Eötvös Loránd Tudományegyetem
Természettudományi Kar

Novák Benedek Bálint

Matematika BSc
Alkalmazott matematikus szakirány

Kernelizált ridge és szupport vektor regresszió

Szakdolgozat

Témavezető:
Dr. Csáji Balázs Csanád
ELTE Valószínűségelméleti és Statisztika Tanszék

Budapest, 2023

NYILATKOZAT

Név: Novák Benedek Bálint

ELTE Természettudományi Kar, szak: Matematika BSc

NEPTUN azonosító: J470LN

Szakdolgozat címe:
Kernelizált ridge és szupport vektor regresszió

A **szakdolgozat** szerzőjeként fegyelmi felelősségem tudatában kijelentem, hogy a dolgozatom önálló szellemi alkotásom, abban a hivatkozások és idézések standard szabályait következetesen alkalmaztam, mások által írt részeket a megfelelő idézés nélkül nem használtam fel.

Budapest, 2023. 06. 04.

Novák Benedek

a hallgató aláírása

Köszönetnyilvánítás

Ezúton szeretnék köszönetet mondani a témavezetőmnek, Csáji Balázs Csanádnak a témakör megismertetéséért, a felmerült kérdések megválaszolásáért és a segítségért amit a félév folyamán kaptam tőle.

Tartalomjegyzék

Bevezetés

A gépi tanulás olyan algoritmusokkal foglalkozik, melyek a rendelkezésre álló adat alapján tudnak különböző összefüggéseket felfedezni. Két főbb iránya van, a felügyelt és a felügyeletlen tanulás. (Illetve egy harmadik, fontos ágazat a megerősítéses tanulás, azonban ez meglehetősen különbözik az előző kettőtől, ezért ezt gyakran külön szokták tárgyalni.) Felügyelt tanulás esetén az a feladat, hogy egy $\mathcal{X}$ halmaz elemeihez szeretnénk hozzárendelni megfelelő $\mathcal{Y}$-beli értékeket. Például lehet $\mathcal{X}$ a kézzel írott számjegyekről készült képek és $\mathcal{Y}$ pedig a $0, \ldots, 9$ számok. Ebben az esetben $\mathcal{Y}$ elemszáma véges, ezt a feladatot klasszifikációnak nevezzük. Amennyiben egy folytonos értéket szeretnénk becsülni, például $\mathcal{Y} = \mathbb{R}$, regresszióról beszélünk. Mindkét esetben a rendelkezésünkre áll egy $(x_1, y_1), \ldots, (x_n, y_n) \in \mathcal{X} \times \mathcal{Y}$ tanító adat, és azt szeretnénk megmondani, hogy még nem látott x-ekhez milyen y tartozhat. Felügyeletlen tanulás esetén csak $\mathcal{X}$-beli minta áll rendelkezésünkre, és ebből próbálunk meg következtetéseket levonni.

Ebben a szakdolgozatban főleg regressziós modellekkel fogunk foglalkozni, de lesz szó klasszifikációról is. Az egyik legegyszerűbb regressziós modell a lineáris regresszió. Ekkor adott egy $(x_1, y_1), \ldots, (x_n, y_n) \in \mathbb{R}^d \times \mathbb{R}$ minta és keressük azt a $\vartheta \in \mathbb{R}^d$-t, amire a $f(x) = \langle \vartheta, x \rangle$ függvény a lehető legjobban illeszkedik ezekre a pontokra. A lineáris regressziónak nagy előnye, hogy gyorsan számítható, azonban csak lineáris összefüggéseket tudunk felfedezni a használatával.

Ezt meg tudjuk oldani úgy, hogy nem az eredeti $\mathcal{X}$ vektortéren alkalmazzuk a modellt, hanem beleképezzük az x_i-ket egy Φ leképezés segítségével egy úgynevezett tulajdonságtérbe és itt alkalmazzuk a regressziós modellt (azaz $f(x) = \langle \vartheta, \Phi(x) \rangle$). Azonban ha a tulajdonságtér dimenziója exponenciálisan növekszik $\mathcal{X}$ dimenziójában, akkor ezt a leképezést nagyon költséges kiszámolni, ráadásul a kiszámolandó ϑ paraméter mérete is kezelhetetlenül nagy lesz.

Erre megoldást nyújt majd a Reprezentációs tétel, ami segítségével megkapjuk, hogy f felírható úgy, mint a mintaelemekkel vett skaláris szorzatok egy lineáris kombinációja, vagyis

$$f(x) = \sum_{i=i}^{n} \alpha_i \langle \Phi(x), \Phi(x_i) \rangle$$

Tehát tudjuk használni f-et anélkül, hogy szükségünk lenne a ϑ paraméterre. Azonban ehhez továbbra is kellene számolni a Φ leképezéseket, majd az ezekkel vett skaláris szorzatot. Azonban ha a

megfelelő α megtalálásához olyan algoritmust használunk, amiben a $\Phi(x_i)$-k csak skaláris szorzáson belül szerepelnek, akkor összevonhatjuk a két műveletet ha találunk rá egy explicit képletet, így egy kernel függvényhez jutva:

$$k(x, x') = \langle \Phi(x), \Phi(x') \rangle$$

Így már akár végtelen dimenziós tulajdonságtereket is tudunk használni, hiszen nem kell kiszámolni a leképezést az optimális α megtalálásához, illetve a kiértékelésnél is használhatjuk a kernel függvényt.

Egy másik probléma még az lehet, hogy ha sok tanítási adatunk van, akkor az összes $k(x, x_i)$-t ki kell számolnunk a függvény kiértékeléséhez, illetve ekkor előfordulhat, hogy sok nagyon kicsi α_i van, ami numerikus hibákhoz is vezethet. Sokkal vonzóbb lenne egy olyan ritka reprezentáció, ami csak kevesebb x_i-t használ nagyobb együtthatókkal. Az utolsó fejezetben szereplő szupport vektor gépek pont pont egy ilyet fognak adni.

1. fejezet

Statisztikai alapok

A regressziós feladatokban feltesszük, hogy létezik egy $f^* : \mathcal{X} \to \mathbb{R}$ függvény és ennek az $\mathcal{X}$ mérhető halmaz különböző pontjaiban felvett értékét szeretnénk becsülni egy $\widehat{f} : \mathcal{X} \to \mathbb{R}$ függvénnyel a rendelkezésünkre álló $(x_1, y_1), \ldots, (x_n, y_n) \in \mathcal{X} \times \mathbb{R}$ adathalmaz segítségével. Ha feltesszük, hogy $y_i = f^*(x_i)$ és azt szeretnénk, hogy $\widehat{f}(x_i) = y_i$ legyen minden i-re, akkor ezt a feladatot interpolációnak nevezzük. Azonban ez egy elég erős feltevés, a gyakorlatban ez általában nincs is így, például lehetnek mérési pontatlanságok. Ez esetben úgy vesszük, hogy $y_i = f^*(x_i) + \varepsilon$, ahol ε egy $\mathbb{R}$ feletti (általában 0 várható értékű és véges szórású) eloszlásból származó zaj.

Tehát szeretnénk egy olyan $\widehat{f}$-et, ami a lehető legjobban közelíti a számunkra ismeretlen f^*-ot. Ez a feladat két részből áll: Először meghatározzuk azt a $\mathcal{F}$ függvényhalmazt, amin $\widehat{f}$-et keressük, majd kiválasztjuk ezen a halmazon a *legjobbat* egy tanulási algoritmus segítségével. Ebben a fejezetben definiáljuk a veszteségfüggvényeket, melyek segítségével az egyes függvények jóságát tudjuk majd mérni, majd bevezetjük a regularizációt, ami segítségével előzetes tudásunk alapján tudjuk torzítani a becslésünket. (Ez a torzítás gyakran annak felel meg, hogy az *egyszerűbb* függvényeket preferáljuk, legyenek például minél *simábbak* vagy *laposabbak*.) Ezután arra térünk ki, hogy hogyan kell megfelelő függvényosztályt választani. Ehhez a választott $\mathcal{F}$ függvényosztály kapacitását vizsgáljuk, vagyis azt, hogy mennyire pontosan lehet függvényt illeszteni $\mathcal{F}$ segítségével tetszőleges adatok esetén. Végül megnézünk egy egyszerűbb regressziós példát, a ridge regressziót $\mathbb{R}^d$ felett.

1.1. Veszteségfüggvények, kockázat és regularizáció

A következőekben definiált fogalmak nem csak a regressziós feladatokban használatosak, ezért általánosságban vezetjük be őket, tehát itt az (x, y) pontpárok tetszőleges $\mathcal{X}$ és $\mathcal{Y}$ mérhető halmazok elemei lehetnek.

Adott tehát egy $(x_1, y_1), \ldots, (x_n, y_n) \subset \mathcal{X} \times \mathcal{Y}$ adathalmaz, és rendelkezésünkre áll $\mathcal{F} \subset \mathcal{F}(\mathcal{X}, \mathcal{Y})$,

ahol $\mathcal{F}(\mathcal{X}, \mathcal{Y})$ az $\mathcal{X}$-ből $\mathcal{Y}$-ba képező függvények halmazát jelöli. Ezek közül szeretnénk kiválasztani azt, amelyik a *legjobb* y becslést adja tetszőleges $x \in \mathcal{X}$ esetén.

Tehát szükségünk van egy olyan számra, ami kifejezi, hogy az egyes $x \in \mathcal{X}$-ekre mekkorát téved egy $f \in \mathcal{F}$ függvény, hogy ezt minimalizálhassuk. Ezt fogja majd kifejezni egy nemnegatív veszteségfüggvény.

1.1.1. Definíció. [7, Definition 3.1]
Legyen $x \in \mathcal{X}$, $y \in \mathcal{Y}$ egy (tipikusan zajos) **minta**, $f(x) \in \mathcal{Y}$ egy **becslés** y-ra x alapján, $(x, y, f(x)) \in \mathcal{X} \times \mathcal{Y} \times \mathcal{Y}$ pedig az ezek által alkotott hármas. Ekkor $c : \mathcal{X} \times \mathcal{Y} \times \mathcal{Y} \to [0, \infty)$ egy **veszteségfüggvény**, ha $c(x, y, y) = 0$ minden $x \in \mathcal{X}$-re és $y \in \mathcal{Y}$-ra.

Megjegyzés. Fontos szempontok lehetnek a veszteségfüggvény választásakor, hogy könnyű legyen kiszámolni, illetve ha $\mathcal{X} \times \mathcal{Y} \times \mathcal{Y}$ egy vektortér, akkor legyen konvex és esetleg néhány pontot kivéve folytonosan differenciálható. (Hasznos eszközök állnak a rendelkezésünkre konvex optimalizálási feladatok megoldására, ezeket majd a negyedik fejezetben tárgyaljuk.)

Esetleg az is szempont lehet, hogy ne legyen túlságosan érzékeny egy-egy kiugró, esetlegesen hibás adatra, ezt a tulajdonságot **robosztusságnak** nevezik.

Fontos megjegyzés, hogy nincsen univerzálisan legjobb veszteségfüggvény, ennek a választása mindig az adott feladattól függ. Például attól, hogy mennyire tartjuk zajosnak az inputtot, mennyire vannak kiugró, esetlegesen hibás adataink, illetve érdemes a veszteségfüggvényt a megoldandó feladathoz igazítani. Például orvosi képklasszifikáció esetén jobban szeretnénk büntetni azt, hogy ha egy rákos sejtről készült felvételre azt mondjuk hogy egészséges, mint fordítva.

Fontos megjegyzés még, hogy miután definiáltuk a feladatunkhoz a megfelelő veszteségfüggvényt, a gyakorlatban nem feltétlen kell ragaszkodnunk ennek a használatához a minimalizáló függvény keresésekor, választhatunk valami könyebben kezelhetőt.

Regresszió esetén c általában valamilyen függvénye $\xi = f(x) - y$-nak, vagyis a becsült és a valós mérés értékeinek különbségének. Általában azt is szeretnénk, hogy a veszteség ne függjön a fenti kifejezés előjelétől, tehát valahogyan az abszolút értékét szeretnénk venni. (De mint ahogy azt a klasszifikációs példában is láttuk, ezen a valós életbeli feltételek változtathatnak)

Az egyik leggyakoribb veszteségfüggvény a **négyzetes veszteség**:

$$c(x, y, f(x)) = \frac{1}{2}(f(x) - y)^2 = \frac{1}{2}\xi^2$$

Ennek a regularizált változatát használja a fejezet végén szereplő ridge regresszió.

Egy másik veszteségfüggvény, amit a szupport vektor gépekhez fogunk használni az ε**-inszenzitív**

veszteség:

$$c(x, y, f(x)) = \max(|f(x) - y| - \varepsilon, 0) = |\xi|_\varepsilon$$

Ennek a veszteségfüggvénynek fontos tulajdonsága, hogy nem tesz különbséget az inszenzitív tartományon belül a becslések között.

Néhány bináris klasszifikációs esetben (ahol $\mathcal{Y} = \{-1, 1\}$) a 0-1 veszteségfüggvényt használunk majd:

$$x(x, y, f(x)) = \frac{1}{2}|f(x) - y|$$

Eddig a veszteségfüggvényt még csak egy (x, y) párra definiáltuk, de ez önmagában még nem túlságosan hasznos, mivel általában nem egy adatpontra, hanem egy n elemű $(x_1, y_1), \ldots, (x_n, y_n)$ mintára próbálunk modellt illeszteni.

1.1.2. Definíció. [7, Definition 3.3]
Ha feltesszük, hogy $\mathcal{X} \times \mathcal{Y}$ egy mérhető tér, ami felett van definiálva egy $\mathbb{P}_{x,y}$ eloszlás, és $c(\cdot, \cdot, f(\cdot))$ egy mérhető veszteségfüggvény, akkor vehetjük minimalizálandó értéknek a **kockázatot**, ami a veszteségfüggvény várható értéke:

$$R[f] \coloneqq E_{x,y}[c(x, y, f(x))] = \int_{\mathcal{X}\times\mathcal{Y}} c(x, y, f(x))\, d\mathbb{P}_{x,y}$$

A gyakorlatban azonban ezt nem tudjuk expliciten kiszámolni, hiszen ha tudnánk $\mathbb{P}_{x,y}$-t, akkor tetszőleges $x \in \mathcal{X}$-re ki lehetne számolni a marginális eloszlást, így készen lenne a feladat. Viszont ezt az eloszlást tudjuk közelíteni a tanító halmaz segítségével úgy, hogy a minta szerinti diszkrét eloszláson, azaz $\mathbb{P}^{\text{emp}}_{x,y} = \frac{1}{n}\sum_{i=1}^{n} \delta_{x_i}(x)\delta_{y_i}(y)$-on nézzük a veszteségfüggvény várható értékét (ahol $\delta_z(w) = 1$ ha $z = w$ és 0 egyébként).

1.1.3. Definíció. [7, Definition 3.4]
Így az **empirikus kockázat** már egy könnyen számolható érték:

$$R_{\text{emp}}[f] \coloneqq \int_{\mathcal{X}\times\mathcal{Y}} c(x, y, f(x))\, d\mathbb{P}^{\text{emp}}_{x,y} = \frac{1}{n}\sum_{i=1}^{n} c(x_i, y_i, f(x_i))$$

Megjegyzés. Definiálhatnánk úgy is a veszteségfüggvényt $(x_1, y_1), \ldots, (x_n, y_n) \subset \mathcal{X} \times \mathcal{Y}$ esetén, mint $L : (\mathcal{X} \times \mathcal{Y})^n \to [0, \infty)$ függvények. Ennek a leggyakrabban használt speciális esete az empirikus kockázat, ezért a továbbiakban veszteségfüggvény alatt az 1.1.1-ben definiált veszteségfüggvényt értjük, kivéve ahol ezt nem emeljük ki külön.

Az empirikus kockázattal óvatosnak kell lenni azonban, hiszen ha bármilyen függvényt használhatunk, akkor könnyen tudjuk úgy definiálni f-et, hogy az empirikus kockázata 0 legyen, például $f(x) = y_i$ ha $x = x_i$ és 0 egyébként. Ennek 0 lesz az empirikus kockázata, de minden még nem látott x-hez 0-t rendel, ami nem túl hasznos. Ekkor azt mondjuk, hogy nem történik érdemi tanulás, hiszen bármelyik másik olyan f' függvényt választhattuk volna, amire teljesül, hogy $f'(x_i) = y_i$,

mivel az is ugyanúgy nulla empirikus kockázattal rendelkezne. Így nem tudunk mondani semmit a még nem látott modellről.

Egy kicsit kevésbé szélsőséges eset az, amikor $\mathcal{X} = \mathcal{Y} = \mathbb{R}$ és egy n elemű mintára interpolálunk egy $n-1$-ed fokú polinomot. Ezt szemlélteti a következő oldalon található ábra.

Tehát mint láthatjuk, az alacsony empirikus kockázat nem feltétlen jelenti azt, hogy a kockázat is alacsony lenne.

Ezt a problémát kétféle képpen lehet megoldani: vagy a priori *elég szűknek* választjuk $\mathcal{F}$-et, ezzel a kapacitását csökkentve, vagy pedig megengedünk egy bővebb függvényosztályt, de valahogyan büntetjük a *komplexitást*. Ezt a két módszert együttesen **induktív torzításnak** nevezzük és gyakran egyszerre használjuk a kettőt: a függvényosztály leszűkítésével először kezelhető méretűvé tesszük a feladatot, majd regularizáció segítségével vonzóbbá teszük az egyszerűbb függvényeket a tanulási algoritmus számára. Minél szűkebb a függvényosztály, illetve minél nagyobb súlyt kap a regularizáció, annál erősebb az induktív torzítás.

1.1.4. Definíció. [7, 4.1 The Regularized Risk Functional]
A **regularizált kockázat** az empirikus kockázat és egy **regularizációs funkcionál** összege:

$$R_{\text{reg}}[f] \coloneqq R_{\text{emp}}[f] + \Omega[f]$$

Ahol a $\Omega : \mathcal{F}(\mathcal{X}, \mathcal{Y}) \to [0, \infty)$ regularizációs funkcionál segítségével tudjuk büntetni $f \in \mathcal{F}(\mathcal{X}, \mathcal{Y})$ komplexitását úgy, hogy hozzáadunk valami nemnegatív értéket a kockázathoz attól függően, hogy mennyire találjuk elfogadhatónak az adott függvényt.

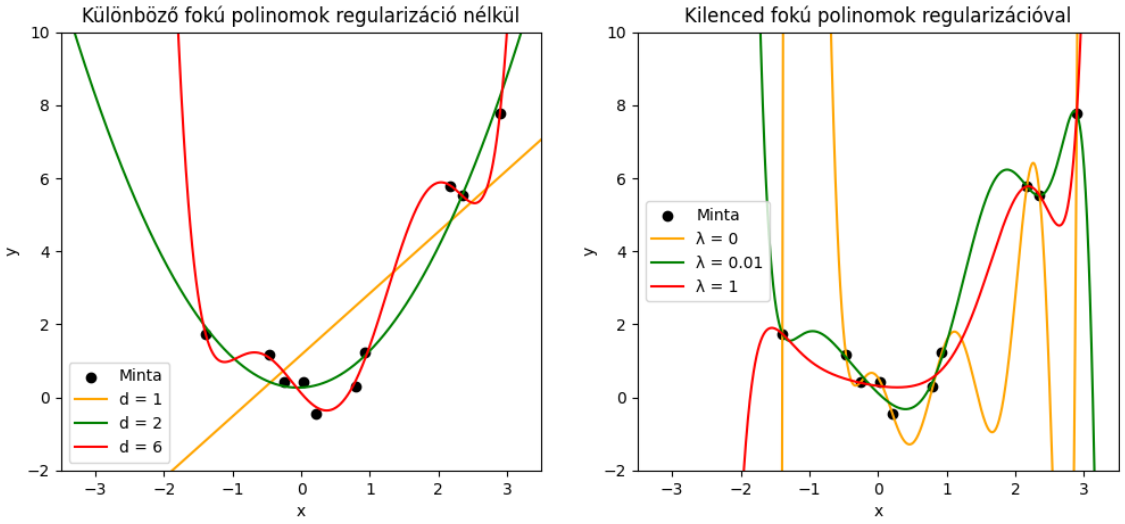
Megjegyzés. Előfordul, hogy azt tudjuk, hogy milyen fajta regulariációt szeretnénk használni, de azt nem hogy milyen súllyal szereplejen az empirikus kockázathoz képest a legjobb eredmény elérése érdekében. Ekkor használhatjuk a következő formát is:

$$R_{\text{reg}}[f] \coloneqq R_{\text{emp}}[f] + \lambda\Omega[f]$$

Ahol a $\lambda > 0$ egy hiperparaméter, ami segítségével ki lehet próbálgatni, hogy a különböző értékeire milyen eredményt kapunk.

Megjegyzés. Általában úgy vesszük, hogy a λ már része az Ω-nak.

1.1.1. Példa. Legyen $\mathcal{X} = [-3, 3], \mathcal{Y} = \mathbb{R}, \mathbb{P}_x \sim U[-3, 3]$ és $y = x^2 + \varepsilon$ ahol $\varepsilon \sim \mathcal{N}(0, 1)$ zaj. A következő ábra szemlélteti különböző fokú polinomok illesztését 10 elemű mintára négyzetes veszteség mellett regularizált, illetve regularizáció nélüli esetben.



Az első ábrán látható, hogy a megengedett polinomok fokának és ezzel együtt a függvényosztály kapacitásának a növelése túltanuláshoz vezethet, mivel ekkor jobban tud illeszkedni a zajra a függvényosztály.

A második ábrán látható a kilenced fokú polinommal való interpoláció (a $\lambda = 0$ eset), aminek nulla az empirikus kockázata, azonban az általa becsült értékek nem túl hihetőek. Viszont látható, hogy már egy kisebb súlyú regularizáció is sokat javít a becslésen.

1.2. Függvényosztályok kapacitása

Mint azt már az előzőekben is láttuk, fontos hogy milyen $\mathcal{F} \subset \mathcal{F}(\mathcal{X}, \mathcal{Y})$ függvényosztály felett próbáljuk minimalizálni az empirikus kockázatot. Arra is láttunk példát, hogy valamilyen jellegű induktív torzítás elengedhetetlen, ha egy olyan empirikus kockázatot minimalizáló algoritmust szeretnénk létrehozni, ami végez érdemi tanulást.

Azonban hogyan lehet eldönteni egy függvényhalmaról, hogy mennyire hajlamos egy felette értelmezett empirikus kockázat minimalizáló algoritmus a túltanulásra? Ennek a kérdésnek a megválaszolásához lehet vizsgálni a függvényosztály **kapacitását**, vagyis azt, hogy mennyire lehet belőle olyan függvényt választani, ami *elég jól* illeszkedik az empirikus adatra. Ennek a definíciónak a pontosítására van több különböző kapacitás koncepció is. Mi ezek közül most a VC dimenziót vizsgáljuk meg.

1.2.1. Definíció. [7, 5.3.3 The Shattering Coefficient]
Legyen $Z_n := ((x_1, y_1), \ldots, (x_n, y_n))$ egy n elemű minta $\mathcal{X} \times \{-1, 1\}$-ből, $\mathcal{F} \subset \mathcal{F}(\mathcal{X}, \{-1, 1\})$ pedig egy függvényosztály. Jelölje az $\mathcal{N}(\mathcal{F}, Z_n)$ azon függvények számát, amik megkülönböztethetőek egymástól az $x_1, \ldots, x_n$-en felvett értékeik segítségével. Illetve legyen a **zúzó együttható** $\mathcal{N}(\mathcal{F}, n)$ ezeknek a maximuma az összes lehetséges n elemű Z_n minták között.

Megjegyzés. $\mathcal{N}(\mathcal{F}, n)$ szemléletesen azt jelenti, hogy adott $x_1, \ldots, x_n$-hez hány különböző $y_1, \ldots, y_n$ címkézésre lehet olyan $f \in \mathcal{F}$-et találni, ami helyesen osztályozza az összeset. Mivel bináris klasszifikációról van szó, ezért $\mathcal{N}(\mathcal{F}, n) \leq 2^n$. Ha a $\mathcal{N}(\mathcal{F}, n) = 2^n$, akkor azt mondjuk, hogy $\mathcal{F}$ **szétzúz** n pontot.

1.2.2. Definíció. [7, 5.5.6 The VC Dimension and Other Capacity Concepts]
Egy $\mathcal{F}$ függvényosztály által szétzúzható pontok maximális száma a **VC dimenziója** vagyis az a $h \in \mathbb{N}$, amire $\mathcal{N}(\mathcal{F}, h) = 2^h$, de $\mathcal{N}(\mathcal{F}, h+1) < 2^{h+1}$.

Megjegyzés. A VC dimenzió lehet végtelen is.

Egy $\mathcal{F}$ függvényosztály VC dimenziójának segítségével már tudunk garanciát adni arra, hogy az empirikus kockázat minimalizálásával csökken a kockázat is.

1.2.3. Tétel. *[7, (1.19), (1.20)]*
Ha $(x_1, y_1), \ldots, (x_n, y_n) \in \mathcal{X} \times \{-1, 1\}$ egy i.i.d. minta a $\mathbb{P}_{x,y}$ eloszlásból és a tanulási algoritmus egy $h < n$ VC dimenziójú $\mathcal{F}$ függvénytéren minimalizálja az empirikus kockázatot, akkor $1 - \delta$ valószínűséggel teljesül, hogy

$$R[f] \leq R_{\text{emp}}[f] + \sqrt{\frac{1}{n}\left(h\left(\log\left(\frac{2n}{h}\right) + 1\right) + \log\left(\frac{4}{\delta}\right)\right)}$$

Itt az $1 - \delta$ valószínűség abból adódik, hogy a $(x_1, y_1), \ldots, (x_n, y_n)$ minta véletlenszerűen generálódik az ismeretlen eloszlásból. Figyeljük meg, hogy a VC dimenzió, vagy a becslés helyességének valószínűségének a növelése növeli a hibatagot, míg a minta elemszámának növelése csökkenti azt, mint ahogyan azt el is várnánk.

1.2.4. Tétel. *[9, Theorem 1.]*
Ha $\mathcal{F}$ egy altere $\mathcal{F}(\mathbb{R}^d, \mathbb{R})$-nek és $\mathcal{F}' = \{\text{sgn}(f) | f \in \mathcal{F}\}$, akkor $\mathcal{F}'$ VC dimenziója megegyezik $\mathcal{F}$ dimenziójával.

1.2.5. Következmény. *[7, 5.5.6 The VC Dimension and Other Capacity Concepts]*
Ha $\mathcal{X} = \mathbb{R}^d$ és $\mathcal{F} = \left\{\text{sgn}(\langle w, \cdot \rangle + b) : w \in \mathbb{R}^d, b \in \mathbb{R}\right\}$, vagyis a hipersík klasszifikátorok. Ekkor $\mathcal{F}$ VC dimenziója $d + 1$

Az ötödik fejezetben szereplő szupport vektor gépek ilyen hipersíkokat fognak majd használni.

1.3. Ridge regresszió

Most nézzünk egy egyszerűbb példát regularizált kockázatot minimalizáló regressziós algoritmusra, ami egyenlőre még csak lineáris összefüggéseket tud majd felfedezni, azonban a későbbi eszközeink használatával ennél már sokkal többre lesz képes.

Legyen $\mathcal{X} = \mathbb{R}^d, \mathcal{Y} = \mathbb{R}$ és adott egy $(x_1, y_1), \dots, (x_n, y_n) \subset \mathbb{R}^d \times \mathbb{R}$ minta a $\mathbb{P}_{x,y}$ eloszlásból.

Legyen a veszteségfüggvény $c(x, y, f(x)) = (y - f(x))^2$, illetve a függvényhalmaz amin a legjobb becslést keressük $\mathcal{F} := \left\{ \langle \vartheta, \cdot \rangle \,\middle|\, \vartheta \in \mathbb{R}^d \right\}$, vagyis az $\mathcal{X}$-ből $\mathcal{Y}$-ba képező lineáris függvények halmaza. A regularizációs tag pedig legyen $\Omega[f] = \lambda \|f\|^2 = \lambda \langle \vartheta, \vartheta \rangle$ ahol $\lambda > 0$. Ekkor a **ridge regressziós** feladat a következőképpen írható fel:

$$\operatorname*{argmin}_{f \in \mathcal{F}} R_{\text{emp}}[f] + \Omega[f] = \frac{1}{n} \sum_{i=1}^{n} (f(x_i) - y_i)^2 + \lambda \|f\|^2$$

Ha felhasználjuk azt, hogy $f(x) = \langle \vartheta, x \rangle$, akkor a fenti kifejezést átírhatjuk a ϑ paraméterek használatával:

$$\operatorname*{argmin}_{\vartheta \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^{n} (\langle \vartheta, x_i \rangle - y_i)^2 + \lambda \langle \vartheta, \vartheta \rangle$$

ami szigorúan konvex ϑ-ban (4.1.9), ezért a minimumot ott fogja felvenni, ahol a kifejezés deriváltja 0. ϑ szerint deriválva a következő egyenletet kell majd megoldanunk:

$$\frac{1}{n} \sum_{i=1}^{n} (2 \left\langle \widehat{\vartheta}, x_i \right\rangle x_i - 2 y_i x_i) + 2\lambda \widehat{\vartheta} = 0$$

Ahol $\widehat{\vartheta}$ jelöli az optimális becslést ϑ-ra a minta alapján. A fenti képletet átrendezve kapunk egy új egyenletet $\widehat{\vartheta}$-ra:

$$(*)\ n\lambda \widehat{\vartheta} = \sum_{i=1}^{n} (y_i - \left\langle \widehat{\vartheta}, x_i \right\rangle) x_i = \sum_{i=1}^{n} y_i x_i - \sum_{i=1}^{n} \left\langle \widehat{\vartheta}, x_i \right\rangle x_i$$

Ahoz, hogy a skaláris szorzásban szereplő $\widehat{\vartheta}$-ot is eltüntessük az egyenlet jobb oldaláról, definiáljuk a következő mátrixokat és vektorokat: $X \in \mathbb{R}^{n \times d}$ i-edik sora tartalmazza x_i-t, illetve $y \in \mathbb{R}^n$ i-edik eleme y_i. Meggondolható, hogy ekkor fenti egyenlet felírható X és y segítségével:

$$n\lambda \widehat{\vartheta} = X^{\mathrm{T}} y - X^{\mathrm{T}} X \widehat{\vartheta}$$

Ezt már könnyen átrendezhetjük $\widehat{\vartheta}$-ra az I identitásmátrix segítségével:

$$(n\lambda I + X^{\mathrm{T}} X) \widehat{\vartheta} = X^{\mathrm{T}} y$$

Mivel $\lambda > 0$, ezért λI egy pozitív definit, illetve $(X^{\mathrm{T}} X)$ egy pozitív szemidefinit mátrix, ezért az összegük is pozitív definit, tehát invertálható:

$$\widehat{\vartheta} = (n\lambda I + X^{\mathrm{T}} X)^{-1} X^{\mathrm{T}} y$$

A kapott kifejezést még tovább lehet alakítani egy lineáris algebrai azonosság segítségével:

1.3.1. Lemma. *[1, Appendix (C.5)]*
Ha $P \in \mathbb{R}^{n \times n}, B \in \mathbb{R}^{m \times n}, R \in \mathbb{R}^{n \times n}$, akkor

$$(P^{-1} + B^{\mathrm{T}} R^{-1} B)^{-1} B^{\mathrm{T}} R^{-1} = P B^{\mathrm{T}} (B P B^{\mathrm{T}} + R)^{-1}$$

Ha $P = \frac{1}{n\lambda}I$, $B = X$, $R = I$-et helyettesítünk a képletbe, akkor azt kapjuk, hogy

$$(n\lambda I + X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}} = \frac{1}{n\lambda}IX^{\mathrm{T}}(X\frac{1}{n\lambda}IX^{\mathrm{T}} + I)^{-1} = X^{\mathrm{T}}(XX^{\mathrm{T}} + n\lambda I)^{-1}$$

Így $\widehat{\vartheta}$ egy másik alakjához jutunk, ami hasznosabb lesz a továbbiakban:

$$\widehat{\vartheta} = X^{\mathrm{T}}(XX^{\mathrm{T}} + n\lambda I)^{-1}y$$

Tehát az optimális ϑ paramétereket ki lehet számolni mátrixszorzás és egy inverálás segítségével. Ha visszatérünk egy korábbi, $(*)$-al jelölt kifejezéshez, akkor az optimális $\widehat{f}$ definió szerint

$$\widehat{f}(x) = \left\langle \widehat{\vartheta}, x \right\rangle = \left\langle \frac{1}{n\lambda}\sum_{i=1}^{n}(y_i - \left\langle \widehat{\vartheta}, x_i \right\rangle)x_i, x \right\rangle = \sum_{i=1}^{n}\frac{1}{n\lambda}(y_i - \left\langle \widehat{\vartheta}, x_i \right\rangle)\left\langle x_i, x \right\rangle = \sum_{i=1}^{n}\alpha_i \left\langle x_i, x \right\rangle$$

Ahol a második egyenlőtlenség a skaláris szorzás linearitásából következik, α_i pedig egy, a $y_i, x_i, \widehat{\vartheta}$ segítségével számolható skalár. Innen látszódik, hogy $\widehat{f}(x)$ az előáll, mint a $\langle x_i, x\rangle$-ek lineáris kombinációja, csak a megfelelő súlyokat kell tudni megválasztani és így telejesen elkerülhető a ϑ paraméter használata. Ezt a megfigyelést általánosítja majd az úgynevezett **Reprezentációs Tétel** (3.4.1). Tehát most szeretnénk egy olyan módszert, ami megadja nekünk az α_i súlyokat anélkül, hogy használnánk a ϑ paramétereket.

Legyen $\alpha \in \mathbb{R}^n$ az a vektor, ami tartalmazza az α_i-ket, ekkor:

$$\alpha = \frac{1}{n\lambda}(y - X\widehat{\vartheta}) = \frac{1}{n\lambda}(y - XX^{\mathrm{T}}(XX^{\mathrm{T}} + n\lambda I)^{-1}y)$$

Megfigyelhetjük, hogy ha $K = XX^{\mathrm{T}}$, akkor $K_{ij} = \langle x_i, x_j\rangle$. Tehát az α-t meg lehet kapni pusztán az x_i-k egymással vett skaláris szorzatának ismeretében:

$$\alpha = \frac{1}{n\lambda}(y - K(K + n\lambda I)^{-1}y)$$

Ez a tény teszi majd lehetővé, hogy a *kernel trükk* segítségével egy sokkal szélesebb függvényosztályon keressünk legjobb becslést ugyanennek az algoritmusnak a használatával, ahogyan azt a következő fejezetben látni fogjuk.

Megjegyzés. A ridge regressziós feladatot szokás még felírni a fent definiált X és y segítségével is:

$$\underset{\vartheta\in\mathbb{R}^d}{\operatorname{argmin}} \frac{1}{n}\left(\vartheta^{\mathrm{T}}X^{\mathrm{T}}X\vartheta - 2y^{\mathrm{T}}X\vartheta + y^{\mathrm{T}}y\right) + \lambda\vartheta^{\mathrm{T}}\vartheta$$

Megjegyzés. Ha elhagyjuk a regularizációs tagot, vagyis $\lambda = 0$, akkor megkapjuk az egyszerű lineáris regressziós feladatot:

$$\underset{\vartheta\in\mathbb{R}^d}{\operatorname{argmin}} \frac{1}{n}(X\vartheta - y)^2 = \frac{1}{n}\left(\vartheta^{\mathrm{T}}X^{\mathrm{T}}X\vartheta - 2y^{\mathrm{T}}X\vartheta + y^{\mathrm{T}}y\right)$$

Az ehhez tartozó α együtthatók is megkaphatóak, de itt mivel $K = XX^{\mathrm{T}}$ nem feltétlen invertálható,

ezért pszeudo-inverzet kell használni:

$$\widehat{f}(x) = \left\langle \widehat{\vartheta}, x \right\rangle = \left\langle X^{\mathrm{T}}(XX^{\mathrm{T}})^{-1}y, x \right\rangle = (X^{\mathrm{T}}(XX^{\mathrm{T}})^{-1}y)^{\mathrm{T}}x = y^{\mathrm{T}}K^{-1}Xx$$

Ahol kihasználtunk, hogy K és így K^{-1} is szimmetrikusak. Tehát $\alpha = y^{\mathrm{T}}K^{-1}$.

2. fejezet

Kernelek

2.1. Polinom kernelek

Az előző fejezetben láttuk, hogy a függvényosztály választása alapvető fontosságú feladat, azonban eddig csak az $\mathbb{R}^d$ feletti lineáris függvények halmazát vizsgáltuk. Hogyan tudnánk kezelni a négyzetes veszteségfüggvény mellett másik, bővebb függvényosztályokat? Az 1.1.1 példában például különböző fokú polinomokra alkalmaztunk négyzetes veszteséget. Felmerülhet a kérdés, hogy hogyan számoltuk ki ezeket a függvényeket. Elsőre talán meglepő lehet, de ugyanazt a ridge és lineáris regressziót használtuk, mint amit a fejezet végén is néztünk.

Adott tehát egy $(x_1, y_1), \dots, (x_n, y_n) \in \mathbb{R}^2$ i.i.d. minta egy $\mathbb{P}_{x,y}$ eloszlásból, és szeretnénk megtalálni azt a d-ed fokú p polinomot, amire $\sum_{i=1}^n (p(x_i) - y_i)^2$ a lehető legkisebb. (Egyenlőre tekintsünk el a regularizációs tagtól.) Könnyen látható, hogy ez egy $d+1$ paraméterű minimalizálási feladat:

$$\operatorname*{argmin}_{a \in \mathbb{R}^{d+1}} \sum_{i=1}^{n} (a_0 + a_1 x_i + a_2 x_i^2 + \dots + a_d x_i^d - y_i)^2$$

Erről viszont már látható, hogy ha az x_i-k helyett az $\Phi(x_i) = [1, x_i, x_i^2, \dots, x_i^d]^{\mathrm{T}}$ vektorokat nézzük, akkor ez egy lineáris regressziós feladat. Tehát egy ügyes Φ leképezéssel visszavezettük lineáris függvényekre a feladatot. Azt a Hilbert-teret, ahol már lineáris a feladat (vagyis ahova a Φ képez) nevezzük **tulajdonságtérnek**.

Ez a megközelítés kisebb dimenziók esetén még működik, azonban ha az eredeti tér $\mathbb{R}^n$ és d dimenziós polinomokat szeretnénk használni, akkor a tulajdonságtér dimenziója $\binom{d+n}{d}$, ami nagyobb n és d esetén már kezelhetetlen lenne. Azonban ha egy olyan algoritmust használunk, ami a tulajdonságtér elemein csak skaláris szorzás műveletet végez és van egy explicit képletünk $k(x, x') = \langle \Phi(x), \Phi(x') \rangle$-re, akkor nem kell kiszámolni a Φ leképezést a mintaelemekre, hiszen ezeknek úgyis csak a skaláris szorzatát vennénk később, és pont ezt a két műveletet vonja össze $k(x, x')$. Ezt a kétváltozós k függvényt nevezzük majd **kernelnek**. Tehát ha lenne egy szép képletünk az n változós, d dimen-

ziós polinomok skaláris szorzatára, akkor ki lehetne kerülni azt a problémát, amit a tulajdonságtér dimenziójának exponenciális növekedése okoz.

2.1.1. Állítás. [7, *Proposition 2.1]*
Legyen $x, x' \in \mathbb{R}^n$, $C_d(x)$ pedig az a vektor, aminek a koordinátái az x koordinátáinak pontosan d elemű rendezett szorzatai. (például ha $x = [x_1, x_2]^{\mathrm{T}}$, akkor $C_2(x) = [x_1^2, x_2^2, x_1x_2, x_2x_1]^{\mathrm{T}}$)

Ekkor $k(x, x') \coloneqq \langle C_d(x), C_d(x')\rangle = \langle x, x'\rangle^d$

Bizonyítás.

$$\begin{aligned}\langle C_d(x), C_d(x')\rangle &= \sum_{j_1=1}^{n}\sum_{j_2=1}^{n}\cdots\sum_{j_d=1}^{n} x_{j_1}x'_{j_1}\, x_{j_2}x'_{j_2}\ldots x_{j_d}x'_{j_d}\\ &= \sum_{j_1=1}^{n}(x_{j_1}x'_{j_1})\sum_{j_2=1}^{n}(x_{j_2}x'_{j_2})\cdots\sum_{j_d=1}^{n}(x_{j_d}x'_{j_d})\\ &= (\sum_{j=1}^{n} x_jx'_j)^d = \langle x, x'\rangle^d\end{aligned}$$

□

Itt az első egyenlőség felírásánál használtuk azt, hogy a szorzatban számít a sorrend, így egy n^d dimenziós térhez jutva. Azonban ha azt szeretnénk, hogy a tulajdonságtérben minden d-ed fokú monom csak egyszer szereplejen, akkor is használható ugyanez a kernel, azonban az ebbe a térbe képező Φ_d leképezést módosítani kell egy kicsit. Mégpedig úgy, hogy ha a $P(x_1, \ldots, x_n)$ polinom N-szer szerepelt a C_d képterében, akkor itt kap egy $\sqrt{N}$ együtthatót. Például ha $x = [x_1, x_2]^{\mathrm{T}}$, akkor $\Phi_2(x) = [x_1^2, x_2^2, \sqrt{2}x_1x_2]$. Meggondolható, hogy így pontosan ugyanahoz a kifejezéshez fogunk jutni mint az előbb, viszont most már a megfelelő vektortér felett.

2.1.2. Definíció. Ha $x, x' \in \mathbb{R}^n$, akkor $k(x, x') = \langle x, x'\rangle^d$ a d-ed fokú **monom kernel**.

A fenti eset nagyon jól szemlélteti, hogy egy kernelhez több tulajdonságtér és Φ leképezés is tartozhat.

Ha azt szeretnénk hogy a tulajdonságtérben a legfeljebb d-ed fokú polinomok legyenek, akkor az is kijön hasonló számolással:

2.1.3. Definíció. [8, Example 16.1]
$k(x, x') = (1 + \langle x, x'\rangle)^d$ a d-ed fokú **polinom kernel**.

2.2. Pozitív definit kernelek

Most, hogy láttunk egy bevezető példát a kernelekre, definiáljuk őket rendesebben, illetve vizsgáljuk meg őket általánosabban.

Legyen $\mathcal{X}$ egy tetszőlges halmaz. Ekkor **kernelnek** nevezünk egy $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ függvényt, ami két $\mathcal{X}$-beli ponthoz egy, a *hasonlóságukat* kifejező számot rendel.

Általában az olyan kétváltozós függvényeket fogadjuk el kerneleknek, amiknek van gyakorlati hasznuk. Azonban ehhez nem szükséges, hogy kernel függvények is legyenek, ami viszont már egy szűkebb, de cserébe jól kezelhető függvényosztály.

2.2.1. Definíció. [7, Kernels (2.2)]
$k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ egy **kernel függvény** $\mathcal{X}$ felett, ha létezik egy olyan $\mathcal{H}$ Hilbert-tér és egy $\Phi : \mathcal{X} \to \mathcal{H}$ leképezés, hogy $k(x, x') = \langle \Phi(x), \Phi(x') \rangle_{\mathcal{H}}$.
Ekkor ezt a Φ-t **tulajdonság leképezésnek**, a $\mathcal{H}$-t pedig **tulajdonságtérnek** nevezzük.

Tehát egy kernelt kétféle képpen lehet megkonstruálni. Az egyik lehetőség az, hogy úgy, mint polinom kerneleknél, először azt találjuk ki, hogy milyen Φ leképezést szeretnénk használni, majd ennek a segítségével próbálunk egy minél egyszerűbben számolható képletet alkotni k-ra.

A másik megközelítés az, hogy a k függvényt ismerjük és azt szeretnénk belátni, hogy létezik egy olyan $\mathcal{H}$ Hilbert-tér és Φ leképezés, amire $k(x, x') = \langle \Phi(x), \Phi(x') \rangle_{\mathcal{H}}$. Ekkor elegendő a létezés bizonyítása, nem szükséges a leképezést és a teret meghatározni. A következőkben nézünk egy szükséges és elégséges feltételt arra, hogy k egy kernel függvény legyen.

2.2.2. Definíció. [8, C.3 Positive Definite Matrices]
$A \in \mathbb{R}^{n \times n}$ mátrix **pozitív szemidefinit** ha $\langle Ax, x \rangle \geq 0 \ \forall x \in \mathbb{R}^n$ Ha ez szigorú egyenlőtlenséggel teljesül minden nemnulla x esetén, akkor azt mondjuk, hogy A **pozitív definit**.

2.2.3. Definíció. [7, Definition 2.3, 2.4, 2.5]
Egy $\mathcal{X}$ halmazon értelmezett $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ kernel **Gram mátrixa** $x_1, \ldots, x_n \in \mathcal{X}$-en az a K mátrix, melynek elemei $K_{ij} \coloneqq k(x_i, x_j)$.

A k kernel **pozitív definit** ha a Gram mátrixa minden $x_1, \ldots, x_n \in \mathcal{X}$-re pozitív szemidefinit.

Hasonlóan, k **szigorúan pozitív definit**, ha a Gram mátrixa minden $x_1, \ldots, x_n \in \mathcal{X}$-re pozitív definit.

2.2.4. Állítás. *[8, Lemma 16.2]*
Egy $\mathcal{X}$ feletti k kernel pontosan akkor kernel függvény, ha pozitív definit.

Bizonyítás. Ha $k = \langle \Phi(x), \Phi(x') \rangle$ egy kernel függvény és K jelöli a Gram mátrixát tetszőleges $x_1, \ldots, x_n \in \mathcal{X}$-re, akkor minden $\alpha \in \mathbb{R}^n$-re:

$$\langle K\alpha, \alpha \rangle = \sum_{i,j=1}^{n} \alpha_i \alpha_j k(x_i, x_j) = \sum_{i,j=1}^{n} \alpha_i \alpha_j \langle \Phi(x_i), \Phi(x_j) \rangle = \sum_{i,j=1}^{n} \langle \alpha_i \Phi(x_i), \alpha_j \Phi(x_j) \rangle = \sum_{i=1}^{n} \| \alpha_i \Phi(x_i) \| \geq 0$$

Tehát pozitív definit.

A másik irányt a következő fejezetben látjuk majd be a 3.3.4 Moore-Aronszajn tétel segítségével, miután bevezettük a reprodukáló magú Hilbert tereket. Minden pozitív definit kernelhez fog egyértelműen tartozni egy ilyen és ez lesz majd egy lehetséges tulajdonságtér. □

Megjegyzés. A kernel függvényekre lehet úgy gondolni, mint egy általánosított skaláris szorzásra. Például könnyen látható, hogy a kernel függvények is szimmetrikusak:

$$k(x, x') = \langle \Phi(x), \Phi(x') \rangle = \langle \Phi(x'), \Phi(x) \rangle = K(x', x)$$

Azonban az első változóban való linearitás nyilván nem teljesül (vagy csak néhány esetben), mert $\mathcal{X}$-től még csak azt sem követeljük meg, hogy vektortér legyen. De vannak további olyan tulajdonságok is, amiket a skaláris szorzás esetén már jól ismerünk, ilyen például a CSB-egyenlőtlenség:

2.2.5. Állítás (Cauchy-Schwarz-Bunyakovszkij-egyenlőtlenség). *[3, 2.3. Állítás]*
$\mathcal{H}$ egy Hilbert-tér, $x, y \in \mathcal{H}$. Ekkor $|\langle x, y \rangle_{\mathcal{H}}| \leq \|x\|_{\mathcal{H}} \|y\|_{\mathcal{H}}$

2.2.6. Állítás (CSB egyenlőtlenség pozitív definit kernelekre). *[7, Proposition 2.7]*
Ha k egy pozitív definit kernel $\mathcal{X}$ felett, $x, y \in \mathcal{X}$, akkor $|k(x_1, x_2)|^2 \leq k(x_1, x_1)k(x_2, x_2)$.

Bizonyítás. Mivel a $K_{ij} = k(x_i, x_j), i, j \in \{1, 2\}$ Gram mátrix pozitív szemidefinit, ezért a determinánsa nemnegatív, azaz

$$0 \leq K_{11}K_{22} - K_{12}K_{21} = K_{11}K_{22} - K_{12}K_{12} = K_{11}K_{22} - |K_{12}|^2$$

Mindkét oldalból $|K_{12}|^2$-et kivonva megkapjuk a keresett egyenlőtlenséget. □

2.3. A kernel függvények alkalmazása

A 2.1 polinomjaihoz hasonlóan bármilyen másik kernelt használhatunk arra, hogy lineáris modellek segítségével nemlineáris összefüggéselet keressünk.

2.3.1. Állítás (A Kernel Trükk). *[7, Remark 2.8]*
Ha van egy algoritmus, ami az $x_1, \ldots, x_n \in \mathcal{X}$ adattal dolgozik és megfogalmazható k_1 kernel függvény használatával, akkor ha a k_1-et lecseréljük egy k_2 kernel függvényre, akkor egy értelmes algoritmust kapunk.

A kernel trükk indokolható azzal, hogy az eredeti algoritmus olyan, mintha a $\Phi_1(x_1), \ldots, \Phi_1(x_n)$ vektorokkal dolgozna egy Hilbert-téren és itt csak a skaláris szorzást használná. A második algoritmus pedig pontosan ugyanezt csinálja, csak a $\Phi_2(x_1), \ldots, \Phi_2(x_n)$ Hilbert-téren.

A kernel trükköt leggyakrabban olyan esetekben szokták alkalmazni, ahol a k_1 egy skaláris szorzás az $\mathcal{X}$ felett. Mi is ezt a megközelítést fogjuk alkalmazni: először belátjuk $\mathbb{R}^n$ felett az algoritmusok

helyességét, majd a kernel trükköt alkalmazva sokkal erősebb modellekhez jutunk. Ezt az eljárást nevezik kernelizálásnak.

A kernel függvények segítségével nem csak lecsökkenthetjük a számítási igényét egy algoritmusnak, hanem olyan tulajdonságtereket is tudunk használni, amikbe a leképezést nem lehetne expliciten kiszámolni, például azért, mert végtelen dimenziósak.

2.3.1. Példa (Gauss RBF kernel). [8, Example 16.2]
Legyen $x, x' \in \mathbb{R}$, $\sigma > 0$ paraméter. Jelölje $\Phi_n(x)$ a $\Phi(x)$ n-edik koordinátáját.

Ha $\Phi_n(x) = \frac{1}{\sqrt{n!}} \exp\left(-\frac{x^2}{2\sigma^2}\right)(\frac{x}{\sigma})^n$, akkor a **Gauss RBF kernel**

$$k(x, x') = \langle \Phi(x), \Phi(x') \rangle = \exp\left(-\frac{\|x - x'\|^2}{2\sigma^2}\right)$$

Bizonyítás.

$$\begin{aligned}
\langle \Phi(x), \Phi(x') \rangle &= \sum_{n=0}^{\infty} \left(\frac{1}{\sqrt{n!}} \exp\left(-\frac{x^2}{2\sigma^2}\right)\left(\frac{x}{\sigma}\right)^n\right)\left(\frac{1}{\sqrt{n!}} \exp\left(-\frac{x'^2}{2\sigma^2}\right)\left(\frac{x'}{\sigma}\right)^n\right) \\
&= \exp\left(-\frac{x^2 + x'^2}{2\sigma^2}\right) \sum_{n=0}^{\infty} \frac{1}{n!}\left(\frac{xx'}{\sigma^2}\right)^n = \exp\left(-\frac{x^2 + x'^2}{2\sigma^2}\right) \exp\left(\frac{xx'}{\sigma^2}\right) \\
&= \exp\left(-\frac{\|x - x'\|^2}{2\sigma^2}\right)
\end{aligned}$$

□

Megjegyzés. Azért használtuk az előbb a $\|\cdot\|$ jelet a $|\cdot|$ abszolútérték jel helyett, mert a Gauss kernelt ugyanígy lehet definiálni $x, x' \in \mathbb{R}^n$-re is, ugyanezzel a kernel függvénnyel. [7, (2.68)]
Fontos tulajdonsága még a Gauss kernelnek, hogy szigorúan pozitív definit. [7, Theorem 2.18]

3. fejezet

Reprodukáló kernel Hilbert terek

A reprodukáló magú Hilbert terek először a funkcionálanalízisben lettek bevezetve, azonban később a matematika többi területén is teret nyertek, például a statisztikában, vagy a komplex függvénytanban.

Számunkra azért lesznek érdekesek, mert ezekben a függvényterekben fogjuk a kockázatot minimalizáló becslést keresni ha kernelizáljuk az algoritmusunkat egy kernel függvénnyel.

A következőek teljesülnek $\mathbb{C}$ illetve $\mathbb{R}$ feletti Hilbert-terek esetén is, ezért jelölje $\mathbb{F}$ tetszőlegesen $\mathbb{R}$-et vagy $\mathbb{C}$-t, így egyszerre tudjuk kezelni a két esetet.

3.1. RKHS definíció

Adott egy $\mathcal{X}$ nemüres halmaz. Jelölje $\mathcal{F}(\mathcal{X}, \mathbb{F})$ az $\mathcal{X}$-ből $\mathbb{F}$-be képező függvények $\mathbb{F}$ feletti vektorterét.

3.1.1. Definíció. [3, 1.39. Definíció]
Ha $\mathcal{F}$ egy normált tér, akkor a $\Phi : \mathcal{F} \to \mathbb{F}$ funkcionál **korlátos**, ha $\exists M \geq 0 : |\Phi(f)| \leq M\|f\| \ \forall f \in \mathcal{F}$

3.1.2. Definíció. [6, Definition 1.1]
$\mathcal{H} \subset \mathcal{F}(\mathcal{X}, \mathbb{F})$ egy **reprodukáló magú Hilbert tér** (a továbbiakban RKHS) $\mathcal{X}$ felett, ha az alábbi három feltétel teljesül rá:
(i) $\mathcal{H}$ egy vektortér
(ii) tartozik $\mathcal{H}$-hoz egy olyan $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ skaláris szorzás, mellyel Hilbert teret alkot
(iii) minden $x \in \mathcal{X}$-re az $E_x : \mathcal{H} \to \mathbb{F}$ kiértékelő funkcionál, azaz $E_x(f) \coloneqq f(x)$ korlátos.

$\mathbb{F} = \mathbb{R}$ esetén azt mondjuk, hogy $\mathcal{H}$ egy **valós RKHS** $\mathcal{X}$ felett.

Megjegyzés. A kiértékelő funkcionál lineáris:
$E_x(\lambda_1 f_1 + \lambda_2 f_2) = (\lambda_1 f_1 + \lambda_2 f_2)(x) = \lambda_1 f_1(x) + \lambda_2 f_2(x) = \lambda_1 E_x(f_1) + \lambda_2 E_x(f_2)$

3.1.3. Tétel (Riesz reprezentációs tétele). *[3, 5.2. Tétel] Adott $\mathcal{H}$ Hilbert téren minden $\Phi : \mathcal{H} \to \mathbb{F}$ korlátos lineáris funkcionálhoz pontosan egy $y \in \mathcal{H}$ létezik úgy, hogy minden $x \in \mathcal{H}$-ra $\Phi(x) = \langle x, y \rangle$*

3.1.4. Következmény. *Ha $\mathcal{H}$ egy RKHS $\mathcal{X}$ felett, akkor minden $x \in \mathcal{X}$-hez létezik pontosan egy olyan $k_x \in \mathcal{H}$, hogy $f(x) = E_x(f) = \langle f, k_x \rangle$ minden $f \in \mathcal{H}$-ra.*

3.1.5. Definíció. [6, Definition 1.2]
Az előbbi k_x az x pont **reprodukáló kernele**.

3.1.6. Definíció. [6, Definition 1.2]
Az a $k : \mathcal{X} \times \mathcal{X} \to \mathbb{F}$ függvény, ami a $k(x, y) = k_y(x) = \langle k_y, k_x \rangle$ hozzárendelési szabállyal rendelkezik, a $\mathcal{H}$ RKHS **reprodukáló kernele**.

Megjegyzés. Az előbbi definíciókból következik, hogy RKHS-ek esetében minden $f \in \mathcal{H}$-ra

$$f(x) = \langle f, k_x \rangle = \langle f, k(\cdot, x) \rangle$$

Ezt nevezzük **reprodukáló tulajdonságnak**.

3.1.7. Definíció. [3, 2.18. Definíció]
Egy $S \subset \mathcal{H}$ rendszer **teljes**, ha teljesül az, hogy $\langle f, s \rangle = 0 \ \forall s \in S$ pontosan akkor ha $f = 0$.

3.1.8. Definíció. Egy $S \subset \mathcal{H}$ **sűrű** $\mathcal{H}$-ban, ha $\overline{S} = \mathcal{H}$, ahol $\overline{S}$ az S lezártját jelöli.

3.1.9. Állítás. *[6, Proposition 2.1] Az $S = \{k_x | x \in \mathcal{X}\}$ rendszer teljes.*

Bizonyítás. Ha $f \in \mathcal{H}$ ortogonális minden k_x-re, akkor az azt jelenti, hogy $\langle f, k_x \rangle = f(x) = 0$ minden x-re, tehát f a konstans 0 függvény. □

3.1.10. Következmény. *[3, 2.20. Állítás] Jelölje $[S]$ az S lineáris burkát. Az előző tétel alapján $[S]^\perp = \{0\}$, tehát $\overline{[S]}^\perp = \{0\}$. Riesz felbontási tétele szerint $\mathcal{H} = \overline{[S]} \times \overline{[S]}^\perp$, tehát $\mathcal{H} = \overline{[S]}$. Tehát S lineáris burka sűrű $\mathcal{H}$-ban.*

3.1.1. Példa. [6, 1.2.1 $\mathbb{C}^n$ as RKHS]
Legyen $\mathcal{X}$ egy tetszőleges megszámlálható halmaz, $\mathcal{H} := \ell^2(\mathcal{X}) = \left\{ f : \mathcal{X} \to \mathbb{F} \middle| \sum_{x \in \mathcal{X}} |f(x)|^2 < \infty \right\}$
Hilbert tér az $\langle f, g \rangle := \sum_{x \in \mathcal{X}} f(x)\overline{g(x)}$ skaláris szorzással. Ekkor a kiértékelő funkcionál $E_x(f)$ korlátos minden x-re, hiszen $|E_x(f)|^2 = |f(x)|^2 \leq \sum_{x \in \mathcal{X}} |f(x)|^2 = \|f\|^2$.

Könnyen ellenőrizhető, hogy $k_x = \mathbb{I}_x$ (x indikátorfüggvénye) minden x-re, hiszen ekkor $\langle f, k_x \rangle = \sum_{x' \in \mathcal{X}} f(x')\overline{k_x(x')} = f(x)$. $\mathcal{H}$ reprodukáló kernele pedig $k(x, y) = k_y(x) = \mathbb{I}_{x=y}$.

3.1.11. Következmény. *$\mathbb{F}^d$ RKHS minden $d < \infty$ esetén. Ekkor $\mathcal{X} = \{1, \dots, d\}$ és az $f \in \mathcal{F}(\mathcal{X}, \mathbb{F})$ függvények felelnek meg az $\mathbb{F}^d$-beli vektoroknak, ahol $f(x)$ az f x-edik koordinátája.*

3.1.2. Példa. [6, 1.2.2 A non-example]
$L^2[0,1]$ nem RKHS, mert a kiértékelő funkcionál nem korlátos.
Ha választunk egy tetszőleges $0 < x < 1$-et, és

$$f_n^x(t) = \begin{cases} (\frac{t}{x})^n & \text{ha } 0 \le t \le x \\ (\frac{1-t}{1-x})^n & \text{ha } x < t \le 1 \end{cases}$$

akkor $\lim_{n\to\infty} \|f_n^x\|_{L^2[0,1]} = 0$, de $f_n^x(x) = 1 \ \forall x$.

Megjegyzés. Egyébként $L^2[0,1]$ elemei nem konkrét függvények, hanem Lebesgue majdnem mindenütt megegyező függvényosztályok, tehát a kiértékelő függvény szigorúan véve nem jól definiált, de ha minden függvényosztályból tetszőlegesen kijelölünk egy függvényt (mondjuk egy olyat, amelyik a legkevesebb szakadási ponttal rendelkezik), hogy reprezentálja az osztályt, akkor a kiértékelő függvény már jóldefiniált, de így sem kapunk RKHS-t, mint ahogy ezt a fenti ellenpélda is mutatja.

3.1.12. Lemma. *[6, Lemma 2.2.]*
$\mathcal{H}$ egy RKHS $\mathcal{X}$-en, $\{f_n\} \subset \mathcal{H}$. Ekkor ha $\lim_{n\to\infty} \|f_n - f\|_{\mathcal{H}} = 0$ akkor $\lim_{n\to\infty} f_n(x) = f(x) \ \forall x \in \mathcal{X}$

Bizonyítás.

$$|f_n(x) - f(x)|^2 = |\langle f_n - f, k_x \rangle|^2 \le \|f_n - f\| \, \|k_x\|$$

ahol $\|k_x\|$ konstans rögzített x-re, tehát ha a jobb oldal tart 0-hoz, akkor a bal oldali kifejezés is. Az egyenlőtlenség a CSB-egyenlőtlenség. □

Megjegyzés. Az előző lemma nagyon jól szemlélteti a különbséget L^2 és az RKHS-ek között. Az utóbbi esetben a norma beli konvergenciából következik a pontonkénti konvergencia is, de L^2-re ez nem igaz, erre épült az ellenpélda is, hiszen $\lim_{n\to\infty} \|f_n^x - 0\|_{\mathcal{H}} = 0$, de $\lim_{n\to\infty} f_n^x(x) = 1 \neq 0$.

Azonban vannak $L^2[0,1]$-nek olyan alterei, melyek már RKHS-ek, nézzünk erre két ilyen ismertebb függvényteret.

3.2. További példák

3.2.1. Példa (Szoboljev-terek). [6, 1.3.1 Sobolev spaces]
Legyen H_0^1 az abszolút folytonos, majdnem mindenütt deriválható valós f függvények halmaza $[0,1]$-en, melyekre $f(0) = f(1) = 0$ és $f' \in L^2[0,1]$. Továbbá legyen a skaláris szorzás $\langle f, g \rangle_{H_0^1} = \int_0^1 f'(x)g'(x)\,dx$. Tehát H_0^1 nem altere az $L^2[0,1]$-nek abban az értelemben, hogy más skaláris szorzatot használ, de a H_0^1-beli függvények mind $L^2[0,1]$-beilek is. Annak a belátásához, hogy H_0^1 egy RKHS, a következő tételt használjuk majd:

3.2.1. Tétel. *[6, 1.3.1 Sobolev spaces]*
f függvény abszolút folytonos pontosan akkor, ha $f'(x)$ létezik majdnem minden x-re és $f = \int f' + c$.

Az RKHS három feltételének ellenőrzése:
$(i) - (ii)$ [3, 10.4. Állítás] szerint H_0^1 egy Hilbert tér.
(iii) A Newton-Leibniz szabályból, az abszolút folytonosságból, majd a CSB-egyenlőtlenségből következik, hogy minden $f \in H_0^1$-ra és $x \in [0,1]$-re

$$|f(x)| = \left|\int_0^x f'(t)\,dt\right| = \left|\int_0^1 f'(t)\mathbb{I}_{[0,x]}(t)\,dt\right| \leq \left(\int_0^1 f'(t)^2\,dt\right)^{\frac{1}{2}} \left(\int_0^1 \mathbb{I}_{[0,x]}(t)\,dt\right)^{\frac{1}{2}} = \|f\|\sqrt{x}$$

Tehát a kiértékelő funkcionál korlátos, így H_0^1 tényleg egy RKHS.

A reprodukáló kernel kiszámításához használjuk a következő tételt:

3.2.2. Tétel. *[6, Theorem 2.4.]*
Ha $\mathcal{H}$ egy RKHS $\mathcal{X}$ felett, és $\{e_n\}$ egy teljes ortonormált rendszer $\mathcal{H}$-ban, akkor

$$k(x,y) = \sum_{n=1}^{\infty} \overline{e_n(y)} e_n(x)$$

Bizonyítás. Vegyük k_y Fourier-sorát $\{e_n\}$ szerint:

$$k_y = \sum_{n=1}^{\infty} \langle k_y, e_n\rangle\, e_n = \sum_{n=1}^{\infty} \overline{\langle e_n, k_y\rangle} e_n = \sum_{n=1}^{\infty} \overline{e_n(y)} e_n$$

Tehát $K(x,y) = k_y(x) = \sum_{n=1}^{\infty} \overline{e_n(y)} e_n(x)$ □

Nézzük az alábbi függvényeket $n \neq 0$-ra:

$$c_n(t) = \frac{1}{\sqrt{2}\pi n}(\cos(2\pi n t) - 1),\ \ s_n(t) = \frac{1}{\sqrt{2}\pi n}\sin(2\pi n t)$$

Ezeknek a deriváltjai $c_n' = -\sqrt{2}\sin(2\pi n t)$, illetve $s_n' = \sqrt{2}\cos(2\pi n t)$, tehát ha $\langle f, c_n\rangle = \langle f, s_n\rangle = 0\ \forall n > 0$, akkor f' ortogonális minden nem konstans $L^2[0,1]$-beli függvényre, mert $\{1, c_n', s_n' : n > 0\}$ egy ortonormált bázis ezen a téren. Így f' csak valami konstans függvény lehet, tehát f egy elsőfokú polinom. Viszont csak egy ilyen f van, ami teljesíti az $f(0) = f(1) = 0$ peremfeltételeket, mégpedig a konsans nulla függvény. Tehát $\{c_n, s_n : n > 0\}$ teljes H_0^1-en. Az ortonormáltság pedig következik a deriváltak $L^2[0,1]$-beli ortonormáltságából, így már alkalmazhatjuk az előző tételt:

$$\begin{aligned} k(x,y) &= \sum_{n=1}^{\infty} \overline{c_n(y)} c_n(x) + \sum_{n=1}^{\infty} \overline{s_n(y)} s_n(x) \\ &= \sum_{n=1}^{\infty} \frac{1}{2\pi^2 n^2}(\cos(2\pi n x)\cos(2\pi n y) - \cos(2\pi n x) - \cos(2\pi n y) + 1 + \sin(2\pi n x)\sin(2\pi n y)) \\ &= \begin{cases} (1-y)x & \text{ha } x \leq y \\ y(1-x) & \text{ha } x > y \end{cases} \end{aligned}$$

3.2.2. Példa (Paley-Wiener terek). [6, 1.3.2 Paley-Wiener spaces]
Érdekes eredmény, hogy az $L^2[-A,A]$-beli függvények nem alkotnak RKHS-t, de ezen függvényeknek a Fourier transzformáltjai viszont már igen, illetve az így kapott tér egy valódi altere $L^2[\mathbb{R}]$-nek.

3.2.3. Definíció. Egy $f:\mathbb{R}\to\mathbb{R}$ függvény **Fourier transzformáltja**

$$\widehat{f}(x)=\int_{-\infty}^{\infty} f(t)e^{-2\pi ixt}\,dt$$

Legyen $A>0$. Ekkor a **Paley-Wiener tér** $PW_A \coloneqq \left\{\widehat{f}\,|f\in L^2[-A,A]\right\}$ vagyis az $L^2[-A,A]$ elemeinek Fourier transzformáltjai.

Megjegyzés. Az integrálásnak köszönhetően az L^2-beli ekvivalenciasztályok helyett itt az $\widehat{f}$-ek már egyértelműek, tehát PW_A elemei konkrét függvények, L^2-vel ellentétben.

Mivel $\left\{e^{2\pi i\frac{n}{A}t}|n\in\mathbb{Z}\right\}$ egy ortonormált bázis $L^2[-A,A]$-ban, ezért ha $f\in L^2[-A,A]$ Fourier transzformáltja $\widehat{f}(n/A)=0$ minden $n\in\mathbb{Z}$-re, akkor f majdnem mindenütt nulla, hiszen ortogonális az előbbi bázis minden elemére. Tehát $\widehat{\cdot}: L^2[-A,A]\to PW_A$ injektív, mert a magtere a majdnem mindenütt nulla függvények.

Mivel PW_A definíciója miatt szürjektív, illetve az integrálás linearitásából következően lineáris is, ezért $\widehat{\cdot}$ egy izomorfizmus, tehát ha úgy definiáljuk a skaláris szorzást, hogy

$$\left\langle \widehat{f},\widehat{g}\right\rangle_{PW_A} \coloneqq \langle f,g\rangle_{L^2[-A,A]}=\int_{-A}^{A} f(t)\overline{g(t)}\,dt$$

akkor PW_A egy Hilbert tér lesz $\mathbb{R}$-en.

3.2.4. Állítás. *PW_A a fenti skaláris szorzással ellátva egy RKHS.*

Bizonyítás. Már csak azt kell belátni, hogy minden kiértékelőfüggvény korlátos:

$$\left|\,\widehat{f}(x)\,\right|=\left|\int_{-A}^{A} f(t)e^{-2\pi ixt}\,dt\right|\le \|f(t)\|_2\left\|e^{-2\pi ixt}\right\|_2=\sqrt{2A}\left\|\,\widehat{f}\,\right\|_{PW_A}$$

□

Megjegyzés. A definiált skaláris szorzás segítségével könnyen megkapható a kernel függvény is: Keressük először azt a k_y-t, amire minden $\widehat{f}\in PW_A$-ra $\left\langle \widehat{f},k_y\right\rangle=\widehat{f}(y)$. A Fourier transzformált definíció szerint

$$\widehat{f}(y)=\int_{-A}^{A} f(t)e^{-2\pi iyt}\,dt=\int_{-A}^{A} f(t)\overline{e^{2\pi iyt}}\,dt=\left\langle f(t),e^{2\pi iyt}\right\rangle_{L^2[-A,A]}$$

Tehát innen következik, hogy

$$k_y(x)=\widehat{e^{2\pi iyx}}=\int_{-A}^{A} e^{2\pi iyt}e^{-2\pi ixt}\,dt=\int_{-A}^{A} e^{2\pi i(y-x)t}\,dt$$

Az integrált kiszámolva azt kapjuk, hogy

$$k_y(x) = k(x,y) = \begin{cases} \frac{1}{\pi}\frac{\sin(2\pi A(x-y))}{x-y} & \text{ha } x \neq y \\ 2A & \text{ha } x = y \end{cases}$$

3.3. Pozitív definit kernelek

Most, hogy formálisan bevezettük az RKHS-eket és néztünk is rájuk néhány példát, illetve alaptulajdonságot, nézzük meg, hogy miért relevánsak a kernelizálás szempontjából. Azt szeretnénk belátni, hogy a kernel függvények (2.2.1), a pozitív definit kernelek (2.2.3) és a valós RKHS-ekhez tartozó reprodukáló kernelek egymással ekvivalens fogalmak.

Azt már láttuk, hogy minden kernel függvény pozitív definit is (2.2.4). Először azt látjuk be, hogy a valós RKHS-ekhez tartozó reprodukáló kernelek kernel függvények, majd azt, hogy pozitív definitek. Ezután a Moore-Aronszajn tétel segítségével konstruálunk minden k pozitív definit kernelhez egy RKHS-t, aminek k a reprodukáló kernele, ezzel belátva a két fogalom közötti ekvivalenviát. És ebből már következik az is, hogy a három fogalom ekvivalens.

3.3.1. Állítás. *Ha $\mathcal{H}$ egy valós RKHS $\mathcal{X}$-en a k reprodukáló kernellel, akkor $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ egy kernel függvény.*

Bizonyítás. $k(x,x') = \langle k_{x'}, k_x \rangle_{\mathcal{H}} = \langle k_x, k_{x'} \rangle_{\mathcal{H}}$ definíció szerint. Tehát ha a $\Phi : \mathcal{X} \to \mathcal{H}$ hozzárendelés $\Phi(x_i) = k_{x_i}$, akkor $k(x,x') = \langle \Phi(x), \Phi(x') \rangle_{\mathcal{H}}$. □

Mivel vannak olyan reprodukáló kernelek, amik felvehetnek komplex értékeket, ezért érdemes definiálni a definitséget komplex kernelekre is.

3.3.2. Definíció. [6, 2.2]
$A \in \mathbb{C}^{n \times n}$ mátrix **pozitív szemidefinit** ha $\langle Ax, x \rangle \geq 0 \ \forall x \in \mathbb{C}^n$ Ha ez szigorú egyenlőtlenséggel teljesül minden nemnulla x esetén, akkor azt mondjuk, hogy A **pozitív definit**.

Ekkor a komplex kernelek definitségét 2.2.3-hez hasonlóan definiáljuk.

3.3.3. Állítás. *[6, Proposition 2.13]*
Ha $\mathcal{H}$ egy RKHS $\mathcal{X}$-en, k reprodukáló kernellel, akkor k pozitív definit.

Bizonyítás. Tetszőleges $x_1, \ldots, x_n \in \mathcal{X}$-re és $\alpha \in \mathbb{F}^n$-re a K Gram mátrixszal:

$$\langle K\alpha, \alpha \rangle = \sum_{i,j=1}^{n} \overline{\alpha_i}\alpha_j k(x_i, x_j) = \sum_{i,j=1}^{n} \overline{\alpha_i}\alpha_j \left\langle k_{x_j}, k_{x_i} \right\rangle = \left\langle \sum_{j=1}^{n} \alpha_j k_{x_j}, \sum_{i=1}^{n} \alpha_i k_{x_i} \right\rangle = \left\| \sum_{i=1}^{n} \alpha_i k_{x_i} \right\| \geq 0$$

□

3.3.4. Tétel (Moore-Aronszajn). *[6, Theorem 2.14, Proposition 2.3]*
Ha egy $\mathcal{X}$ halmaz feletti k kernel pozitív definit, akkor egyértelműen létezik egy olyan $\mathcal{H}$ RKHS $\mathcal{X}$ felett, aminek k a reprodukáló kernele.

Bizonyítás. Rendeljük hozzá minden $x \in \mathcal{X}$-hez azt a $k_x : \mathcal{X} \to \mathbb{F}$ függvényt, ami minden $y \in \mathcal{X}$-hez hozzárendeli $k(y,x)$-t, azaz $k_x = k(\cdot, x)$. Vegyük először ezen k_x-ek által generált vektorteret, legyen ez W. Ezután definiáljuk a skaláris szorzást W-n a következőképpen:

$$\langle f, g\rangle_W = \left\langle \sum_j \alpha_j k_{x_j}, \sum_i \beta_i k_{x_i} \right\rangle_W := \sum_{i,j} \alpha_j \overline{\beta_i} k(x_i, x_j)$$

Mivel f többféle képpen is előállhat, mint k_{x_i}-k lineáris kombinációja, ezért először azt kell belátni, hogy $\langle\cdot,\cdot\rangle_W$ jóldefiniált, vagyis hogy nem függ az értéke az α_j-k és β_i-k választásától. Legyen $f = \sum_j \alpha_j^1 k_{x_j} = \sum_j \alpha_j^2 k_{x_j}$ és $g = \sum_i \beta_i^1 k_{x_i} = \sum_i \beta_i^2 k_{x_i}$ f és g tetszőleges felbontásai.

Ha feltesszük, hogy csak az f-nek különbözik a két felbontása (vagyis $\beta_i^1 = \beta_i^2 = \beta_i$), akkor:

$$\left\langle \sum_j \alpha_j^1 k_{x_j}, \sum_i \beta_i k_{x_i} \right\rangle_W = \sum_{i,j} \alpha_j^1 \overline{\beta_i} k(x_i, x_j) = \sum_i \overline{\beta_i} \sum_j \alpha_j^1 k_{x_j}(x_i) = \sum_i \overline{\beta_i} f(x_i)$$

Ami már nem függ attól, hogy f-et hogyan bontjuk fel. Ezt hasonlóan megtehetjük g esetén is, itt azonban még azt is ki kell használni, hogy K Gram mátrixa konjugáltan szimmetrikus:

$$\left\langle \sum_j \alpha_j k_{y_j}, \sum_i \beta_i^1 k_{x_i} \right\rangle_W = \sum_{i,j} \alpha_j \overline{\beta_i^1 k(x_j, x_i)} = \sum_j \alpha_j \sum_i \overline{\beta_i^1 k_{x_i}(x_j)} = \sum_j \alpha_i \overline{g(x_j)}$$

Tehát ha az egyik felbontást rögzítjük, akkor a másikat teszőlegesen változtathatjuk, így:

$$\left\langle \sum_j \alpha_j^1 k_{x_j}, \sum_i \beta_i^1 k_{x_i} \right\rangle_W = \left\langle \sum_j \alpha_j^2 k_{x_j}, \sum_i \beta_i^1 k_{x_i} \right\rangle_W = \left\langle \sum_j \alpha_j^2 k_{x_j}, \sum_i \beta_i^2 k_{x_i} \right\rangle_W$$

Tehát $\langle\cdot,\cdot\rangle_W$ jóldefiniált. Most lássuk be, hogy tényleg skaláris szorzás:

1. Linearitás az első változóban:

$$\begin{aligned}
\langle \lambda_1 f_1 + \lambda_2 f_2, g\rangle_W &= \left\langle \lambda_1 \sum_j \alpha_j^1 k_{x_j} + \lambda_2 \sum_j \alpha_j^2 k_{x_j}, \sum_i \beta_i k_{x_i} \right\rangle_W \\
&= \left\langle \sum_j (\lambda_1 \alpha_j^1 + \lambda_2 \alpha_j^2) k_{x_j}, \sum_i \beta_i k_{x_i} \right\rangle_W \\
&= \sum_{i,j} (\lambda_1 \alpha_j^1 + \lambda_2 \alpha_j^2) \overline{\beta_i} k(x_i, x_j) \\
&= \lambda_1 \sum_{i,j} \alpha_j^1 \overline{\beta_i} k(x_i, x_j) + \lambda_2 \sum_{i,j} \alpha_j^2 \overline{\beta_i} k(x_i, x_j) \\
&= \lambda_1 \langle f_1, g\rangle_W + \lambda_2 \langle f_2, g\rangle_W
\end{aligned}$$

2. A konjugált szimmetria következik abból, hogy Gram mátrix a pozitív szemidefinitség miatt

önadjungált:

$$\langle f,g\rangle_W = \sum_{i,j}\alpha_j\overline{\beta_i}k(x_i,x_j) = \sum_{i,j}\overline{\beta_i}\alpha_j\overline{k(x_j,x_i)} = \overline{\sum_{i,j}\beta_i\overline{\alpha_j}k(x_j,x_i)} = \overline{\langle g,f\rangle}_W$$

3. Az, hogy $\langle f,f\rangle_W = \sum_{i,j}\alpha_j\overline{\alpha_i}k(x_i,x_j) \geq 0$ következik k pozitív definitségéből, így már csak az van hátra, hogy ez pontosan akkor teljesül egyenlőséggel, ha $f=0$. Egyrészt ha $f=0$, akkor

$$\langle f,f\rangle_W = \sum_{i,j}\alpha_j\overline{\alpha_i}k(x_i,x_j) = \sum_i\overline{\alpha_i}\sum_j\alpha_j k_{x_j}(x_i) = \sum_i\overline{\alpha_i}f(x_i) = 0$$

Másrészt, ha $\langle f,f\rangle_W = 0$ akkor ez átírható $\sum\alpha_j\overline{\alpha_i}k(x_i,x_j) = \alpha^*K\alpha$ alakba, ahol $K_{ij} = k(x_i,x_j)$ a k Gram mátrixa. Ekkor mivel K pozitív szemidefinit, ezért létezik a négyzetgyöke:

$$0 = \alpha^*K^{\frac{1}{2}}K^{\frac{1}{2}}\alpha = (K^{\frac{1}{2}}\alpha)^*K^{\frac{1}{2}}\alpha = \left\langle K^{\frac{1}{2}}\alpha, K^{\frac{1}{2}}\alpha\right\rangle_2 = \left\|K^{\frac{1}{2}}\alpha\right\|_2^2 = 0$$

Ahol a * az adjungáltat jelöli. Tehát $K^{\frac{1}{2}}\alpha = 0$, így tetszőleges $g\in W$-re:

$$\langle f,g\rangle_W = \sum_{i,j}\alpha_j\overline{\beta_i}k(x_i,x_j) = \beta^*K\alpha = \beta^*K^{\frac{1}{2}}K^{\frac{1}{2}}\alpha = \beta K^{\frac{1}{2}}0 = 0$$

Ami csak akkor lehetséges, ha $f=0$. Ezzel beláttuk, hogy $\langle\cdot,\cdot\rangle_W$ egy skaláris szorzás W-n.

Következő lépésként tegyük teljessé W-t, így megkapva $\mathcal{H}$-t, majd az szeretnénk belátni, hogy az így definiált $\mathcal{H}$ minden eleme megfeleltethető egy $\mathcal{F}(\mathcal{X},\mathbb{F})$-beli függvénynek.

Ehhez először definiáljuk $\widehat{h}(x) \coloneqq \langle h,k_x\rangle_W$ függvényeket, majd ezek segítségével a $\widehat{\mathcal{H}} \coloneqq \left\{\widehat{h}\middle| h\in\mathcal{H}\right\}$ halmazt. Ekkor $\widehat{\mathcal{H}}\subset\mathcal{F}(\mathcal{X},\mathbb{F})$ és a $\widehat{\cdot}$ operétor lineáris, tehát $\widehat{\mathcal{H}}$ még vektortér is. Azt szeretnénk belátni, hogy $\widehat{\cdot}$ bijekció.

Mivel a leképezés lineáris, ezért az injektivitáshoz elég azt belátni, hogy ha $\widehat{h}(x)=0\ \forall x\in\mathcal{X}$, akkor $h=0$. Tegyük fel tehát, hogy $\widehat{h}(x)=0\ \forall x\in\mathcal{X}$. Ekkor $\langle h,k_x\rangle_W = 0\ \forall x\in\mathcal{X}$, tehát $h\perp W$. De mivel W sűrű $\mathcal{H}$-ban ($\mathcal{H}$ konstruálása miatt), ezért $h=0$, tehát a leképezés injektív.

A szürjektivitás következik $\widehat{\mathcal{H}}$ definíciójából, tehát $\widehat{\cdot}$ bijekció, így definiálhatunk rajta a következőképpen skaláris szorzást: $\left\langle\widehat{h_1},\widehat{h_2}\right\rangle = \langle h_1,h_2\rangle_W$, és ezzel $\widehat{\mathcal{H}}$ az $\mathcal{X}$-en értelmezett függvények egy Hilbert-tere lesz. Könnyen kiszámolható, hogy ezen a téren a kiértékelő funkcionál korlátos:

$$E_x(\widehat{h}) = \widehat{h}(x) = \langle h,k_x\rangle_W = \left\langle\widehat{h},\widehat{k_x}\right\rangle \leq \left\|\widehat{h}\right\|\left\|\widehat{k_x}\right\|$$

Tehát $\widehat{\mathcal{H}}$ egy RKHS $\mathcal{X}$-en, és a reprodukáló kernele pedig $\widehat{k_y}(x) = k_y(x) = k(k_y,k_x) = k(x,y)$

Már csak az egyértelműséget kell belátni: Tegyük fel, hogy $\mathcal{H}_1,\mathcal{H}_2$ is RKHS-ek $\mathcal{X}$ felett a k kernellel. Legyen W_i a $\{k_x\in\mathcal{H}_i | x\in\mathcal{X}\}$ lineáris burka. Ekkor $f(x) : f\in W_i$ értéke nem függ attól, hogy W_1-ből vagy W_2-ből vesszük, mert $f(x) = \sum_j\alpha_j k_{x_j}(x) = \sum_j\alpha_j k(x,x_j)$.

Illetve tetszőleges $f\in W_i$-re:

$$\|f\|_{\mathcal{H}_1}^2 = \sum_{i,j}\alpha_i\overline{\alpha_j}\left\langle k_{x_i},k_{x_j}\right\rangle_{\mathcal{H}_1} = \sum_{i,j}\alpha_i\overline{\alpha_j}k(x_j,x_i) = \|f\|_{\mathcal{H}_2}^2$$

Tehát $\|f\|_{\mathcal{H}_1} = \|f\|_{\mathcal{H}_2}\ \forall f \in W_i$. Mivel W_1 sűrű $\mathcal{H}_1$-ben a 3.1.10 következmény miatt, ezért létezik minden $f \in \mathcal{H}_1$-hez létezik egy $(f_n) \subset W_1$ sorozat, amire $\|f - f_n\|_{\mathcal{H}_1} \to 0$

Mivel (f_n) egy Cauchy sorozat W_1-ben, ezért (f_n) Cauchy sorozat W_2-ben is (mert a normák megegyeznek). Tehát létezik egy $g \in \mathcal{H}_2 : \|g - f_n\|_{\mathcal{H}_2} \to 0$ Mivel RKHS-ek esetében a norma beli konvergenciából következik a pontonkénti konvergencia a 3.1.12 lemma szerint, ezért $f(x) = \lim_n f_n(x) = g(x)\ \forall x \in \mathcal{X}$, így ha $f \in \mathcal{H}_1$ akkor $f \in \mathcal{H}_2$, vagyis $\mathcal{H}_1 = \mathcal{H}_2$ □

Tehát a valós RKHS-ek és a (valós) pozitív definit kernelek egyértelműen megfeleltethetőek egymásnak úgy, hogy az RKHS reprodukáló kernele a neki megfelelő pozitív definit kernel legyen. Így ezzel beláttuk, hogy a pozitív definit kernelek reprodukáló kernelek és így kernel függvények is.

A fejezet elején láttunk már arra példát, hogy hogyan lehet az RKHS-ből meghatározni a kernel függvényt, szóval most nézzük a másik irányt, az úgynevezett **rekonstrukciós problémát**. Ez általában egy nehezebb feladat, mint a kernel meghatározása a az RKHS-ből.

3.3.1. Példa (Skaláris szorzás által indukált RKHS). [6, 2.3.4]
Legyen $\mathcal{L}$ egy $\mathbb{C}$ feletti Hilbert tér a $\langle\cdot,\cdot\rangle_{\mathcal{L}}$ skaláris szorzással, $k(x,y) = \langle x, y\rangle_{\mathcal{L}}$ a kernel függvény és az ehhez tartozó $\mathcal{H}(k)$ RKHS-t szeretnénk megkeresni.

Bizonyítás. Legyen K a k kernel Gram mátrixa egy tetszőleges $x_1, \dots, x_n \subset \mathcal{L}$ halmazon, azaz $K_{ij} = \langle x_i, x_j\rangle$. Ekkor minden $\alpha \in \mathbb{C}^n$-re:

$$\langle K\alpha, \alpha\rangle_{\mathbb{C}^n} = \sum_{i,j=1}^{n} \langle x_i, x_j\rangle_{\mathcal{L}}\, \alpha_j\overline{\alpha_i} = \sum_{i,j=1}^{n} \langle \overline{\alpha_i}x_i, \overline{\alpha_j}x_j\rangle_{\mathcal{L}} = \left\langle \sum_{i=1}^{n} \overline{\alpha_i}x_i, \sum_{j=1}^{n} \overline{\alpha_j}x_j \right\rangle_{\mathcal{L}} = \left\| \sum_{i=1}^{n} \alpha_i x_i \right\|_{\mathcal{L}}^2 \geq 0$$

Tehát k egy pozitív definit kernel, így létezik egy hozzá tartozó $\mathcal{H}(k)$ RKHS.

A CSB-egyenlőtlenség miatt minden $x \in \mathcal{L}$-re $f_x = \langle\cdot, x\rangle_{\mathcal{L}}$ egy korlátos funkcionál, illetve a lineáris kombinációik is, így az a megérzésünk, hogy ezek fogják alkotni $\mathcal{H}(k)$-t. Ez megegyezik az $\mathcal{L} \to \mathbb{C}$ korlátos lineáris funkcionálok $\mathcal{H}$ vektorterével, hiszen a Riesz reprezentációs tétel szerint minden $f \in \mathcal{H}$-hoz egyértelműen létezik $w \in \mathcal{L}$ úgy, hogy $f = \langle\cdot, w\rangle_{\mathcal{L}} = f_w$. Mivel $\mathcal{H}$ konjugáltan lineárisan izomorf $\mathcal{L}$-el:

$$\lambda_1 f_{w_1} + \lambda_2 f_{w_2} = \overline{\lambda_1}\,\langle\cdot, w_1\rangle_{\mathcal{L}} + \overline{\lambda_2}\,\langle\cdot, w_2\rangle_{\mathcal{L}} = \left\langle\cdot, \overline{\lambda_1}w_1 + \overline{\lambda_2}w_2\right\rangle_{\mathcal{L}}$$

ezért tényleg egy vektortér, illetve az $\langle f_{w_1}, f_{w_2}\rangle_{\mathcal{H}} = \langle w_2, w_1\rangle_{\mathcal{L}}$ skaláris szorzással Hilbert-tér is.

Ha veszünk egy tetszőleges $x \in \mathcal{L}$ pontot, akkor itt az $E_x : \mathcal{H} \to \mathbb{C}$ kiértékelő funkcionál abszolútértéke korlátos:

$$|E_x(f_w)| = |f_w(x)| = |\langle x, w\rangle_{\mathcal{L}}| \leq \|x\|_{\mathcal{L}}\, \|w\|_{\mathcal{L}} = \|x\|_{\mathcal{L}}\, \|f_w\|_{\mathcal{H}}$$

Tehát $\mathcal{H}$ egy RKHS. Mivel $f_w(x) = \langle x, w\rangle_{\mathcal{L}} = \langle f_w, f_x\rangle_{\mathcal{H}}$, ezért a kernel függvény az x pontban

$k_x = f_x$. Tehát $\mathcal{H}$ kernel függvénye:

$$k_{\mathcal{H}}(x, y) = k_y(x) = f_y(x) = \langle x, y \rangle_{\mathcal{L}} = k(x, y)$$

Így az RKHS egyértelműsége miatt $\mathcal{H}(k) = \mathcal{H}$. □

3.3.5. Állítás (Szigorúan pozitív definit kernelek). *Legyen $\mathcal{H}$ egy valós RKHS $\mathcal{X}$ felett a k reprodukáló kernellel. Ha k szigorúan pozitív definit, akkor tetszőleges $(x_1, y_1), \ldots, (x_n, y_n) \in \mathcal{X} \times \mathbb{R}$ mintára létezik olyan $f \in \mathcal{H}$-t, ami interpolálja a mintaelemeket, vagyis $f(x_i) = y_i$ $\forall i \in [n]$-re.*

Bizonyítás. Az x_i pontokhoz tartozó k_{x_i} reprodukáló kernelek által kifeszített altéren pontosan akkor létezik ilyen $f = \sum_{j=1}^{n} \alpha_j k_{x_j}$, ha minden i-re:

$$y_i = f(x_i) = \langle f, k_{x_i} \rangle = \left\langle \sum_{j=1}^{n} \alpha_j k_{x_j}, k_{x_i} \right\rangle = \sum_{j=1}^{n} \alpha_j \left\langle k_{x_j}, k_{x_i} \right\rangle = \sum_{j=1}^{n} \alpha_j k(x_i, x_j)$$

Ami a $K_{ij} = k(x_i, x_j)$ Gram mátrix használatával a $K\alpha = y$ lineáris egyenletrendszerhez vezet. Mivel k szigorúan pozitív definit, ezért a K mátrix pozitív definit, így invertálható. Tehát az $\alpha = K^{-1}y$ együtthatók segítségével $f(x_i) = \sum_{j=1}^{n} \alpha_j k_{x_j}(x_i) = y_i$. □

3.3.6. Következmény. *Szigorúan pozitív definit kernelek használata esetén szükséges a regularizáció használata, különben interpolálnák a mintát.*

3.4. A reprezentációs tétel

Most térjünk egy kicsit vissza az első fejezetben tárgyalt regularizált kockázathoz (1.1.4). (Igazából ennek egy általánosabb esetét vizsgáljuk, hiszen a tételben tetszőleges veszteségfüggvényt használhatunk a minta felett.)

A következőekben ki fogjuk mondani a Reprezentációs tételt, illetve egy speciális változatát, melyek kezelhetővé teszik majd a végtelen dimenziós RKHS-eken való minimalizálást.

3.4.1. Tétel. *[7, Theorem 4.2]*
Legyen $\mathcal{H}$ egy valós RKHS $\mathcal{X}$ felett, $\Omega : [0, \infty] \to \mathbb{R}$ egy szigorúan monoton növekvő függvény, és $L : (\mathcal{X} \times \mathbb{R}^2)^n \to \mathbb{R}$ egy tetszőleges veszteségfüggvény.

Ekkor ha az $\widehat{f} \in \mathcal{H}$ minimalizálja a $J(f) = L((x_1, y_1, f(x_1)), \ldots, (x_n, y_n, f(x_n))) + \Omega(\|f\|_{\mathcal{H}})$ funkcionált $\mathcal{H}$-n, akkor felírható

$$\widehat{f}(x) = \sum_{i=1}^{n} \alpha_i k(x_i, x) = \sum_{i=1}^{n} \alpha_i k_{x_i}(x)$$

alakban, ahol k a $\mathcal{H}$ RKHS-hez tartozó kernel függvény.

Bizonyítás. Először is, mivel az x^2 függvény szigorúan monoton nő a $[0,\infty)$ intervallumon, ezért feltehető, hogy $\overline{\Omega}(\|f\|_{\mathcal{H}}^2) = \Omega(\|f\|_{\mathcal{H}})$, mivel a kettő pontosan ugyanakkor lesz szigorúan monoton növő. Ha S a $k_{x_1},\dots,k_{x_n}$ által generált altér, akkor $\widehat{f}$ felbontható egy S-beli g és egy S-re ortogonális h komponensre, vagyis

$$\widehat{f}(x) = g(x) + h(x) = \sum_{i=1}^{n} \alpha_i k_{x_i}(x) + h(x)$$

Tehát azt szeretnénk belátni, hogy $h = 0$. Mivel

$$\widehat{f}(x_i) = g(x_i) + h(x_i) = \langle g, k_{x_i}\rangle_{\mathcal{H}} + \langle h, k_{x_i}\rangle_{\mathcal{H}} = \langle g, k_{x_i}\rangle_{\mathcal{H}}$$

ezért L nem függ h-tól, tehát h választása csak az $\Omega(\|f\|_{\mathcal{H}})$ tag értékén változtat. Itt azonban

$$\Omega(\|f\|_{\mathcal{H}}) = \overline{\Omega}(\|f\|_{\mathcal{H}}^2) = \overline{\Omega}(\|g\|_{\mathcal{H}}^2 + \|h\|_{\mathcal{H}}^2) \geq \overline{\Omega}(\|g\|_{\mathcal{H}}^2) = \Omega(\|g\|_{\mathcal{H}})$$

$\overline{\Omega}$ szigorú monotinitásából következik az utolsó egyenlőtlenség, illetve az is, hogy ez pontosan akkor teljesül egyenlőséggel, ha $h = 0$. Mivel $\widehat{f}$ minimalizálta J-t, ezért ebből következik $\widehat{f} = g$. □

Tehát így a minimalizáló függvényt elegendő az esetlegesen végtelen dimenziós $\mathcal{H}$ helyett az adatokhoz tartozó kernelfüggvények által kifeszített véges dimenziós altéren keresni, ezzel kezelhetővé téve a feladatot. Azonban a tétel bizonyításakor feltettük, hogy létezik az optimalizálandó feladatnak minimuma a $\mathcal{H}$ RKHS-en. A következő tétel kimond elégséges feltételeket a minimum létezésének és egyértelműségének érdekében az $\Omega(f) = \lambda\|f\|_{\mathcal{H}}^2$ esetben.

3.4.2. Tétel. *[6, Theorem 8.8]*
$\mathcal{H}$ egy valós RKHS $\mathcal{X}$ felett, $L((x_i, y_i, f(x_i))_{i\in[n]})$ konvex és keressük azt az $\widehat{f} \in \mathcal{H}$-t, ami minimalizálja $J(f) \coloneqq L(f) + \lambda\|f\|_{\mathcal{H}}^2$-t. ($\lambda > 0$) Ekkor ez az $\widehat{f}$ létezik és egyértelmű.

Bizonyítás. A létezés belátásához elegendő belátni, hogy a $W \coloneqq \{f \in \mathcal{H} \mid \exists \alpha \in \mathbb{R}^n : f = \sum_{i=1}^n \alpha_i k_{x_i}\}$ halmazon létezik a minimum, hiszen a reprezentációs tétel bizonyításához hasonló érveléssel látható, hogy egy $\mathcal{H} \setminus W$-beli függvény W-re vett projekciója nem növeli a veszteséget. Minden $f \in W$ felírható $f = \sum_{i=1}^n \alpha_i k_{x_i}$ alakban, ezért

$$\|f\|_{\mathcal{H}}^2 = \left\langle \sum_{j=1}^{n} \alpha_j k_{x_j}, \sum_{i=1}^{n} \alpha_i k_{x_i} \right\rangle_{\mathcal{H}} = \sum_{i,j} \alpha_i \alpha_j k(x_i, x_j) = \alpha^{\mathrm{T}} K \alpha$$

ahol K a $\mathcal{H}$-hoz tartozó kernel Gram mátrixa az x_i-k felett, $\alpha \in \mathbb{R}^n$. Mivel K pozitív szemidefinit, ezért az $S_c = \left\{\alpha \in \mathbb{R}^n \middle| \alpha^{\mathrm{T}} K \alpha \leq c\right\}$ szinthalmazok zárt ellipszoidok $\mathbb{R}^n$-ben.

Vegyünk egy tetszőleges $f \in W$-t és legyen $c^* \coloneqq J(f)$. Jelölje $f_\alpha \coloneqq \sum_{i=1}^n \alpha_i k_{x_i}$-t. Így ekkor az $S_J \coloneqq \{\alpha \in \mathbb{R}^n | J(f_\alpha) \leq c^*\}$ egy nemüres halmaz. Mivel az L veszteségfüggvény nemnegatív, ezért $J(g) \geq \lambda\|g\|_{\mathcal{H}}^2$ minden $g \in W$-re, tehát $J(f_\alpha) \leq c^*$ esetén:

$$c^* \geq J(f_\alpha) \geq \lambda\|f_\alpha\|_{\mathcal{H}}^2 = \lambda \alpha^{\mathrm{T}} K \alpha$$

Így $S_{c^*/\lambda} = \left\{\alpha \in \mathbb{R}^n \middle| \alpha^{\mathrm{T}} K \alpha \leq \frac{c^*}{\lambda}\right\}$ egy olyan konvex, korlátos, zárt halmaz, ami tartalmazza az összes olyan α pontot, ahol a $J(f_\alpha) \leq c^*$. Mivel J konvex α-ban egy konvex halmazon, ezért folytonos is ([5, 2.30 Lemma]). Folytonos függvényeknek pedig van minimumuk korlátos és zárt halmazokon a Weierstrass-tétel szerint, tehát létezik a minimumhely.

Mivel $\|f\|_{\mathcal{H}}^2$ szigorúan konvex, $L(f)$ pedig konvex, így az összegük, J is szigorúan konvex (4.1.8), amiből következik a minimumhely egyértelműsége. □

3.4.1. Példa (Kernelizált Ridge Regresszió). Legyen $\mathcal{H}$ egy valós RKHS $\mathcal{X}$ felett a k kernellel és adott egy $(x_1, y_1), \ldots, (x_n, y_n) \in \mathcal{X} \times \mathbb{R}$ i.i.d. minta. Ekkor az $\mathbb{R}^d$ feletti ridge regresszióhoz hasonlóan most $\mathcal{H}$-n keressük azt függvényt, ami minimalizálja a regularizált veszteséget ($\lambda > 0$):

$$\operatorname*{argmin}_{f \in \mathcal{H}} L[f] + \lambda \|f\|_{\mathcal{H}}^2 = \frac{1}{n}\sum_{i=1}^{n}(f(x_i) - y_i)^2 + \lambda \|f\|_{\mathcal{H}}^2 = \frac{1}{n}\sum_{i=1}^{n}(\langle f, k_{x_i}\rangle_{\mathcal{H}} - y_i)^2 + \lambda \|f\|_{\mathcal{H}}^2$$

Mivel L konvex f-ben (4.1.9), ezért Reprezentációs tétel, illetve 3.4.2 miatt az optimális $\widehat{f}$ létezik és felbontható az x_j-khez tartozó kernel függvények segítségével:

$$\widehat{f} = \sum_{j=1}^{n} \alpha_j k_{x_j}$$

Tehát a feladat felírható a következő formában is:

$$\begin{aligned}\operatorname*{argmin}_{\alpha \in \mathbb{R}^n} \frac{1}{n}\sum_{i=1}^{n}\left(\left\langle \sum_{j=1}^{n}\alpha_j k_{x_j}, k_{x_i}\right\rangle_{\mathcal{H}} - y_i\right)^2 + \lambda \left\|\sum_{j=1}^{n}\alpha_j k_{x_j}\right\|_{\mathcal{H}}^2 \\ = \frac{1}{n}\sum_{i=1}^{n}\left(\sum_{j=1}^{n}\alpha_j k(x_i, x_j) - y_i\right)^2 + \lambda \sum_{i,j=1}^{n}\alpha_i \alpha_j k(x_i, x_j)\end{aligned}$$

A fenti feladat átírható a $K_{i,j} = k(x_i, x_j)$ Gram mátrix segítségével lineáris algebrai formára:

$$\operatorname*{argmin}_{\alpha \in \mathbb{R}^n} \frac{1}{n}(K\alpha - y)^{\mathrm{T}}(K\alpha - y) + \lambda \alpha K \alpha = \frac{1}{n}(\alpha^{\mathrm{T}} K^{\mathrm{T}} K \alpha - 2\alpha^{\mathrm{T}} K^{\mathrm{T}} y + y^{\mathrm{T}} y) + \lambda \alpha K \alpha$$

Mivel a fenti kifejezés differenciálható és szigorúan konvex α-ban, ezért a minimum abban az $\widehat{\alpha}$-ban vétetik fel, ahol a derivált nulla:

$$\frac{2}{n} K^{\mathrm{T}} K \widehat{\alpha} - \frac{2}{n} K^{\mathrm{T}} y + 2\lambda K \widehat{\alpha} = 0$$

Tehát az optimális $\widehat{\alpha} \in \mathbb{R}^d$ és ezzel $\widehat{f}$ is megkapható az alábbi lineáris egyenletrendszer megoldásával:

$$(K^{\mathrm{T}} K + \lambda n K)\widehat{\alpha} = K^{\mathrm{T}} y \tag{3.4.1}$$

Ha K pozitív definit, akkor a megoldás egyértelmű.

Ha viszont K pozitív szemidefinit, akkor több α megoldása is lehet a feladatnak, azonban ezek ugyanahoz az $f = \sum_{i=1}^{n} \alpha_i k_{x_i}$-hez vezetnek:

Legyen α és α' olyanok, hogy kielégítik a fenti egyenletrendszert. Ekkor:

$$(K^{\mathrm{T}}K + \lambda nK)\alpha = K^{\mathrm{T}}y = (K^{\mathrm{T}}K + \lambda nK)\alpha'$$

Ezt rendezzük a bal oldalra, majd használjuk K pozitív szemidefinitségéből adódó szimmetriáját:

$$(K + \lambda nI)K(\alpha - \alpha') = 0$$

Mivel $(K + \lambda nI)$ pozitív definit, ezért ez a kifejezés csak akkor lehet nulla, ha $K(\alpha - \alpha') = 0$.

Legyen $f = \sum_{i=1}^n \alpha_i k_{x_i}$ és $f' = \sum_{i=1}^n \alpha_i' k_{x_i}$ a két megoldás által meghatározott függvények. Ekkor

$$\|f - f'\|_{\mathcal{H}} = \left\| \sum_{i=1}^n \alpha_i k_{x_i} - \sum_{i=1}^n \alpha_i' k_{x_i} \right\|_{\mathcal{H}} = \left\| \sum_{i=1}^n (\alpha_i - \alpha_i') k_{x_i} \right\|_{\mathcal{H}} = \left\langle \sum_{i=1}^n (\alpha_i - \alpha_i') k_{x_i}, \sum_{i=1}^n (\alpha_i - \alpha_i') k_{x_i} \right\rangle_{\mathcal{H}}$$
$$= \sum_{i,j=1}^n (\alpha_i - \alpha_i')(\alpha_j - \alpha_j') \left\langle k_{x_i}, k_{x_j} \right\rangle_{\mathcal{H}} = (\alpha - \alpha')K(\alpha - \alpha') = (\alpha - \alpha')0 = 0$$

Tehát $f = f'$, vagyis α és α' is ugyanazt a függvényt definiálják, így nincs ellentmondás a 3.4.2 tétellel.

Megjegyzés. Az 1.3 fejezetben is levezettünk már egy képletet α-ra:

$$\bar{\alpha} = \frac{1}{n\lambda}(y - K(K + n\lambda I)^{-1}y)$$

Ahol K $\mathbb{R}^n$-beli vektorok Gram mátrixa volt, de azt állítottuk, hogy a kernel trükk (2.3.1) segítségével kernelizálható a feladat. Ez azt jelenti, hogy ha most egy $\mathcal{H}$ valós RKHS-en keressük a regularizált kockázatot minimalizáló függvényt, akkor az megkapható úgy is, hogy az előző képletben az $\mathbb{R}^n$-beli skaláris szorzást lecseréljük a k kernelre. Ebben az esetben K helyébe a $\mathcal{H}$-beli $k_{x_1}, \ldots, k_{x_n}$ reproduláló kernelek Gram-mátrixát írjuk, azaz:

$$K_{ij} = \left\langle k_{x_i}, k_{x_j} \right\rangle_{\mathcal{H}} = k(x_j, x_i) = k(x_i, x_j)$$

Tehát azt szeretnénk belátni, hogy $\bar{\alpha}$ így kielégíti a 3.4.1 egyenletet, vagyis:

$$(K^2 + \lambda nK)\frac{1}{n\lambda}(y - K(K + n\lambda I)^{-1}y) = Ky$$
$$K^2y + n\lambda Ky - K^3(K + n\lambda I)^{-1}y - n\lambda K^2(K + n\lambda I)^{-1}y = n\lambda Ky$$
$$K^2(I - K(K + n\lambda I)^{-1} - n\lambda(K + n\lambda I)^{-1})y = 0$$
$$K^2(I - (K + n\lambda I)(K + n\lambda I)^{-1})y = 0$$
$$K^2(I - I)y = 0$$

Ami teljesül K értékétől függetlenül, tehát az 1.3 fejezetbeli ridge kernelizálása tényleg működik.

Megjegyzés. A reprezentációs tételt kernelizált ridge regresszióra már 1970-ben bebizonyította Kimeldorf és Wahba ([4, Lemma 2.2]), azonban ekkor még nem így nevezték a feladatot. Az általánosabb változatot csak 2001-ben bizonyította be Schölkopf, Herbrich és Smola.

4. fejezet

Konvex optimalizálás

Az első fejezetben láttuk, hogy $\mathcal{X} \to \mathbb{R}$ regressziós feladatok esetén a probléma általában a kockázat vagy a regularizált kockázat minimalizálása egy $\mathcal{H} \subset \mathcal{F}(\mathcal{X}, \mathbb{R})$ halmaz felett. A harmadik fejezetben szereplő (3.4.1) reprezentációs tétel következménye, hogy ha $\mathcal{H}$ egy RKHS, akkor n elemű minta esetén egy n dimenziós Hilbert-téren keressük a megoldást. Ezek viszont izomorfak $\mathbb{R}^n$-el, ezért azok a módszerek, amik az $\mathbb{R}^n$ feletti minimalizálás során működnek, kernelizálás után is fognak.

Ebben a fejezetben azt fogjuk megvizsgálni, hogy hogyan lehet kezelni az $\mathbb{R}^n$ feletti megszorításos minimalizálási feladatokat, illetve hogy miért érdemes konvexnek és differenciálhatónak választani a veszteségfüggvényt. Továbbá pótolunk néhány hiányosságot korábbról, vagyis definiáljuk, hogy mit jelent a konvexitás és belátunk néhány olyan állítást, melyeket használtunk már korábban, bizonyításuk azonban elmaradt.

4.1. Alapfogalmak

4.1.1. Definíció. [7, Definition 6.1]
Adott $\mathcal{X}$ vektortér. $X \subset \mathcal{X}$ **konvex**, ha $\forall x, x' \in X, \lambda \in [0,1] : \lambda x + (1-\lambda)x' \in X$

4.1.2. Definíció. [7, Definition 6.2]
$f : X \to \mathbb{R}$ függvény **konvex**, ha $\forall x, x' \in X, \lambda \in [0,1]$:

$$\lambda x + (1-\lambda)x' \in X \Rightarrow f(\lambda x + (1-\lambda)x') \leq \lambda f(x) + (1-\lambda) f(x')$$

f **szigorúan konvex**, ha a fenti egyenlőtlenség szigorú egyenlőtlenséggel teljesül minden $\lambda \in (0,1)$ esetén.

4.1.3. Definíció. [7, Corollary 6.6]
Ezek segítségével már tudjuk definiálni a **feltételes konvex optimalizálási feladatot**:

Adott egy X konvex halmaz és $f, c_1, \ldots, c_n : X \to \mathbb{R}$ konvex függvények. A továbbiakban jelölje $c(x)$ a $[c_1(x), \ldots, c_n(x)]^{\mathrm{T}}$ vektort. Keressük f minimumát a $c(x) \leq 0$ feltétel mellett az X halmazon, vagyis:

$$\operatorname*{argmin}_{x \in X} f(x) \tag{4.1.1}$$

$$\text{ahol } c(x) \leq 0 \tag{4.1.2}$$

Megjegyzés. A konvex optimalizálási feladathoz hozzá lehetne venni még további, $e_j(x) = 0$ feltételeket is.

Most, hogy definiáltuk a feladatot, nézzük meg, hogy milyen halmazon teljesül az összes feltétel, illetve hogy az optimális megoldások halamaza milyen tulajdonsággal rendelkezik.

4.1.4. Lemma. *[7, Lemma 6.3]*
X konvex halmaz, $c : X \to \mathbb{R}$ konvex függvény.
Ekkor az $\{x \in X | c(x) \leq z\}$ halmaz konvex minden $z \in \mathbb{R}$-re.

Bizonyítás. Tetszőleges $x, x' \in X$-re és $\lambda \in [0, 1]$-re:

$$c(\lambda x + (1 - \lambda)x') \leq \lambda c(x) + (1 - \lambda)c(x') \leq \lambda z + (1 - \lambda)z = z$$

ahol az első egyenlőtlenség c konvexitásából, a második a feltételből következik. □

4.1.5. Lemma. *[7, Lemma 6.4]*
Konvex halmazok metszete is konvex.

Bizonyítás. Legyenek $X_1, X_2 \subset \mathcal{X}$ konvex halmazok, $x, x' \in X_1 \cap X_2, \lambda \in [0, 1]$.
Ekkor $\lambda x + (1 - \lambda)x'$ eleme mind X_1-nek, mind X_2-nek, így a metszetüknek is. □

4.1.6. Következmény. *$C \coloneqq \{x \in X | c_i(x) \leq 0 \ \forall i \in [n]\}$, vagyis az a halmaz, amelyre az összes megszorítás teljesül, konvex.*

4.1.7. Állítás. *[7, Theorem 6.5]*
Ha létezik f-nek minimuma C-n, akkor azok a C-beli pontok, ahol f minimális konvex halmazt alkotnak. Ha f szigorúan konvex, akkor pontosan egy elemből áll.

Bizonyítás. Legyen c^* f minimuma C-n. Ekkor $X_m \coloneqq \{x \in \mathcal{X} | f(x) \leq c^*\}$ konvex, tehát az optimális megoldások halmaza, $X_m \cap C$ is konvex. Ha f szigorúan konvex, akkor különböző $x, x' \in X_m \cap C$ esetén a konvex kombinációjuk $\lambda \in (0, 1)$-re:

$$f(\lambda x + (1 - \lambda)x') < \lambda f(x) + (1 - \lambda)f(x') = \lambda c^* + (1 - \lambda)c^* = c^*$$

Tehát a konvex kombinációjuk kisebb értéket venne fel, mint c^*, ami ellentmondás. □

4.1.8. Állítás. *Legyen $\mathcal{X}$ egy vektortér, $f : \mathcal{X} \to \mathbb{R}$ konvex, $g : \mathcal{X} \to \mathbb{R}$ szigorúan konvex függvények, $a \geq 0$, $b > 0$. Ekkor $af + bg$ is szigorúan konvex.*

Bizonyítás. Legyen $\lambda \in (0,1)$, $x, x' \in X$, ekkor:

$$(af + bg)(\lambda x + (1-\lambda)x') = af(\lambda x + (1-\lambda)x') + bg(\lambda x + (1-\lambda)x') <$$
$$a(\lambda f(x) + (1-\lambda)f(x')) + b(\lambda g(x) + (1-\lambda)g(x')) = \lambda(af + bg)(x) + (1-\lambda)(af + bg)(x')$$

Ahol az egyenlőtlenség szigorúsága g szigorú konvexitásából, illetve $b > 0$-ból következik. □

4.1.9. Állítás. *Ha $\mathcal{H}$ egy $\mathbb{R}$ feletti Hilbert tér, akkor a következőképpen definiált $J : \mathcal{H} \to \mathbb{R}$ funkcionál*

$$J(f) = \frac{1}{n}\sum_{i=1}^{n}(\langle f, \varphi_i\rangle - y_i)^2 + C\|f\|^2$$

szigorúan konvex $f \in \mathcal{H}$-ban minden rögzített $(\varphi_1, y_1), \ldots, (\varphi_n, y_n) \in \mathcal{H} \times \mathbb{R}$ és $C > 0$ esetén.

Bizonyítás. Mivel konvex függvények összege is konvex, ezért elegendő a tagokról külön-külön belátni, hogy konvexek. Továbbá a $C, \frac{1}{n} > 0$ konstansokkal való szorzás se változtat a konvexitáson. Legyen $\lambda \in (0,1)$ és $f_1 \neq f_2$. Ekkor a szumma tagjai:

$$(\langle \lambda f_1 + (1-\lambda)f_2, x_i\rangle - y_i)^2 = (\lambda(\langle f_1, \varphi_i\rangle - y_i) + (1-\lambda)(\langle f_2, \varphi_i\rangle - y_i))^2$$
$$< \lambda(\langle f_1, \varphi_i\rangle - y_i) + (1-\lambda)(\langle f_2, \varphi_i\rangle - y_i)^2$$

Ahol az egyenlőtlenség abból következik, hogy $(\langle f_k, \varphi_i\rangle - y_i)$ valós számok $k \in \{1,2\}$-re. Azt pedig már tudjuk, hogy az x^2 függvény szigorúan konvex $\mathbb{R}$-en.

Az $\|f\|^2$-es tag is szigorúan konvex a háromszög-egyenlőtlenég illetve $\lambda \in (0,1)$ miatt:

$$\|\lambda f_1 + (1-\lambda)f_2\|^2 \leq \|\lambda f_1\|^2 + \|(1-\lambda)f_2\|^2 = \lambda^2\|f_1\|^2 + (1-\lambda)^2\|f_2\|^2 < \lambda\|f_1\|^2 + (1-\lambda)\|f_2\|^2$$

□

4.2. Karush-Kuhn-Tucker (KKT) feltételek

4.2.1. Definíció. [7, Corollary 6.6]
Legyen $X = \mathbb{R}^m$. Adottak $f, c_1, \ldots, c_n : \mathbb{R}^m \to \mathbb{R}$ konvex függvények. A továbbiakban **primál feladatnak** nevezzük az alábbi konvex optimalizációs problémát:

$$\operatorname*{argmin}_{x \in \mathbb{R}^m} f(x)$$
$$\text{ahol } c(x) \leq 0$$

Ha nincsenek megszorítások, akkor könnyen minimalizálhatjuk f-et (ha differenciálható) úgy, hogy megoldjuk a $\partial_x f(x) = 0$ egyenletet. Azonban ha már van $c(x) \leq 0$ feltétel is, akkor már nem ilyen

egyszerű a probléma, hiszen például lehet, hogy azok a pontok, ahol a derivált 0 nem teljesítik a feltételeket. Azonban ez egy jó ötletet ad arra, hogy hogyan lehetne megközelíteni a feladatot. Mégpedig úgy, hogy ha nem megszorítjuk a feladatot, hanem csak büntetjük azt, hogy ha egy feltétel nem teljesül.

4.2.2. Definíció. [2, 5.1.1]
Így jutunk az úgynevezett **Lagrange-függvényhez**:

$$L(x, \alpha) = f(x) + \sum_{i=1}^{n} \alpha_i c_i(x)$$

ahol $\alpha \geq 0$ (Itt és a továbbiakban is az α az α_i-ket tartalmazó vektort jelöli, illetve az $\alpha \geq 0$ azt jelenti, hogy az összes eleme nagyobb, mint 0.)

Megjegyzés. Ha

$$f'(x) \coloneqq \sup_{\alpha \geq 0} (L(x, \alpha)) = \sup_{\alpha \geq 0} \left(f(x) + \sum_{i=1}^{n} \alpha_i c_i(x) \right)$$

akkor látható, hogy $f'(x) = f(x)$ minden $x \in \mathbb{R}^m$-re, amire teljesül az összes $c_i(x) \leq 0$ feltétel és $f'(x) = \infty$ ha van legalább egy $c_i(x) > 0$. Tehát az

$$\inf_{x \in \mathbb{R}^m} f'(x) = \inf_{x \in \mathbb{R}^m} \left(\sup_{\alpha \geq 0} (L(x, \alpha)) \right)$$

értéke megegyezik a 4.2.1 primál feladat optimumában felvett értékkel.

Tehát az ötlet az, hogy $L(x, \alpha)$-t egyszerre szeretnénk minimalizálni x-ben és maximalizálni α-ban, ezzel egy nyeregpontot keresve $\mathbb{R}^{n+m}$-ben. Azonban nem mindegy, hogy milyen sorrendben nézzük a kettőt:

4.2.3. Definíció. [2, 5.1.2]
A Lagrange-függvény segítségével definiáljuk a **Lagrange-duálist**:

$$g(\alpha) \coloneqq \inf_{x \in \mathbb{R}^m} (L(x, \alpha)) = \inf_{x \in \mathbb{R}^m} \left(f(x) + \sum_{i=1}^{n} \alpha_i c_i(x) \right)$$

Ekkor a **Lagrange duális feladat** a következő:

$$\operatorname*{argmax}_{\alpha \in \mathbb{R}^n} g(\alpha)$$

$$\text{ahol } \alpha \geq 0$$

4.2.4. Állítás. *[2, 5.1.3]*
Ha p^* a primál feladat infimuma, és d^* a Lagrange duális feladat szuprémuma, akkor $d^* \leq p^*$. Azaz $g(\alpha) \leq p^* \ \forall \alpha \geq 0$

Bizonyítás. Legyen $\bar{x} \in \mathbb{R}^m$ olyan, hogy $c(\bar{x}) \leq 0$ és $f(\bar{x}) = p^*$. Ekkor:

$$g(\alpha) = \inf_{x \in \mathbb{R}^m} L(x, \alpha) \leq L(\bar{x}, \alpha) = f(\bar{x}) + \sum_{i=1}^{n} \alpha_i c_i(\bar{x}) \leq f(\bar{x}) = p^*$$

□

Megjegyzés. Fontos észrevétel, hogy az előbbiekben sehol sem használtuk ki a függvények konvexitását, ezért a fenti érvelés teljesül tetszőleges f, c_i függvények esetén is.

A következő tétel kimondja, hogy $L(x, \alpha)$ nyeregpontja valóban optimális megoldása a primál feladatnak.

4.2.5. Tétel (Elégséges KKT nyeregpont feltétel). *[7, Theorem 6.21]*
$f, c_1, \ldots, c_n : \mathbb{R}^m \to \mathbb{R}$ tetszőleges függvények, $0 \leq \alpha \in \mathbb{R}^n$, $L(x, \alpha) := f(x) + \sum_i \alpha_i c_i(x)$
Ha létezik $(\bar{x}, \bar{\alpha}) : L(\bar{x}, \alpha) \leq L(\bar{x}, \bar{\alpha}) \leq L(x, \bar{\alpha}) \quad \forall x \in \mathbb{R}^m, \ \forall \alpha \geq 0$ azaz $(\bar{x}, \bar{\alpha})$ egy nyeregpont, akkor $\bar{x}$ minimalizálja $f(x)$-et a $c(x) \leq 0$ megszorítás mellett.

Bizonyítás. I. Az így kapott megoldás megengedett, azaz $c(\bar{x}) \leq 0$:
A bal oldali egyenlőtlenség Lagrange-függvényét felírva az $f(\bar{x}) + \sum_i \alpha_i c_i(\bar{x}) \leq f(\bar{x}) + \sum_i \bar{\alpha}_i c_i(\bar{x})$ kifejezéshez jutunk, melyet bal oldalra rendezve a következző egyenlőtlenséget kapjuk:

$$\sum_i (\alpha_i - \bar{\alpha}_i) c_i(\bar{x}) \leq 0$$

Mivel ez fennáll minden $\alpha \geq 0$-ra, ezért ha úgy definiáljuk α^j-t tetszőleges $j \in [n]$-re, hogy $\alpha_i^j = \bar{\alpha}_i + 1$ ha $i = j$ és $\alpha_i^j = \bar{\alpha}_i$ ha $i \neq j$, akkor a fenti egyenlőtlenségbe behelyettesítve azt kapjuk, hogy $c_j(\bar{x}) \leq 0$ minden $j \in [n]$-re. Tehát $\bar{x}$ megengedett megoldása a feladatnak.

II. Az így kapott megoldás optimális, azaz $f(\bar{x}) \leq f(x) \ \forall x : c(x) \leq 0$:
Tekintsük megint a fenti egyenlőtlenséget. Ha most α^j-t úgy definiáljuk, hogy $\alpha_i^j = 0$ ha $i = j$ és $\alpha_i^j = \bar{\alpha}_i$ ha $i \neq j$, akkor behelyettesítve, majd mindkét oldalt -1-el beszorozva $\bar{\alpha}_j c_j(\bar{x}) \geq 0$-t kapunk. De mivel $c_j(\bar{x}) \leq 0$ és $\bar{\alpha}_j \geq 0$, ezért ez csak úgy teljesülhet, ha $\bar{\alpha}_j c_j(\bar{x}) = 0$ minden $j \in [n]$-re.
Tekintsük most a jobb oldali egyenlőtlenséget: Mivel $\bar{\alpha}_i c_i(\bar{x}) = 0 \ \forall i \in [n]$ és $\bar{\alpha}_i c_i(x) \leq 0 \ \forall i \in [n]$, ezért $f(\bar{x}) = f(\bar{x}) + \sum_i \bar{\alpha}_i c_i(\bar{x}) \leq f(x) + \sum_i \bar{\alpha}_i c_i(x) \leq f(x)$, ahol a középső egyenlőtlenség $L(\bar{x}, \bar{\alpha}) \leq L(x, \bar{\alpha})$-ból következik. □

Tehát elégséges feltételt tudunk mondani tetszőleges függvények esetén is, azonban ha feltesszük, hogy konvexek a függvények, illetve hogy teljesülnek a *Slater-feltételek*, (amik a gyakorlatban konvex függvények esetén majdnem mindig teljesülnek) akkor a nyeregpont feltétel már szükséges is.

4.2.6. Lemma (Slater feltételek). *[7, Lemma 6.23]*
$\mathcal{X} \subset \mathbb{R}^m$ konvex, $c_1, \ldots, c_n : \mathcal{X} \to \mathbb{R}$ konvex függvények, $C := \{x \in \mathcal{X} | c(x) \leq 0\}$. Ekkor a következő

két állítás ekvivalens:
(i) $\exists x \in \mathcal{X}$ *úgy, hogy* $c(x) < 0$ *(vagyis van* C*-nek belső pontja)*
(ii) minden nemnulla $\alpha \geq 0$ $\exists x \in \mathcal{X} : \sum_i \alpha_i c_i(x) < 0$

Megjegyzés. az $(i) \Rightarrow (ii)$ triviális, a másik irányt nem bizonyítjuk.

4.2.7. Tétel (Szükséges KKT feltételek). *[7, Theorem 6.25]*
$f, c_1, \ldots, c_n : \mathbb{R}^m \to \mathbb{R}$ *függvények, melyek konvexek az* $X \subset \mathbb{R}^m$ *konvex halmazon, melynek részhalmaza a* $C \coloneqq \{x \in \mathbb{R}^m | c(x) \leq 0\}$ *és a* c_i*-k megfelelnek a Slater feltételeknek. Ekkor ha* $\bar{x}$ *optimalitási pontja a feladatnak, akkor létezik* $\bar{\alpha}^* \geq 0$ *úgy, hogy:*

$$L(\bar{x}, \alpha) \leq L(\bar{x}, \bar{\alpha}^*) \leq L(x, \bar{\alpha}^*) \ \forall x \in \mathbb{R}^m, \alpha \geq 0$$

Bizonyítás. Tegyük fel, hogy $\bar{x}$ megoldása a konvex optimalizálási feladatnak.
Ekkor $X' \coloneqq X \cap \{x \in C | f(x) \leq f(\bar{x})\}$ konvex a 4.1.4 és 4.1.5 lemmák miatt, illetve $\bar{x} \in X'$. Tehát nem létezik olyan $x \in X'$, amire $f(x) - f(\bar{x}) < 0$ és $c(x) < 0$, különben $\bar{x}$ nem lenne optimális. Viszont ekkor a 4.2.6 alapján létezik $(\bar{\alpha}_0, \bar{\alpha}) \in \mathbb{R}^{n+1}$ nemnegatív, nemnulla vektor, melyre

$$(1)\ \bar{\alpha}_0(f(x) - f(\bar{x})) + \sum_i \bar{\alpha}_i c_i(x) \geq 0 \ \forall x \in X'$$

x helyébe $\bar{x}$-t helyettesítve azt kapjuk, hogy $\sum_i \bar{\alpha}_i c_i(\bar{x}) \geq 0$, de mivel $c_i(\bar{x}) \leq 0, \bar{\alpha}_i \geq 0 \ \forall i \in [n]$, ezért:

$$(2)\ \sum_i \bar{\alpha}_i c_i(\bar{x}) = 0$$

(1)-et átrendezve, majd a jobb oldalához (2)-t hozzáadva azt kapjuk, hogy

$$(3)\ \bar{\alpha}_0 f(x) + \sum_i \bar{\alpha}_i c_i(x) \geq \bar{\alpha}_0 f(\bar{x}) + \sum_i \bar{\alpha}_i c_i(\bar{x})$$

Továbbá, mivel $\alpha_i c_i(\bar{x}) \leq 0 \ \forall i \in [n]$, ezért

$$(4)\ \bar{\alpha}_0 f(\bar{x}) + \sum_i \alpha_i c_i(\bar{x}) \leq \bar{\alpha}_0 f(\bar{x}) + 0 = \bar{\alpha}_0 f(\bar{x}) + \sum_i \bar{\alpha}_i c_i(\bar{x}) \leq \bar{\alpha}_0 f(x) + \sum_i \bar{\alpha}_i c_i(x)$$

Ez már majdnem a keresett $L(\bar{x}, \alpha) \leq L(\bar{x}, \bar{\alpha}^*) \leq L(x, \bar{\alpha}^*)$ alak.

Ha $\bar{\alpha}_0 > 0$, akkor $\bar{\alpha}^* = \frac{\bar{\alpha}}{\bar{\alpha}_0}$ jó lenne. Tehát már csak azt kell belátni, hogy $\bar{\alpha}_0$ nem lehet nulla. Ehhez tegyük fel, hogy $\bar{\alpha}_0 = 0$. Ekkor ezt (3)-ba behelyettesítve azt kapjuk, hogy $\sum_i \bar{\alpha}_i c_i(x) \geq \sum_i \bar{\alpha}_i c_i(\bar{x}) = 0$. Mivel azonban $(\bar{\alpha}_0, \bar{\alpha})$ nemnulla és $\bar{\alpha}_0 = 0$, ezért van egy olyan i, amire $\bar{\alpha}_i > 0$. Tehát van egy olyan $\bar{\alpha}$ nemnulla vektor, amire $\sum_i \bar{\alpha}_i c_i(x) \geq 0$ minden $x \in X'$-re, de ez ellentmond a (ii) Slater feltételnek, így ez a lehetőség nem állhat fenn. □

Megjegyzés. [2, 5.2.3]
Ha teljesülnek a Slater feltételek, akkor a primál és a Lagrange duál optimumok megegyeznek, vagyis $d^* = p^*$. Ezt nevezzük **erős dualitásnak**.

Most már vannak szükséges és elégséges feltételeink konvex függvényekre, azonban a gyakorlatban gyakran differenciálható függvényekkel dolgozunk, így ezt kihasználva egy egyenletrendszerré tudjuk alakítani a feladatot, ami előnyös mind gyakorlati, mind elméleti szempontból.

4.2.8. Tétel (Elégséges KKT feltételek differenciálható estben). *[7, Theorem 6.26]*

$f, c_1, \dots, c_n : \mathbb{R}^m \to \mathbb{R}$ konvex, differenciálható függvények. Ekkor $\bar{x}$ minimalizálja $f(x)$-et a $c(x) \leq 0$ feltétel mellett, ha létezik egy olyan $\bar{\alpha} \geq 0$, amire:
(i) $\partial_x L(\bar{x}, \bar{\alpha}) = \partial_x f(\bar{x}) + \sum_i \bar{\alpha}_i \partial_x c_i(\bar{x}) = 0$
(ii) $\partial_{\alpha_i} L(\bar{x}, \bar{\alpha}) = c_i(\bar{x}) \leq 0 \ \forall i \in [n]$
(iii) $\sum_i \bar{\alpha}_i c_i(\bar{x}) = 0$

Megjegyzés. Mivel $x \in \mathbb{R}^m$, ezért $\partial_x L$ egy m dimenziós vektor.

Bizonyítás. Azt fogjuk belátni, hogy minden $x \in C := \{x \in \mathbb{R}^m | c(x) \leq 0\}$-ra $f(x) \geq f(\bar{x})$.

$$f(x) - f(\bar{x}) \underset{1}{\geq} \partial_x f(\bar{x})^{\mathrm{T}}(x - \bar{x}) \underset{2}{=} (-\sum_i \bar{\alpha}_i \partial_x c_i(\bar{x}))^{\mathrm{T}}(x - \bar{x}) \underset{3}{=} -\sum_i \bar{\alpha}_i \partial_x (c_i(\bar{x})^{\mathrm{T}}(x - \bar{x}))$$
$$\underset{4}{\geq} -\sum_i \bar{\alpha}_i (c_i(x) - c_i(\bar{x})) \underset{5}{=} -\sum_i \bar{\alpha}_i c_i(x) \underset{6}{\geq} 0$$

Az 1. és a 4. egyenlőtlenség a függvények konvexitásából, a 2. egyenlőség (i)-ből, az 5. (iii)-ból, a 6. pedig abból, hogy $x \in C$. □

Megjegyzés. [10, 3. The duality theorems]
Ha a Slater feltételek is teljesülnek, akkor $(i) - (iii)$ szükséges is, ezek közül kettő triviális:
(ii)-nek $\bar{\alpha}$-tól függetlenül teljesülnie kell $\bar{x}$-ra ahhoz, hogy megoldás lehessen.
(iii) a 4.2.7 bizonyításában szereplő (2) egyenlőségből következik.

4.3. A Wolfe duális

4.3.1. Definíció. [10, 2. The problems]
Adottak $f, c_1, \dots, c_n : \mathbb{R}^m \to \mathbb{R}$ konvex, differenciálható függvények. Ekkor a 4.2.1 primál feladathoz tartozó **Wolfe duális feladat** a következő:

$$\underset{x,\alpha}{\operatorname{argmax}} \ L(x, \alpha) = f(x) + \sum_{i=1}^{n} \alpha_i c_i(x)$$
$$\text{ahol } \partial_x L(x, \alpha) = \partial_x f(x) + \sum_{i=1}^{n} \alpha_i \partial_x c_i(x) = 0$$
$$\text{és } \alpha \geq 0$$

4.3.2. Tétel (KKT rés). *[10, Theorem 1]*
Ha a Wolfe duális feladatban p^* az f infimuma a primál feladatbeli megszorítások, illetve d^* az L szuprémuma a duál megszorítások mellett, akkor $d^* \leq p^*$.

Bizonyítás. Legyen x^* primál megengedett, illetve (x, α) duál megengedett. Ekkor:

$$f(x^*) - f(x) \underset{1}{\geq} \partial_x f(x)(x^* - x) \underset{2}{=} -\sum_{i=1}^{n} \alpha_i \partial_x c_i(x)(x^* - x) \underset{3}{\geq} -\sum_{i=1}^{n} \alpha_i (c_i(x^*) - c_i(x)) \underset{4}{\geq} \sum_{i=1}^{n} \alpha_i c_i(x)$$

Ahol az 1 és 3 egyenlőtlenségel az f, illetve a c konvexitásából, 2 az első duál feltételből, 4 pedig a primál feltételből következik. Az egyenlőtlenség mindkét oldalához $f(x)$-et adva azt kapjuk, hogy $f(x^*) \geq f(x) + \sum_{i=1}^{n} \alpha_i c_i(x)$, ami minden primál megengedett x^*-ra és duál megengedett (x, α)-ra teljesül, így az infimumukra és a szuprémumukra is. □

A kettő közötti különbség a KKT rés. Ez azért hasznos, mert így akármilyen algoritmust is használunk az optimális (x, α) meghatározására, ennek a segítségével meg tudjuk határozni, hogy milyen közel vagyunk az optimális megoldáshoz, mivel ha eltűnik a különbség, azaz $v = V$, akkor egy optimális pontban vagyunk, hiszen a primál feladat nem vehetne fel ennél kisebb értéket. Ennek a következménye, hogy a KKT feltételben kapott (x, α) pár optimális megoldása a duális feladatnak, hiszen itt a harmadik pont azt mondja ki, hogy ez a rés eltűnik.

Azonban ennek a teljesülése nem csak elégséges, hanem szükséges feltétele is az optimalitásnak a Slater feltételek mellett:

4.3.3. Tétel (Wolfe Tétele). *[10, Theorem 2]*
Ha $f, c_1, \ldots, c_n$ konvexek, differenciálhatóak és teljesülnek a Slater feltételek, akkor az optimumban eltűnik a KKT rés, azaz ha $\bar{x}$ optimális megoldása a primál feladatnak, akkor létezik olyan $\bar{\alpha}$, hogy $f(\bar{x}) = L(\bar{x}, \bar{\alpha})$.

4.3.4. Definíció. [7, (6.72)]
A konvex optimalizálásnak egy speciális esete a **Kvadratikus programozás**, ahol a primál feladatban a feltételek lineárisak, a minimalizálandó függvény pedig négyzetes x-ben:

$$\underset{x}{\operatorname{argmin}} \ \frac{1}{2} x^{\mathrm{T}} K x + c^{\mathrm{T}} x$$
$$\text{ahol } Ax + d \leq 0$$

Ahol $K \in \mathbb{R}^{m \times m}$ egy pozitív definit mátrix, $x, c \in \mathbb{R}^m$, $A \in \mathbb{R}^{n \times m}$ és $d \in \mathbb{R}^n$. Figyeljük meg, hogy itt az összes feltételt egyszerre tároljuk az A és a d segítségével.

Ekkor a Lagrange-függvény a következő lesz:

$$L(x, \alpha) = \frac{1}{2} x^{\mathrm{T}} K x + c^{\mathrm{T}} x + \alpha^{\mathrm{T}} (Ax + d)$$

Mivel mind a minimalizálandó függvény, mind a feltételek konvexek és differenciálhatóak, ezért

használható az 4.2.8 tétel, ami szerint ha x optimális megoldása a primál feladatnak, akkor létezik $\alpha \geq 0$:

$$\begin{aligned}\partial_x L(x,\alpha) &= Kx + c + A^{\mathrm{T}}\alpha = 0\\ \partial_\alpha L(x,\alpha) &= Ax + d \leq 0\\ \alpha^{\mathrm{T}}(Ax+d) &= 0\end{aligned}$$

és ekkor ez az (x,α) pár optimális megoldása a duális feladatnak.

Azonban az első feltétel segítségével ki lehet fejezni α-ból x-et, mivel K pozitív definit, és így invertálható: $x = K^{-1}(-A^{\mathrm{T}}\alpha - c)$

Ennek a segítségével a duális feladatot fel tudjuk írni csak α-ra. Tehát a maximalizálandó tag:

$$\begin{aligned}L(x,\alpha) &= \frac{1}{2}x^{\mathrm{T}}Kx + c^{\mathrm{T}}x + \alpha^{\mathrm{T}}(Ax+d)\\ &= \frac{1}{2}(K^{-1}(-A^{\mathrm{T}}\alpha - c))^{\mathrm{T}}KK^{-1}(-A^{\mathrm{T}}\alpha - c) + c^{\mathrm{T}}K^{-1}(-A^{\mathrm{T}}\alpha - c) + \alpha^{\mathrm{T}}(AK^{-1}(-A^{\mathrm{T}}\alpha - c) + d)\\ &= -\frac{1}{2}\alpha^{\mathrm{T}}AK^{-1}A^{\mathrm{T}}\alpha - (c^{\mathrm{T}}K^{-1}A^{\mathrm{T}} + d)\alpha - \frac{1}{2}c^{\mathrm{T}}K^{-1}c\end{aligned}$$

Ahol mindössze elemi lineáris algebrai átalakításokat végeztünk, illetve kihasználtuk, hogy a K mátrix inverze is pozitív definit, tehát szimmetrikus. Az α-tól független konstans tagot lehagyhatjuk, így a duális probléma a következő:

$$\begin{aligned}&\operatorname*{argmax}_{\alpha} \; -\frac{1}{2}\alpha^{\mathrm{T}}AK^{-1}A^{\mathrm{T}}\alpha - (c^{\mathrm{T}}K^{-1}A^{\mathrm{T}} + d)\alpha\\ &\text{ahol } \alpha \geq 0\end{aligned}$$

Ami egy olyan szempontból egy szebb alak, mint a primál, hogy sokkal egyszerűbbek a feltételek, itt mindössze egy $\alpha \geq 0$ szerepel.

Megjegyzés. A kvadratikus programozási feladatokra már léteznek hatékony megoldóprogramok.

5. fejezet

Szupport vektor regresszió

5.1. Szupport vektor gépek

[7, 1.4 Hyperplane Classifiers]
A Szupport Vektor Gépek (a továbbiakban SVM-ek) eredetileg klasszifikációs feladatokra lettek kitalálva, így érdemes ennek az esetetnek egy egyszerűbb változatát is megvizsgálni az alapötlet bemutatása érdekében.

Adott egy $(x_1, y_1), \dots, (x_n, y_n) \in \mathbb{R}^d \times \{-1, 1\}$ minta egy $\mathbb{P}_{x,y}$ eloszlásból, és ezek segítségével szeretnénk $\mathbb{R}^d$-t kettéosztani egy hipersíkkal úgy, hogy az egyik oldalán legyenek a $+1$-es, másik oldalán a -1-es y_i címkével rendelkező pontok. Minden hipersík felírható $\langle w, x\rangle + b = 0$ alakban és azt, hogy egy pont melyik oldalán helyezkedik el, azt az egyenletbe való helyettesítés előjele mondja meg. Tehát az összes hipersík általi klasszifikáció felírható $f(x) := \operatorname{sgn}(\langle w, x\rangle + b)$ alakban.

Egyenlőre feltesszük, hogy létezik legalább egy olyan hipersík, ami úgy osztja ketté $\mathbb{R}^d$-t, hogy mindkét oldalán csak egyfajta y_i címkével rendelkező x_i pontok vannak. Az ilyenek közül azt szeretnénk majd kiválasztani, amelyik a lehető legjobban általánosítható még nem látott x-ekre. Azonban hogyan lehetne ezt elérni? Az ötlet az, hogy próbáljuk próbáljuk meg a hipersíkot minél távolabb felvenni a pontoktól, vagyis legyen a **margó** a lehető legszélesebb.

Legyen $m > 0$, és jelölje $\mathcal{V}_m$ azon hipersíkok halmazát, melyek legalább m széles margóval rendelkeznek, azaz $\operatorname{dist}(x_i, V) \geq m$ minden $V \in \mathcal{V}_m$-re. Látható, hogy m növelésével $\mathcal{V}_m$ mérete csökken, és így a kapacitása is, amivel csökken a túltanulás lehetősége. Sőt, ekkor a VC dimenzió is csökken, így garantált, hogy alacsonyabb lesz az empirikus kockázat és a kockázat közötti különbség. (1.2.3)

5.1.1. Tétel. *[7, Theorem 5.5]*
Vegyük az olyan hipersíkokat, amelyeknek az egyenlete $\langle w, x\rangle = 0$ alakú és adott $x_1, \dots, x_n \in \mathbb{R}^d$ pontokra $\min_{i\in[n]} |\langle w, x_i\rangle| = 1$. Jelölje $\mathcal{F}$ az olyan klasszifikátorokat, amik ezeket a hipersíkokat

használják döntési határként, vagyis $f_w(x) = \operatorname{sgn}(\langle w, x\rangle)$ *alakúak. Ekkor minden* $\Lambda \in \mathbb{R}$*-re* $\mathcal{F}_\Lambda := \{f_w \in \mathcal{F} | \, \|w\| \leq \Lambda\}$ *VC dimenziója* $h \leq R^2\Lambda^2$ *ahol* R *a legkisebb olyan 0 középpontú kör sugara, ami tartalmazza az összes* x_i*-t. (Az optimalizálási feladat felírása után belátjuk, hogy alacsonyabb* $\|w\|$ *miért vezet szélesebb margóhoz.)*

Az optimális hipersíkhoz tartozó normálvektor megkeresése felírható a következő formában:

$$\begin{aligned} &\underset{w\in\mathbb{R}^d, b\in\mathbb{R}}{\operatorname{argmin}} \ \tau(w) = \frac{1}{2}\|w\|^2 \\ &\text{ahol } y_i(\langle w, x_i\rangle + b) \geq 1 \ \forall i \in [n] \end{aligned}$$

A fenti konvex optimalizálási feladatban mindkét sor von némi magyarázatot maga után. Először nézzük a második sort: az eredeti feltétel az volt, hogy $y_i = \operatorname{sgn}(\langle w, x_i\rangle + b)$ és mivel $y_i = \pm 1$, ezért ez ekvivalens az $y_i(\langle w, x_i\rangle + b) > 0$ feltétellel. Azonban így nem lenne minimuma $\tau(w)$-nek, mivel ha feltesszük, hogy van egy optimális (w, b) pár, akkor ezt $1 > \lambda > 0$-val beszorozva azt kapjuk, hogy

$$y_i(\langle \lambda w, x_i\rangle + \lambda b) = \lambda y_i(\langle w, x_i\rangle + b) > 0$$

Másrészt $\|\lambda w\|^2 = \lambda^2 \|w\|^2 < \|w\|^2$ tehát így $\|w\|^2$-et tetszőlegesen kicsire tudnánk csökkenteni. Ennek a kiküszöbölésére használjuk azt, hogy megengedjük az egyenlőséget is, de ezt egy nullánál nagyobb konstansra követeljük meg. Így ekkor nem fordul elő ez a probléma. Mivel a minta elemszáma véges, ezért ez nem változtat a feltételen.

A másik kérdés az, hogy $\|w\|$ minimalizálása miért felel meg a margó szélességének maximalizálásának. Ez abból következik, hogy egy x_i pont távolsága a $H_{w,b}$ hipersíktól megadható az alábbi képlettel:

5.1.2. Lemma. *[6, Lemma 8.2] A* $V = \{x \in \mathcal{H} : \langle x, v\rangle = c\}$ *affin hipersík és egy* x_i *pont távolsága megadható az alábbi képlettel:*

$$\operatorname{dist}(x_i, V) = \frac{|\langle x_i, v\rangle - c|}{\|v\|}$$

Tehát a $v = w$ és $c = -b$ helyettesítéssel azt kapjuk, hogy

$$\operatorname{dist}(H_{w,b}, x_i) = \frac{|\langle w, x_i\rangle + b|}{\|w\|} = \frac{y_i(\langle w, x_i\rangle + b)}{\|w\|} \geq \frac{1}{\|w\|}$$

Ez margón lévő pontokra fog egyenlőséggel teljesülni, tehát $\|w\|$ fordítottan arányos a hipersíkhoz legközelebb lévő pontok hipersíktól való távolságával.

A feladat egy feltételes konvex optimalizálási feladat, tehát kezdjük a megoldását a Lagrange-függvény felírásával:

$$L(w, b, \alpha) = \frac{1}{2}\|w\|^2 - \sum_{i=1}^{n} \alpha_i(y_i(\langle x_i, w\rangle + b) - 1)$$

Ezután használjuk ki, hogy a feladat differenciálható is, így a KKT-feltételeket használva azt kap-

juk, hogy a nyeregpontban $\partial_w L(w,b,\alpha) = w - \sum_{i=1}^n \alpha_i y_i x_i = 0$ tehát

$$(1)\ w = \sum_{i=1}^n \alpha_i y_i x_i$$

Vagyis w előáll az x_i-k lineáris kombinációjaként. Másrészt

$$(2)\ \partial_b L(w,b,\alpha) = \sum_{i=1}^n \alpha_i y_i = 0 \text{ illetve } (3)\ \alpha_i[y_i(\langle x_i, w\rangle + b) - 1] = 0$$

is következnek a KKT feltételekből és fontos tulajdonságaira mutatnak rá az optimális hipersíknak.

(3) szerint minden i-re vagy az $\alpha_i = 0$ vagy $y_i(\langle x_i, w\rangle + b) = 1$, tehát az összes nem a margón lévő x_i pontra $\alpha_i = 0$. Az (1) szerint ez viszont azt jelenti, hogy w értéke meghatározható a hipersíkhoz legközelebb eső pontok segítségével, ami megegyezik az elvárásainkkal. Ráadásul így ha nem egyszerre érkeznek az adataink, akkor elegendő egyszerre csak ezeket az x_i-ket tárolni, hiszen a minta többi eleme már nem lehet hatással a megoldásra. (Azonban ez a tulajdonság sajnos már nem fog teljesülni a későbbi gyenge margó, illetve a regressziós esetben) Ezeket az x_i-ket nevezzük **szupport vektoroknak**, hiszen ezek „tartják" az optimális hipersíkot.

További érdekesség, hogy (2)-ből következik, hogy ha az összes x_i-ből hatna egy α_i méretű erő merőlegesen a hipersíkra, akkor ezek összege 0 lenne. Ráadásul még az is megkapható, hogy a perdület is 0 lenne, így a hipersík helyben maradna, a szupport vektorok tényleg „tartanák".

A Lagrange-függvénybe (1) segítségével w-be helyettesítve, majd (2)-t kihasználva azt kapjuk, hogy:

$$\frac{1}{2}\sum_{i,j=1}^n \alpha_i\alpha_j y_i y_j \langle x_i, x_j\rangle - \sum_{i=1}^n \alpha_i(y_i(\sum_{j=1}^n \langle x_i, \alpha_j y_j x_j\rangle + b) - 1) =$$
$$-\frac{1}{2}\sum_{i,j=1}^n \alpha_i\alpha_j y_i y_j \langle x_i, x_j\rangle - b\sum_{i=1}^n \alpha_i y_i + \sum_{i=1}^n \alpha_i = -\frac{1}{2}\sum_{i,j=1}^n \alpha_i\alpha_j y_i y_j \langle x_i, x_j\rangle + \sum_{i=1}^n \alpha_i$$

Így megkaphatjuk, hogy a duális feladat a következő:

$$\underset{\alpha\in\mathbb{R}^n}{\operatorname{argmax}}\ W(\alpha) = -\frac{1}{2}\sum_{i,j=1}^n \alpha_i\alpha_j y_i y_j \langle x_i, x_j\rangle + \sum_{i=1}^n \alpha_i$$
$$\text{ahol } \alpha > 0 \text{ és } \sum_{i=1}^n \alpha_i y_i = 0$$

Mivel ez egy kvadratikus programozási feladat (ha a célfüggvény ellentetjét minimalizáljuk), ezért már be lehet adni egy megoldónak. Az α-ból megkapjuk w-t (1) segítségével, majd (3)-ból kiszámolhatjuk b-t egy olyan i-t választva, ahol α_i nemnulla.

Az előző esetben feltettük, hogy az adathalmaz szétválasztható, viszont ez sajnos a gyakorlatban gyakran nem teljesül. Azonban meg lehet oldani, hogy az SVM-ek működjenek ebben az esetben is. Ehhez relaxálni kell a feltételeket slack változók segítségével, majd büntetni ezek használatát.

Így kapjuk a **gyenge margó** változat primál feladatát:

$$\begin{gathered}\underset{w\in\mathbb{R}^d, b\in\mathbb{R}, \xi\in\mathbb{R}^n}{\operatorname{argmin}} \quad \tau(w,\xi) = \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}\xi_i \\ \text{ahol } y_i(\langle w, x_i\rangle + b) \geq 1 - \xi_i \ \forall i \in [n] \\ \xi_i \geq 0\end{gathered}$$

Ahol a C hiperparaméter segítségével tudjuk szabályozni, hogy mennyire legyen nagy súlya a hibáknak. Érdemes megfigyelni, hogy a célfüggvényben ha az ξ_i-ket tekintjük a minta i-edik elemének veszteségének és a $\|w\|^2$-et pedig regularizációs tagnak, akkor az pont a 3.4.2 tétel formátumában lesz, tehát w egyértelműen előáll, mint az x_i-k lineáris kombinációja és ekkor a szeparálható esethez hasonlóan a megfelelő oldalon lévő, a margón kívül eső pontok 0 együtthatóval rendelkeznek.

5.2. ε-SV regresszió

[7, 9.2.1 ε-Insensitive Loss]
A ridge regresszióhoz hasonlóan ezt is $\mathbb{R}^d$-beli lineáris függvényeken vezetjük be és utánna ugyanúgy lehet majd kernelizálni. A különbség a két algoritmus között az, hogy más veszteségfüggvényt használnak és ebből már egy teljesen más módszer adódik. Mivel a veszteségfüggvény ebben az esetben is konvex, így most is a reprezentációs tételből adódó lineáris kombinációt keressük. Jelölje α az ezeket az együtthatókat tartalmazó vektort.

Legyen $(x_1, y_1), \ldots, (x_n, y_n) \in \mathbb{R}^d \times \mathbb{R}$ i.i.d minta egy $\mathbb{P}_{x,y}$ eloszlásból. A ridge regresszió segítségével kapott α-val az a probléma, hogy ha a minta elemszáma nagy, akkor az $f(x) = \sum_{i=1}^{n}\alpha_i \langle x, x_i\rangle$ kiértékeléséhez az összes mintaelemmel vett skaláris szorzatot ki kell számolnunk. Azonban mi azt szeretnénk, hogy a SV klasszifikációhoz hasonlóan itt is legyenek olyan x_i-k, amik nem járulnak hozzá a megoldás értékéhez, hiszen így még ha lassabb is megkapni ezt a felbontást, a kiértékelés sokkal gyorsabb lesz, mert kevesebb skaláris szorzást kell elvégeznünk, illetve numerikusan is vonzóbb, hiszen a ridge regresszió esetén előfordulhat, hogy sok nagyon kicsi α_i-vel kell szoroznunk.

Ennek érdekében lett kitalálva az ε-inszenzitív veszteségfüggvény: $L(x, y, f(x)) = |y - f(x)|_\varepsilon$. Így kapunk itt is egy olyan tartományt a klasszifikációhoz hasonlóan, ahol 0 veszteséget rendelünk hozzá a pontokhoz, ebben az esetben ez a $\langle w, x\rangle + b$ körüli, ε széles sáv lesz. (Ez az alapja annak, hogy ritka felbontást kaphassunk, hiszen az itt lévő pontok nem számítanak bele a veszteségfüggvénybe, így létezésük nem befolyásolja, hogy mi lesz az optimális w és b. De hogy melyek ezek a pontok, azt természetesen csak az algoritmus futtatása után tudjuk meg.)

ε-inszenzitív veszteségfüggvény, és az $\Omega[f] = \frac{\lambda}{2}\|w\|^2$ regularizációs funkcionál használatával azt kapjuk, hogy a regularizált kockázat:

$$R_{\text{reg}}[f] = R_{\text{emp}} + \Omega[f] = \frac{1}{n}\sum_{i=1}^{n}|y_i - \langle w, x_i\rangle - b|_\varepsilon + \frac{\lambda}{2}\|w\|^2$$

Ahol $\lambda > 0$ hiperparaméter segítségével lehet szabályozni a regularizáció súlyát. Ha leosztunk λ-val, akkor a minimumhely nem változik, így a $C = \frac{1}{\lambda}$ helyettesítéssel a kifejezés minimuma:

$$\min_{w\in\mathbb{R}^d, b\in\mathbb{R}} \frac{1}{2}\|w\|^2 + \frac{C}{n}\sum_{i=1}^{n} |y_i - \langle w, x_i\rangle - b|_\varepsilon$$

Mivel ez egy konvex feladat, ezért a minimumhely keresését felírhatjuk konvex optimalizálási feladatként további $\xi^{(*)}$ slack változók hozzávételével a következő alakban:

$$\begin{aligned} \operatorname*{argmin}_{w\in\mathbb{R}^d, b\in\mathbb{R}, \xi^{(*)}\in\mathbb{R}^n} \tau(w, \xi^{(*)}) &= \frac{1}{2}\|w\|^2 + \frac{C}{n}\sum_{i=1}^{n}(\xi_i + \xi_i^*) \\ \text{ahol } (\langle w, x_i\rangle + b) - y_i &\le \varepsilon + \xi_i \\ y_i - (\langle w, x_i\rangle + b) &\le \varepsilon + \xi_i^* \\ \xi_i^{(*)} &\ge 0 \end{aligned}$$

Ahol $\xi^{(*)}$ egyszerre jelöli ξ-t és ξ^*-ot. Az L Lagrange-függvény felírásához használjuk a $\eta^{(*)}, \alpha^{(*)} \ge 0$ duális változókat:

$$\frac{1}{2}\|w\|^2 + \frac{C}{n}\sum_{i=1}^{n}(\xi_i+\xi_i^*) - \sum_{i=1}^{n}\alpha_i(\varepsilon+\xi_i+y_i-\langle w, x_i\rangle - b) - \sum_{i=1}^{n}\alpha_i^*(\varepsilon+\xi_i^*-y_i+\langle w, x_i\rangle + b) - \sum_{i=1}^{n}(\eta_i\xi_i+\eta_i^*\xi_i^*)$$

Az (i) KKT nyeregpont feltétel miatt a Lagrange-függvény deriváltja 0 a primál változókra nézve:

$$\begin{aligned} (1) \quad & \partial_b L = \sum_{i=1}^{n}(\alpha_i - \alpha_i^*) = 0 \\ (2) \quad & \partial_w L = w - \sum_{i=1}^{n}(\alpha_i^* - \alpha_i)x_i = 0 \\ (3) \quad & \partial_{\xi_i^{(*)}} L = \frac{C}{n} - \alpha_i^{(*)} - \eta_i^{(*)} = 0 \end{aligned}$$

(3)-ból következik, hogy $\eta_i^{(*)} = \frac{C}{n} - \alpha_i^{(*)}$, tehát mivel $\eta^{(*)} \ge 0$, ezért $0 \le \alpha^{(*)} \le \frac{C}{m}$ kell, hogy legyen. Illetve (2)-ből megkapjuk a már ismerős $w = \sum_{i=1}^{n}(\alpha_i^* - \alpha_i)x_i$ felbontást. Ezt helyettesítsük be a Lagrange-függvénybe, majd végezzünk néhány átrendezést:

$$
\begin{aligned}
L =& \frac{1}{2}\left\langle \sum_{i=1}^{n}(\alpha_i^* - \alpha_i)x_i, \sum_{j=1}^{n}(\alpha_j^* - \alpha_j)x_j \right\rangle + \frac{C}{n}\sum_{i=1}^{n}(\xi_i + \xi_i^*) \\
& - \sum_{i=1}^{n}\alpha_i(\varepsilon + \xi_i + y_i - \left\langle \sum_{j=1}^{n}(\alpha_j^* - \alpha_j)x_j, x_i \right\rangle - b) \\
& - \sum_{i=1}^{n}\alpha_i^*(\varepsilon + \xi_i^* - y_i + \left\langle \sum_{j=1}^{n}(\alpha_j^* - \alpha_j)x_j, x_i \right\rangle + b) - \sum_{i=1}^{n}(\eta_i\xi_i + \eta_i^*\xi_i^*) \\
=& \frac{1}{2}\sum_{i,j=1}^{n}(\alpha_i^* - \alpha_i)(\alpha_j^* - \alpha_j)\left\langle x_i, x_j \right\rangle - \varepsilon\sum_{i=1}^{n}(\alpha_i + \alpha_i^*) + \sum_{i=1}^{n} y_i(\alpha_i^* - \alpha_i) \\
& + \frac{C}{n}\sum_{i=1}^{n}(\xi_i + \xi_i^*) - \sum_{i=1}^{n}(\alpha_i\xi_i + \alpha_i^*\xi_i^*) - \sum_{i=1}^{n}(\eta_i\xi_i + \eta_i^*\xi_i^*) \\
& - \sum_{i=1}^{n}(\alpha_i^* - \alpha_i)\left(\left\langle \sum_{j=1}^{n}(\alpha_j^* - \alpha_j)x_j, x_i \right\rangle + b\right) \\
=& \frac{1}{2}\sum_{i,j=1}^{n}(\alpha_i^* - \alpha_i)(\alpha_j^* - \alpha_j)\left\langle x_i, x_j \right\rangle - \varepsilon\sum_{i=1}^{n}(\alpha_i + \alpha_i^*) + \sum_{i=1}^{n} y_i(\alpha_i^* - \alpha_i) \\
& + \sum_{i=1}^{n}\left(\xi_i\left(\frac{C}{n} - \alpha_i - \eta_i\right) + \xi_i^*\left(\frac{C}{n} - \alpha_i^* - \eta_i^*\right)\right) \\
& - \sum_{i,j=1}^{n}(\alpha_i^* - \alpha_i)(\alpha_j^* - \alpha_j)\left\langle x_i, x_j \right\rangle - b\sum_{i=1}^{n}(\alpha_i^* - \alpha_i)
\end{aligned}
$$

(1)-ből következik, hogy a b-t tartalmazó tag kiesik, illetve (3)-ból pedig az, hogy a középső sor nulla. Így megkapjuk a duális feladatot pusztán $\alpha^{(*)}$ függvényeként:

$$
\begin{aligned}
\underset{\alpha^{(*)}\in\mathbb{R}^{2n}}{\operatorname{argmax}} L(\alpha^{(*)}) = & -\frac{1}{2}\sum_{i,j=1}^{n}(\alpha_i^* - \alpha_i)(\alpha_j^* - \alpha_j)\left\langle x_i, x_j \right\rangle - \varepsilon\sum_{i=1}^{n}(\alpha_i + \alpha_i^*) + \sum_{i=1}^{n} y_i(\alpha_i^* - \alpha_i) \\
& \text{ahol } \sum_{i=1}^{n}(\alpha_i - \alpha_i^*) = 0 \\
& 0 \le \alpha^{(*)} \le \frac{C}{n}
\end{aligned}
$$

Mivel ez egy kvadratikus programozási feladat, ezért ezt már be lehet adni egy kvadratikus programozási feladat megoldónak.

Megjegyzés. (2)-ből következik, hogy $f(x) = \langle w, x\rangle + b = \sum(\alpha_i^* - \alpha_i)\langle x_i, x\rangle + b$ formában is megkapható, tehát $\alpha^{(*)}$ és b ismeretében f kiértékeléséhez nem szükséges w explicit kiszámolása. α megkapható a fenti feladat megoldásával, már csak b-t kéne megkapni valahogy.

A (*iii*) KKT nyeregpont feltétel miatt a primál feladatban minden feltételnél vagy az adott feltétel,

vagy a hozzá tartozó duál változó 0:

$$\alpha_i[\varepsilon + \xi_i - (y_i - \langle w, x_i\rangle - b)] = 0$$
$$\alpha_i^*[\varepsilon + \xi_i^* + (y_i - \langle w, x_i\rangle - b)] = 0$$
$$\eta_i\xi_i = (\frac{C}{n} - \alpha_i)\xi_i = 0$$
$$\eta_i^*\xi_i^* = (\frac{C}{n} - \alpha_i^*)\xi_i^* = 0$$

Ebből egyrészt látszódik, hogy ha egy (x_i, y_i) páros kívül esik az ε-inszenzitív sávon ($\xi_i^{(*)} > 0$), akkor a hozzá tartozó $\alpha_i^{(*)} = \frac{C}{n}$ a megfelelő oldalon.

Másrészt ha $|y_i - \langle w, x_i\rangle - b| < \varepsilon$, vagyis a mintaelem az ε-sávon belül van, akkor $\alpha_i = \alpha_i^* = 0$, mivel $\xi_i^{(*)} \geq 0$.

Harmadrészt $\alpha_i\alpha_i^* = 0$. Ez onnan látszódik, hogy $\varepsilon > 0$ és $\xi_i^{(*)} \geq 0$, tehát $(y_i - \langle w, x_i\rangle - b)$ előjelétől függően legfeljebb az egyik szögletes zárójelen belüli kifejezés lehet 0, tehát α_i és α_i^* közül legalább az egyiknek 0-nak kell lennie.

Tehát az $\alpha_i^{(*)}$-ok hasonlóan viselkednek a klasszifikációs esethez, ha α_i az x_i pontból az $f(x)$ síkra felfele, α_i^* pedig a lefele ható erő: $\alpha_i\alpha_i^* = 0$, tehát egy pontból csak az egyik irányba hathat, illetve $\sum(\alpha_i - \alpha_i^*) = 0$ tehát az eredőjük 0. Továbbá azok a pontok, amik „elég jó" becsléssel rendelkeznek (azaz az ε-sávon belül vannak) nem befolyásolják w értékét, illetve a sávon kívül esők pedig $\frac{C}{n}$ súllyal rendelkeznek, a pont a határon lévők értéke pedig változhat. Az utóbbiak közül egy $\alpha_i^{(*)} \in (0, \frac{C}{n})$-et választva megkaphatjuk b értékét, hiszen ebben az esetben $\xi_i^{(*)} = 0$ miatt

$$b = y_i - \langle w, x_i\rangle \pm \varepsilon$$

attól függően, hogy az α_i vagy az α_i^* nem nulla, és a $\langle w, x_i\rangle$-t pedig megintcsak ki lehet számolni $\sum_{j=1}^{n}(\alpha_j^* - \alpha_j)\langle x_j, x_i\rangle$ formában.

Megjegyzés. Az, hogy az ε-sávon belül lévő pontok nem járulnak hozzá végül $f(x) = \langle w, x\rangle + b$ kiértékeléséhez volt a célunk az ε-inszenzitív veszteség használatával. Az algoritmus pontosan ugyanezt az eredményt adta volna, ha nincsenek ezek a pontok, vagy ha akárhány másik pontot vettünk volna fel itt. A maradék pontok fogják „tartani" a kiértékelő függvényt, ezek lesznek a szupport vektorok.

Megjegyzés. Ha az ε-inszenzitív sávon kívül eső egyik pontot egy kicsit megváltoztatjuk (csak annyira, hogy továbbra se érjen hozzá a margóhoz), akkor a hozzá tartozó $\alpha_i^{(*)}$ továbbra is C/n marad, tehát a megoldás nem változik.

Az intuíciónknak megfelelően, ha ε-t növeljük, vagyis nagyobb hibát is megengedünk a veszteségfüggvényben, akkor kevesebb szupport vektor lesz, de így egy pontatlanabb becsléshez jutunk. Azonban nyilván azért használjuk ezt a módszert, mert szeretnénk minél kevesebb SV segítségével

egy elég jó becslést kapni. Azonban C-t és ε-t nekünk kell a priori megmondanunk, mint hiperparamétert, így ennek a megfelelő megválasztása problémás lehet, mert nem mindig tudjuk előre, hogy mekkora hibát engedünk meg, csak azt szetnénk, hogy ez a lehető legkisebb legyen.

Ennek a problémának a megoldására találták ki a ν-SV regressziót, ahol a ν hiperparaméter segítségével lehet megmondani, hogy a minta hanyadrésze szerepeljen SV-ként.

5.3. ν-SV regresszió

[7, ν-SV Regression]
A ν-SV regresszióban hozzávesszük az ε-t a primál változók közé, de felveszünk helyette egy új, $\nu \geq 0$ hiperparamétert, ami azt szabályozza, hogy ε-t milyen súllyal büntessük a $\xi_i^{(*)}$-okhoz képest, a C pedig továbbra is a model komplexitást szabályozza. Így annak ellenére használunk ε-inszenzitív veszteségfüggvényt, hogy még nem tudjuk ezt a paramétert, de szerencsére ennek az optimális eredményét kiszámolja az algoritmus a C és a ν paraméterek alapján.

A primál feladat ekkor a következő lesz:

$$\begin{aligned} \underset{w,\varepsilon,b,\xi^{(*)}}{\operatorname{argmin}} \ \tau(w,\xi^{(*)},\varepsilon) &= \frac{1}{2}\|w\|^2 + C\left(\nu\varepsilon + \frac{1}{n}\sum_{i=1}^{n}(\xi_i+\xi_i^*)\right) \\ \text{ahol } (\langle w,x_i\rangle + b) - y_i &\leq \varepsilon + \xi_i \\ y_i - (\langle w,x_i\rangle + b) &\leq \varepsilon + \xi_i^* \\ \xi_i^{(*)} &\geq 0 \\ \varepsilon &\geq 0 \end{aligned}$$

Megjegyzés. Ha $\nu > 1$, akkor az optimum biztosan $\varepsilon = 0$-ban vétetik fel, mivel ekkor a $\xi_i^{(*)}$-ket 'olcsóbb' növelni, mint az ε-t. Ez onnan látható, hogy minden i-re ξ_i vagy ξ_i^* nulla, ezért legfeljebb n $\xi_i^{(*)}$ tagot adunk össze, amit ezután leosztunk n-el. Tehát számunkra csak a $\nu \leq 1$ eset érdekes.

A Lagrange-függvény megintcsak felírható az $\alpha^{(*)}, \eta^{(*)}, \beta \geq 0$ duális változók használatával:

$$\begin{aligned} &L(w,b,\xi^{(*)},\varepsilon,\alpha^{(*)},\eta^{(*)},\beta) = \\ &= \frac{1}{2}\|w\|^2 + C\nu\varepsilon + \frac{C}{n}\sum_{i=1}^{n}(\xi_i+\xi_i^*) - \sum_{i=1}^{n}\alpha_i(\varepsilon+\xi_i+y_i-\langle w,x_i\rangle - b) - \\ &- \sum_{i=1}^{n}\alpha_i^*(\varepsilon+\xi_i^*-y_i+\langle w,x_i\rangle + b) - \sum_{i=1}^{n}(\eta_i\xi_i+\eta_i^*\xi_i^*) - \beta\varepsilon \end{aligned}$$

Megintcsak a KKT-feltételek szerint a Lagrange függvény primál változók szerinti deriváltjai nullák

az optimalitási pontban, így itt a következőknek kell teljesülni:

$$\partial_w L = w - \sum_{i=1}^{n}(\alpha_i^* - \alpha_i)x_i = 0 \tag{5.3.1}$$

$$\partial_b L = \sum_{i=1}^{n}(\alpha_i - \alpha_i^*) = 0 \tag{5.3.2}$$

$$\partial_{\xi_i^{(*)}} L = \frac{C}{n} - \alpha_i^{(*)} - \eta_i^{(*)} = 0 \tag{5.3.3}$$

$$\partial_\varepsilon L = C\nu - \sum_{i=1}^{n}(\alpha_i + \alpha_i^*) - \beta = 0 \tag{5.3.4}$$

Ezen egyenlőségek felhasználásával az ε-SV regresszióhoz hasonlóan megkaphatjuk L-et pusztán α_i-k segítségével, így a Wolfe-duális:

$$\underset{\alpha^{(*)}\in\mathbb{R}^{2n}}{\operatorname{argmax}} W(\alpha^{(*)}) = -\frac{1}{2}\sum_{i,j=1}^{n}(\alpha_i^* - \alpha_i)(\alpha_j^* - \alpha_j)\langle x_i, x_j\rangle + \sum_{i=1}^{n} y_i(\alpha_i^* - \alpha_i) \tag{5.3.5}$$

$$\text{ahol } \sum_{i=1}^{n}(\alpha_i - \alpha_i^*) = 0 \tag{5.3.6}$$

$$\sum_{i=1}^{n}(\alpha_i + \alpha_i^*) \le C\nu \tag{5.3.7}$$

$$0 \le \alpha^{(*)} \le \frac{C}{n} \tag{5.3.8}$$

Ezt pedig megintcsak be lehet adni már egy kvadratikus programozás megoldónak, ezzel megkapva α-t. Ugyancsak az ε-SV algoritmushoz hasonlóan b és ε kiszámolhatóak egy $0 < \alpha_i^{(*)} < C/n$-hez tartozó primál feltétel segítségével. Geometrilag ez azt jelenti, hogy ha tudjuk, hogy melyik pontok helyezkednek el pontosan a margón, illetve ismert egy, a hipersík normálvektorának megfelelő irányba mutató vektor, akkor ebből ki tudjuk számolni a sík eltolásának mértékét (b) és a margó szélességét (ε).

5.3.1. Állítás. *[7, Proposition 9.2] A ν-SV regressziót egy n elemű $(x_1, y_1), \ldots, (x_n, y_n)$ adathalmazra alkalmazva $0 \le \nu \le 1$ esetén ha az optimumban $\varepsilon \neq 0$, az ε-inszenzitív sávon kívül elhelyezkedő pontok száma N, a szupport vektorok száma M, akkor:*
(i) ν egy felső korlát N/n-re
(ii) ν egy alsó korlát M/n-re

Bizonyítás. (i) Az ε-SVR-hez hasonlóan a sávon kívül eső pontokhoz $\alpha_i = C/n$ vagy $\alpha_i^* = C/n$ duál változó tartozik, ezek összege NC/n. Azonban (5.3.7) miatt ez kisebb, mint $C\nu$, tehát $N/n \le \nu$
(ii) A (iii) KKT feltételből és $\varepsilon > 0$-ból következik, hogy $\beta = 0$. Így (5.3.4) szerint $\sum_{i=1}^{n}(\alpha_i + \alpha_i^*) = C\nu$. Azonban (5.3.8) miatt legalább $n\nu$ nemnulla $\alpha_i^{(*)}$-nak kell hogy legyen, tehát $M/n \ge \nu$ □

Megjegyzés. (5.3.6)-ből és (5.3.7)-ból következik, hogy $\sum_{i=1}^{n}\alpha_i \le C\nu/2$ (illetve $\sum_{i=1}^{n}\alpha_i^* \le C\nu/2$), tehát a 5.3.1 állítás alkalmazható külön-külön a sáv alatt, illetve felett elhelyezkedő pontokra $\nu/2$

aránnyal.

5.3.2. Állítás. *[7, Proposition 9.3] Ha a ν-SV regresszió primál feladatának megoldásai ν és C hiperparaméterek mellett $\bar{\varepsilon}, \bar{w}, \bar{b}$, akkor ha ε-SV regressziót futtatnánk ugyanezen a mintán C és $\bar{\varepsilon}$ hiperparaméterekkel, akkor annak a megoldása is $\bar{w}$, illetve $\bar{b}$ lenne.*

Bizonyítás. Mivel a ν-SVR primál célfüggvénye csak egy $+C\nu\varepsilon$ tagban különbözik az ε-SVR célfüggvényétől. Ezért ha az előbbi ad egy optimális primál $\bar{\varepsilon}, \bar{w}, \bar{b}$ megoldást, akkor ha ε-t lerögzítjük $\bar{\varepsilon}$-ban és ezt a kifejezést minimalizáljuk a többi változó felett, akkor nyilván nem változik a megoldás. A második futtatás azonban megfelel egy ε-SVR-nak $\bar{\varepsilon}$ és C paraméterekkel (hiszen a $C\nu\varepsilon$ tag ekkor konsans lesz), tehát ennek a megoldásai is $\bar{w}$ és $\bar{b}$. □

Összefoglalás

A szakdolgozat első fejezetében bevezettük az olyan alapfogalmakat, mint a veszteségfüggvény, kockázat és az empirikus kockázat. Ezután megnéztük a VC dimenziót, mint kapacitás koncepciót, majd levezettük a ridge regresszió megoldóképletét, amiről a következő fejezetben beláttuk, hogy tényleg kernelizálható. A kernel függvényeket a második és a harmadik fejezet tárgyalta. Beláttuk a pozitív definitségen keresztül, hogy a kernel függvények és a valós reprodukáló kernelek ugyanazok, csak más megközelítésben. A kernel függvényeket úgy vezettük be, mint egy $\mathcal{H}$ képtérbe vett vetítés és skaláris szorzás összevonása:

$$k(x, x') = \langle \Phi(x), \Phi(x') \rangle_{\mathcal{H}}$$

A harmadik fejezetben beláttuk, hogy ez a $\mathcal{H}$ lehet a k kernelhez tartozó RKHS is akár és ekkor a $\Phi(x) = k_x$ egy jó választás. Ekkor a reprezentációs tétel miatt az regularizált empirikus kockázatot minimalizáló függvény előáll, mint a mintához tartozó reprodukáló kernelek lineáris kombinációja:

$$\widehat{f} = \sum_{i=1}^{n} \alpha_i k(\cdot, x_i) = \sum_{i=1}^{n} \alpha_i k_{x_i}$$

Tehát kernelizálás után az optimális függvényt a k-hoz tartozó RKHS-en keressük attól függetlenül, hogy mi az algoritmus, ennek a választása az α_i-k értékét befolyásolja csak.

A ridge regresszió α-ját gyors és könnyű kiszámolni, mivel csak egy mátrixinvertálás és egy mátrixszorzás, azonban a kiértékelésnél gondok akadhatnak, mivel az összes együttható hozzájárul $f(x)$ értékéhez egy kicsit, ezért az összes $k_{x_i}(x)$-et ki kell számolnunk. Ezzel szemben a SV-regresszió által meghatározott α-t sokkal lasabb kiszámolni, mivel egy kvadratikus programozási feladatot kell megoldani, viszont cserébe egy ritka reprezentációt kapunk, amivel sokkal gyorsabb lesz az $f(x)$ kiértékelése. Ráadásul a ν-SV regresszióval még a szupport vektorok számát is tudjuk szabályozni.

Irodalom

[1] Christopher M Bishop. *Information Science and Statistics*. Springer, 2006.

[2] Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

[3] Karátson János. *Numerikus funkcionálanalízis*. Typotex Kiadó, 2014.

[4] George S Kimeldorf and Grace Wahba. "A correspondence Between Bayesian Estimation on Stochastic Processes and Smoothing by Splines". In: *The Annals of Mathematical Statistics* 41.2 (1970), pp. 495–502.

[5] Etienne Klerk, Cornelis Roos, and Terlaky Tamás. *Nemlineáris optimalizálás*. Aula Kiadó, 2004.

[6] Vern I. Paulsen and Mrinal Raghupathi. *An Introduction to the Theory of Reproducing Kernel Hilbert Spaces*. Cambridge University Press, 2016.

[7] Bernhard Schölkopf and Alexander J Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, 2002.

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

[9] Eduardo D Sontag. "VC Dimension of Neural Networks". In: *NATO ASI Series F Computer and Systems Sciences* 168 (1998), pp. 69–96.

[10] Philip Wolfe. "A Duality Theorem for Non-Linear Programming". In: *Quarterly of Applied Mathematics* 19.3 (1961), pp. 239–244.