \text{ és } h(c_0) = 0 \\ \text{és } \exists h' \in \mathcal{H} \text{ amelyre } h'(c_i) = g(c_i)\ \forall c_i \in C' \text{ és } h'(c_0) = 1\big\}.\end{aligned}$$

Y_0 tehát azokból a függvényekből áll, amelyek C'-n megegyeznek valamely $\mathcal{H}$-beli függvénnyel. Ez pontosan $\mathcal{H}$ megszorítása C'-re, vagyis $Y_0 = \mathcal{H}_{C'}$.
Y_1 elemei pedig olyan függvények, amelyek C'-n megyegyeznek két olyan $\mathcal{H}$-beli függvénnyel, amelyek c_0-on különböző értéket vesznek fel.

Most bebizonyítjuk, hogy $|\mathcal{H}_C| = |Y_0| + |Y_1|$. Először is $|\mathcal{H}_C|$ az összes olyan $\{0,1\}^m$ vektor száma, amelyet $\mathcal{H}$ függvényei C-n előállítanak. Az Y_0 elemeivel azt számoljuk meg, hogy C'-n hány különböző $\{0,1\}^{m-1}$ vektort kapunk. Minden ilyen vektort

kiterjeszthetünk $\{0,1\}^m$ vektorrá úgy, hogy $\mathcal{H}_C$ egy-egy elemét kapjuk (előfordulhat, hogy bizonyos elemeket többféleképpen is kiterjeszthetnénk). Ha megszámoljuk továbbá azokat a vektorokat amiket kétféleképpen is kit tudtunk terjeszteni, az pontosan Y_1 elemszáma. Így végül megkaptuk $\mathcal{H}_C$ elemszámát.

Mivel Y_0 és Y_1 elemei az m elemű C' halmazon ható függvények, így alkalmazható rájuk az indukciós feltevés.

$$\big|Y_0\big| = \big|\mathcal{H}_{C'}\big| \leq \big|\{B \subseteq C' : \mathcal{H} \text{ telíti } B\text{-t}\}\big| = \big|\{B \subseteq C : c_0 \notin B \text{ és } \mathcal{H} \text{ telíti } B\text{-t}\}\big|.$$

Legyen $\mathcal{H}' \subseteq \mathcal{H}$ a következőképpen meghatározva

$$\mathcal{H}' := \big\{h \in \mathcal{H} : \exists h' \in \mathcal{H} \text{ hogy } h(c_i) = h'(c_i)\ \forall c_i \in C' \text{ és } h(c_0) \neq h'(c_0)\big\}.$$

Látható, hogy $\mathcal{H}'$ pontosan akkor telít egy B halmazt, ha telíti a $B \cup \{c_0\}$ halmazt is, illetve $Y_1 = \mathcal{H}'_{C'} = \big\{(h(c_1), ..., h(c_m)) : h \in \mathcal{H}'\big\}$. Ezért

$$\begin{aligned}\big|Y_1\big| = \big|\mathcal{H}'_{C'}\big| &\leq \big|\{B \subseteq C' : \mathcal{H}' \text{ telíti } B\text{-t}\}\big| \\ &= \big|\{B \subseteq C' : \mathcal{H}' \text{ telíti } B \cup \{c_0\}\text{-t}\}\big| \\ &= \big|\{B \subseteq C : \mathcal{H}' \text{ telíti } B\text{-t és } c_0 \in B\}\big| \\ &\leq \big|\{B \subseteq C : c_0 \in B \text{ és } \mathcal{H} \text{ telíti } B\text{-t}\}\big|.\end{aligned}$$

A fentiekből következik, hogy

$$\begin{aligned}\big|\mathcal{H}_C\big| &= \big|Y_0\big| + \big|Y_1\big| \\ &\leq \big|\{B \subseteq C : c_0 \notin B \text{ és } \mathcal{H} \text{ telíti } B\text{-t}\}\big| + \big|\{B \subseteq C : c_0 \in B \text{ és } \mathcal{H} \text{ telíti } B\text{-t}\}\big| \\ &= \big|\{B \subseteq C : \mathcal{H} \text{ telíti } B\text{-t}\}\big|.\end{aligned}$$

□

Megjegyzés. Az $m \leq d$ esetben egyenlőség teljesül, hiszen ekkor

$$\sum_{i=0}^{d} \binom{m}{i} = \sum_{i=0}^{m} \binom{m}{i} = 2^m = \tau_{\mathcal{H}}(m).$$

A lemma jelentőségét a következő összefüggés segítségével láthatjuk. Eszerint ugyanis $m > d$ esetén $\tau_{\mathcal{H}}(m) \leq (em/d)^d = \mathcal{O}(m^d)$, tehát a hipotézisosztály VC-dimenziójánál nagyobb m-ekre a növekedési függvény exponenciális helyett polinomiális növekedést mutat. Itt felhasználtuk az alábbi állítást, amit teljes indukcióval és a Stirling-formulával lehet bizonyítani.

4.8. Állítás. *Legyenek $m, d \in \mathbb{N}$ melyekre $d \leq m-2$. Ekkor*

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

5. fejezet

A statisztikai tanuláselmélet alaptétele

A fejezetben kimondjuk és bebizonyítjuk a statisztikai tanuláselmélet alaptételét, amely kapcsolatot teremt az eddigiekben vizsgált fogalmak között.

5.1. Tétel. *A statisztikai tanuláselmélet alaptétele*
Álljon a $\mathcal{H}$ hipotézisosztály $\mathcal{X} \to \{0,1\}$ függvényekből, és legyen a veszteségfüggvény az ℓ_{0-1}. Ekkor a következő állítások ekvivalensek:

(1). $\mathcal{H}$ egyenletesen konvergens

(2). Bármilyen ERM*-algoritmus sikeres agnosztikus PAC-tanuló*

(3). $\mathcal{H}$ agnosztikusan PAC-megtanulható

(4). Bármilyen ERM*-algoritmus sikeres PAC-tanuló*

(5). $\mathcal{H}$ PAC-megtanulható

(6). $\mathrm{VCdim}(\mathcal{H}) < \infty$

A bizonyítás egyszerű eseteit itt tárgyaljuk:
$(1) \Rightarrow (2)$ a 2.14 következmény szerint teljesül.
$(2) \Rightarrow (3)$ és $(4) \Rightarrow (5)$ triviális, hiszen $\mathcal{H}$ (agnosztikus) PAC-megtanulhatóságához azt követeljük meg, hogy létezzen sikeres tanulóalgoritmus.
$(2) \Rightarrow (4)$ és $(3) \Rightarrow (5)$ abból adódik, hogy az agnosztikus PAC-tanulhatóság magába foglalja a PAC-tanulhatóságot, mivel annak általánosítása.
$(5) \Rightarrow (6)$ bizonyításához tegyük fel, hogy $\mathrm{VCdim}(\mathcal{H}) = \infty$ és $\mathcal{H}$ PAC-megtanulható. Ekkor tetszőlegesen nagy m-re létezik olyan $C \subset \mathcal{X}$, amelyet $\mathcal{H}$ telít, és $|C| = 2m$. Ám ekkor a 4.4 következmény szerint $\mathcal{H}$ nem PAC-megtanulható.
A $(6) \Rightarrow (1)$ irány a nehéz, ehhez további tételek kellenek.

5.1. A bizonyításhoz szükséges lemmák és tételek

Ebben az alfejezetben több, főként valószínűségszámítási lemmát és tételt mondunk ki és látunk be, amelyek mindegyikét felhasználjuk a későbbiekben. Kiemelkedő jelentőségű a 5.6 tétel illetve az előző részben bizonyított Sauer-Shelah-Perles lemma, amelyek az alaptétel bizonyításának meghatározó részét alkotják.

5.2. Lemma. *Legyen X egy valószínűségi változó és $x' \in \mathbb{R}$. Tegyük fel, hogy létezik $a > 0$ és $b \geq e$, hogy minden $t \geq 0$ esetén $\mathbb{P}\big(|X - x'| \geq t\big) \leq 2b \exp\left(-\frac{t^2}{a^2}\right)$. Ekkor*

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

Bizonyítás. Legyen $t_i = a\big(i + \sqrt{\ln(b)}\big) (i = 0,1,2,...)$ sorozat.Mivel a t_i sorozat monoton növekvő, a várható érték felülről becsülhető a következővel:

$$\mathbb{E}\big(|X - x'|\big) \leq t_0 \mathbb{P}\big(|X - x'| \geq 0\big) + \sum_{i=1}^{\infty} t_i \mathbb{P}\big(|X - x'| \geq t_{i-1}\big) \leq$$
$$\leq a\sqrt{\ln(b)} + \sum_{i=1}^{\infty} t_i \mathbb{P}\big(|X - x'| \geq t_{i-1}\big).$$

Az itt szereplő második tagot az alábbi módon felülről tudjuk becsülni a lemmában szereplő feltevés, és egyszerű átalakítások segítségével:

$$\sum_{i=1}^{\infty} t_i \mathbb{P}\big(|X - x'| \geq t_{i-1}\big) \leq \sum_{i=1}^{\infty} a\big(i + \sqrt{\ln(b)}\big) 2be^{-\left(i-1+\sqrt{\ln(b)}\right)^2}$$
$$\leq 2ab \int_{1+\sqrt{\ln(b)}}^{\infty} x e^{-(x-1)^2} dx$$
$$= 2ab \int_{\sqrt{\ln(b)}}^{\infty} (y+1) e^{-y^2} dy$$
$$\leq 4ab \int_{\sqrt{\ln(b)}}^{\infty} y e^{-y^2} dy = 2ab\Big[-e^{-y^2}\Big]_{\sqrt{\ln(b)}}^{\infty}$$
$$= \frac{2ab}{b} = 2a.$$

Ezt visszahelyettesítve a lemma állítását kapjuk. □

5.3. Tétel. *Markov-egyenlőtlenség*
Legyen Z egy nemnegatív valószínűségi változó. Ekkor minden $a \geq 0$ esetén

$$\mathbb{P}(Z \geq a) \leq \frac{\mathbb{E}(Z)}{a}.$$

Bizonyítás. A várható érték definíciója, és Z nemnegativitása miatt

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

Mivel $\mathbb{P}(Z \geq a)$ monoton csökkenő, ezért minden $a \geq 0$ esetén

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

amit átrendezve a fenti egynelőtlenséghez jutunk. □

5.4. Lemma. *Hoeffding lemma*
Legyen X egy valószínűségi változó, amelyre $\mathbb{E}(X) = 0$ *és* $\mathbb{P}(a \leq X \leq b) = 1$. *Ekkor minden* $\lambda > 0$ *esetén*

$$\mathbb{E}\big(e^{\lambda X}\big) \leq \exp\left(\frac{\lambda^2(b-a)^2}{8}\right).$$

Bizonyítás. Az $f(x) = e^{\lambda x}$ egy konvex függvény, ezért minden $\alpha \in [0,1]$ esetén

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

Legyen $x \in [a,b]$ és $\alpha = \frac{b-x}{b-a} \in [0,1]$. Ekkor

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

Mindkét oldal várható értékét véve

$$\mathbb{E}\big(e^{\lambda x}\big) \leq \frac{b-\mathbb{E}(X)}{b-a}e^{\lambda a} + \frac{\mathbb{E}(X)-a}{b-a}e^{\lambda b} = \frac{b}{b-a}e^{\lambda a} + \frac{a}{b-a}e^{\lambda b}. \tag{5.1}$$

Legyen

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

Ekkor az (5.1) egyenlőtlenség jobb oldalán szereplő kifejezés pontosan $e^{L(h)}$. Így a lemma igazolásához elég belátnunk, hogy $L(h) \leq \frac{h^2}{8}$. Felhasználjuk, hogy valamely $c \in [0,h]$ esetén:

$$L(h) = L(0) + hL'(0) + \frac{h^2}{2!}L''(c).$$

Mivel $L(0) = L'(0) = 0$ és minden h-ra teljesül, hogy $L''(h) \leq \frac{1}{4}$, ezzel a lemmát beláttuk. □

5.5. Tétel. *Hoeffding-egyenlőtlenség*
Legyenek $Z_1, \ldots Z_m$ *független, azonos eloszlású valószínűségi változók, és legyen* $\bar{Z} = \frac{1}{m}\sum_{i=1}^{m} Z_i$.
Tegyük fel, hogy $\mathbb{E}(\bar{Z}) = \mu$ *és minden i-re:* $\mathbb{P}(a \leq Z_i \leq b) = 1$. *Ekkor minden* $\varepsilon \geq 0$ *esetén*

$$\mathbb{P}\left(\left|\frac{1}{m}\sum_{i=1}^{m} Z_i - \mu\right| \geq \varepsilon\right) \leq 2\exp\left(\frac{-2m\varepsilon^2}{(b-a)^2}\right).$$

Bizonyítás. Legyen $X_i = Z_i - \mathbb{E}(Z_i)$ és $\bar{X} = \frac{1}{m}\sum_{i=1}^{m} X_i$. Ekkor

$$\bar{X} = \frac{1}{m}\sum_{i=1}^{m}\big(Z_i - \mathbb{E}(Z_i)\big) = \frac{1}{m}\sum_{i=1}^{m} Z_i - \mathbb{E}\bigg(\frac{1}{m}\sum_{i=1}^{m} Z_i\bigg) = \frac{1}{m}\sum_{i=1}^{m} Z_i - \mu.$$

Vagyis azt kell belátnunk, hogy

$$\mathbb{P}\big(|\bar{X}| \geq \varepsilon\big) \leq 2\exp\left(\frac{-2m\varepsilon^2}{(b-a)^2}\right).$$

Az exponenciális függvény monotonitását és a Markov-egyenlőtlenséget felhasználva, minden $\lambda > 0$ és $\varepsilon \geq 0$ esetén

$$\mathbb{P}\big(\bar{X} \geq \varepsilon\big) = \mathbb{P}\Big(e^{\lambda\bar{X}} \geq e^{\lambda\varepsilon}\Big) \leq e^{-\lambda\varepsilon}\mathbb{E}\big(e^{\lambda\bar{X}}\big).$$

Továbbá kihasználjuk, hogy mivel a Z_i-k független valószínűségi változók, így X_i-k is, és független valószínűségi változók szorzatának várható értéke a várható értékeik szorzatával egyenlő:

$$\mathbb{E}\big(e^{\lambda\bar{X}}\big) = \mathbb{E}\Big(e^{\lambda\frac{1}{m}\sum X_i}\Big) = \mathbb{E}\bigg(\prod_{i=1}^{m} e^{\frac{\lambda X_i}{m}}\bigg) = \prod_{i=1}^{m}\mathbb{E}\big(e^{\frac{\lambda X_i}{m}}\big).$$

A Hoeffding lemma szerint minden i-re

$$\mathbb{E}\big(e^{\frac{\lambda X_i}{m}}\big) \leq e^{\frac{\lambda^2(b-a)^2}{8m^2}},$$

vagyis

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

ahol $\lambda = \frac{4m\varepsilon}{(b-a)^2}$ választással

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

A $-\bar{X}$ valószínűségi változóra az előző lépéseket alkalmazva adódik, hogy

$$\mathbb{P}\big(\bar{X} \leq -\varepsilon\big) \leq e^{-\frac{2m\varepsilon^2}{(b-a)^2}}.$$

Tehát

$$\mathbb{P}\big(|\bar{X}| \geq \varepsilon\big) = \mathbb{P}\Big(\big(\bar{X} \geq \varepsilon\big) \cup \big(\bar{X} \leq -\varepsilon\big)\Big) \leq \mathbb{P}\big(\bar{X} \geq \varepsilon\big) + \mathbb{P}\big(\bar{X} \leq -\varepsilon\big) \leq 2e^{-\frac{2m\varepsilon^2}{(b-a)^2}}.$$

□

5.6. Tétel. *Legyen a $\mathcal{H}$ hipotézisosztály növekedési függvénye $\tau_{\mathcal{H}}$. Ekkor tetszőleges $\mathcal{D}$ eloszlás és $\delta \in (0,1)$ esetén:*

$$\mathop{\mathbb{P}}_{S\sim\mathcal{D}^m}\left(\left|L_{\mathcal{D}}(h)-L_S(h)\right| \leq \frac{4+2\sqrt{\ln(\tau_{\mathcal{H}}(2m))}}{\delta\sqrt{2m}}\right) \geq 1-\delta.$$

Bizonyítás. Elegendő belátni a következőt:

$$\mathop{\mathbb{E}}_{S\sim\mathcal{D}^m}\left(\sup_{h\in\mathcal{H}}\left|L_{\mathcal{D}}(h)-L_S(h)\right|\right) \leq \frac{4+2\sqrt{\ln(\tau_{\mathcal{H}}(2m))}}{\sqrt{2m}}. \tag{5.2}$$

Ekkor ugyanis $Z = \sup_{h\in\mathcal{H}} |L_{\mathcal{D}}(h)-L_S(h)|$ egy nemnegatív valószínűségi változó, így alkalmazható rá a Markov-egyenlőtlenség. Legyen

$$a = \frac{4+2\sqrt{\ln(\tau_{\mathcal{H}}(2m))}}{\delta\sqrt{2m}}.$$

Ekkor a fenti egyenlőtlenséget felhasználva

$$\mathbb{P}(Z \geq a) \leq \frac{\mathbb{E}(Z)}{a} \leq \frac{\frac{4+2\sqrt{\ln(\tau_{\mathcal{H}}(2m))}}{\sqrt{2m}}}{\frac{4+2\sqrt{\ln(\tau_{\mathcal{H}}(2m))}}{\delta\sqrt{2m}}} = \delta.$$

Visszahelyettesítés után a következő összefüggést kapjuk:

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

amelyből közvetlenül adódik a tétel állítása.

Az 5.2 egyenlőtlenség igazolásához először is vegyük észre, hogy minden $h \in \mathcal{H}$ és $S' = (z'_1, ..., z'_m)$ független, azonos eloszlású minta esetén $L_{\mathcal{D}}(h) = \mathop{\mathbb{E}}_{S'\sim\mathcal{D}^m}\left[L_{S'}(h)\right]$, ugyanis

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

Az utolsó egyenlőség azért teljesül, mert S' egy független, azonos eloszlású minta, így tetszőleges z'_i elemére vonatkozó veszteség várható értéke megegyezik azzal, mint ha csak egyetlen elemet mintavételeznénk ugyanabból a $\mathcal{D}$ eloszlásból, és annak várható veszteségét tekintenénk. Így, az előbbit behelyettesítve, illetve a Jensen-egyenlőtlenséget alkalmazva:

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

Mivel S és S' ugyanabból az eloszlásból származó független minták, ezért a z_i és z_i' jelöléseket szabadon felcserélhetjük, a fenti egyenlőtlenségnek az utolsó tagja nem változik. Felcseréléskor az $\ell(h,z_i')-\ell(h,z_i)$ tag helyett ennek -1-szerese fog szerepelni a fenti összegben. Emiatt minden $\boldsymbol{\sigma}\in\{\pm1\}^m$ esetén:

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

Ez minden $\boldsymbol{\sigma}\in\{\pm1\}^m$ esetén teljesül, így ha $\boldsymbol{\sigma}$ komponenseit egymástól függetlenül mintavételezzük a $\{\pm1\}$ feletti egyenletes eloszlásból ($E_\pm$), akkor ez megegyezik a következővel:

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

ami a várható érték linearitása miatt pontosan

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

Rögzítsük az S és S' halmazokat, és jelölje C ezek elemeinek halmazát. Világos, hogy elég azt vizsgálnunk hogy a h hipotézisek a C elemein milyen viselkedést mutatnak, ezért elég a szuprémumot a $\mathcal{H}_C$ hipotéziosztály elemeire vizsgálni, vagyis

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

Rögzítsük $h\in\mathcal{H}_C$-t és legyen $\theta_h=\frac{1}{m}\sum_{i=1}^m \sigma_i\big(\ell(h,z_i')-\ell(h,z_i)\big)$. Ekkor $\mathbb{E}(\theta_h)=0$ és θ_h független, azonos eloszlású, korlátos valószínűségi változók átlaga, így a Hoeffding-egyenlőtlenség szerint:

$$\forall \rho \geq 0 \text{ esetén} \quad \mathbb{P}\big(|\theta_h| \geq \rho\big) \leq 2\exp\left(-\frac{m\rho^2}{2}\right).$$

Ekkor

$$\mathbb{P}\left(\max_{h\in\mathcal{H}_C}|\theta_h| \geq \rho\right) = \mathbb{P}\left(\bigcup_{h_i\in\mathcal{H}_C}\big|\theta_{h_i}\big| \geq \rho\right) \leq \big|\mathcal{H}_C\big|\mathbb{P}\big(|\theta_{h_i}| \geq \rho\big)$$
$$\leq 2\big|\mathcal{H}_C\big|\exp\left(-\frac{m\rho^2}{2}\right).$$

Alkalmazzuk az 5.2 lemmát $a = \sqrt{2/m}$, $b = |\mathcal{H}_c|$ és $t = \rho$ választással. Így

$$\mathop{\mathbb{E}}_{\boldsymbol{\sigma}\sim E_{\pm}^m}\left(\max_{h\in\mathcal{H}_C}|\theta_h|\right) \leq \frac{4+2\sqrt{\ln|\mathcal{H}_C|}}{\sqrt{2m}}$$

A fentiekből, és $\tau_{\mathcal{H}}$ definíciójából adódik, hogy

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

□

5.2. Az alaptétel bizonyítása

A (6) ⇒ (1) irányhoz azt kell belátnunk, hogy $\mathrm{VCdim}(\mathcal{H}) = d < \infty$ esetén létezik olyan $m_{\mathcal{H}}^{UC} : (0,1)^2 \to \mathbb{N}$ függvény, hogy $m \geq m_{\mathcal{H}}^{UC}(\varepsilon,\delta)$ méretű minta esetén $\mathcal{H}$ egyenletesen konvergens. Tehát olyan (ε-tól és δ-tól függő) alsó korlátot fogunk mutatni m értékére, hogy a legalább ekkora elemszámú S tanulóhalmaz legalább $(1-\delta)$ valószínűséggel ε-reprezentatív legyen. Az 5.6 tétel szerint:

$$\mathop{\mathbb{P}}_{S\sim\mathcal{D}^m}\left(\big|L_{\mathcal{D}}(h) - L_S(h)\big| \leq \frac{4+2\sqrt{\ln(\tau_{\mathcal{H}}(2m))}}{\delta\sqrt{2m}}\right) \geq 1-\delta.$$

Mivel esetünkben $\mathrm{VCdim}(\mathcal{H}) = d < \infty$, ezért alkalmazhatjuk a Sauer-Shelah-Perles lemmát, amely szerint $m > d$ esetén $\tau_{\mathcal{H}}(2m) \leq (2em/d)^d$, így

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

Ekkor már csak azt kell megmutatnunk, hogy alkalmas m-re

$$\frac{4+2\sqrt{d\ln(2em/d)}}{\delta\sqrt{2m}} \leq \varepsilon.$$

Mivel $\lim_{m\to\infty}\ln(m)/m = 0$, ezért létezik ilyen m. □

Összefoglalás

A szakdolgozatban a statisztikai tanuláselmélet alapvető eredményeit vettük végig. Bevezetésként a gépi tanulás főbb motivációival, kérdéseivel ismerkedtünk meg. Feladatunk matematikailag precíz felírása után a tanuláselmélet PAC-féle modelljével foglalkoztunk. A PAC-megtanulható hipotézisosztályok esetén a pontossági- illetve konfidenciaparaméterek a tanulóhalmaz méretével garantálják a tanulás sikerességét. A Nincs ingyen ebéd tételből láthattuk, hogy nem létezik olyan univerzális tanulóalgoritmus, amelyik minden feladaton jól teljesítene. Megállapítottuk, hogy habár igaz, hogy a véges hipotézisosztályok mind PAC-megtanulhatóak, a tanulhatóságot mégsem az osztályok számossága határozza meg, hanem a Vapnik-Chervonenkis dimenziójuknak a végessége. Szemléletes, egyszerűen értelmezhető példákon keresztül ismertettük ezt a nehezen megfogható fogalmat. A Sauer-Shelah-Perles lemma következményeként láttuk, hogy a VC-dimenziót túllépve a növekedési függvény polinomiálisan növekszik tovább. Végül bebizonyítottuk a statisztikai tanuláselmélet alaptételét, aminek legfontosabb állítása, hogy bináris klasszifikáció esetében a $0-1$ veszteségfüggvény mellett egy hipotézisosztály pontosan akkor PAC-megtanulható, ha a VC-dimenziója véges.

Irodalomjegyzék

[1] Christopher M Bishop. *Pattern recognition and machine learning*. Springer, 2006.

[2] Pedro Domingos. „The role of Occam's razor in knowledge discovery". *Data mining and knowledge discovery* 3.4 (1999), 409–425. old.

[3] Ian Goodfellow, Yoshua Bengio és Aaron Courville. *Deep learning*. MIT press, 2016.

[4] Tom M. Mitchell. *Machine learning*. McGraw-Hill, 1997.

[5] Kevin P Murphy. *Machine learning: a probabilistic perspective*. MIT press, 2012.

[6] Hajnal Péter. *Halmazrendszerek*. Polygon, 2002.

[7] Shai Shalev-Shwartz és Shai Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, 2014.

[8] Richard S Sutton és Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

[9] Leslie Valiant. „A theory of the learnable". *Proceedings of the sixteenth annual ACM symposium on Theory of computing*. ACM. 1984, 436–445. old.

[10] Leslie Valiant. *Probably Approximately Correct: Nature's Algorithms for Learning and Prospering in a Complex World*. Basic Books (AZ), 2013.

[11] Vladimir Vapnik. *The nature of statistical learning theory*. Springer, 2013.