Metoda potpornih vektora s primjenama u ekstrakciji informacije

Jovanović, Ivan

SVEUČILIŠTE U ZAGREBU

PRIRODOSLOVNO–MATEMATIČKI FAKULTET

MATEMATIČKI ODSJEK

Ivan Jovanović

METODA POTPORNIH VEKTORA S PRIMJENAMA U EKSTRAKCIJI INFORMACIJE

Diplomski rad

Voditelj rada:
prof. dr. sc. Zlatko Drmač

Zagreb, rujan, 2019.

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

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

Povjerenstvo je rad ocijenilo ocjenom _____.

Potpisi članova povjerenstva:

1. _____
2. _____
3. _____

Svojim roditeljima

Sadržaj

Uvod

U svakodnevnom životu okruženi smo velikom količinom podataka. Razvoj računarske znanosti i unaprjeđenje tehnologije omogućili su znanstvenicima i inženjerima da, pomoću matematičke teorije, obrade tako velike skupove informacija i iz njih izvedu određene zaključke. Zbog toga je strojno učenje, koje se kao grana umjetne inteligencije bavi automatskom obradom podataka, izrazito zanimljivo područje istraživanja.

Osnovni cilj strojnog učenja je izvođenje znanja iz skupa podataka kojim raspolažemo i korištenje računala u svrhu donošenja odluka, to jest na osnovi viđenih podataka konstruirati model koji će predvidjeti svojstva novih, neviđenih podataka. Strojno učenje se može podijeliti na tri osnovne grupe s obzirom na skup problema koje rješavaju: nadzirano, nenadzirano i polunadzirano strojno učenje.

Nadzirano učenje podrazumijeva da su podaci dani kao parovi, odnosno podrazumijeva postojanje ulazne i izlazne vrijednosti, a naša je zadaća odrediti ovisnost među njima. S obzirom na izlazne vrijednosti razlikujemo klasifikaciju i regresiju. U slučaju nenadziranog učenja nemamo znanja o ciljnoj varijabli, a potrebno je odrediti neku pravilnost u podacima. Neke od zadaća nenadziranog učenja su grupiranje, smanjanje dimenzionalnosti i procjena gustoće. U ovom radu ću se usredotočiti na nadzirano učenje, odnosno na jednu njegovu tehniku modeliranja koju nazivamo metoda potpornih vektora. Metoda jednostavnim postupkom dolazi do određenog podskupa podataka iz ukupnog skupa kojim raspolažemo kako bi konstruirala model koji će vršiti predikciju izlaznih vrijednosti novih podataka.

U prvom poglavlju predstavit ću osnovne zadaće klasifikacije i regresije. Kako su za metodu potpornih vektora ključna znanja iz teorije optimizacije, u drugom će poglavlju biti riječ o nekim rezultatima iz tog područja. U trećem i četvrtom poglavlju prikazat ću kako se metoda potpornih vektora koristi u problemima klasifikacije i regresije. Stvarni problemi često zahtjevaju nelinearna rješenja pa se algoritmi za učenje moraju tome prilagoditi. Velika prednost metode potpornih vektora je njena sposobnost transformiranja nelinearnog problema pomoću jezgrenog trika. O tom zanimljivom svojstvu više ću reći u petom poglavlju. U šestom poglavlju predstavio sam neke implementacijske probleme i algoritam sekvencijalne minimalne optimizacije koji se jako često koristi za rješavanje problema kvadratnog programiranja. U sedmom poglavlju možete vidjeti kako je metoda

potpornih vektora vrlo moćan alat u ekstrakciji informacije na primjerima analize tekstualnih podataka.

Poglavlje 1

Nadzirano učenje

Algoritmi nadziranog učenja pokušavaju utvrditi ulazno-izlaznu vezu u skupu podataka, $\mathcal{D} = \{(x_1, y_1), \ldots, (x_N, y_N)\} \subset \mathcal{X} \times \mathcal{Y}$, koji je unaprijed zadan i kojeg nazivamo *skup primjera za učenje*. $\mathcal{X}$ je *ulazni prostor* ili *prostor primjera*, a $\mathcal{Y}$ nazivamo *izlazni prostor* ili *prostor oznaka*. Osnovne zadaće nadziranog učenja su *klasifikacija* i *regresija*. Kod klasifikacije primjerima pridružujemo oznaku klase kojoj pripadaju, dok im kod regresije pridružujemo neku kontinuiranu vrijednost.

1.1 Klasifikacija

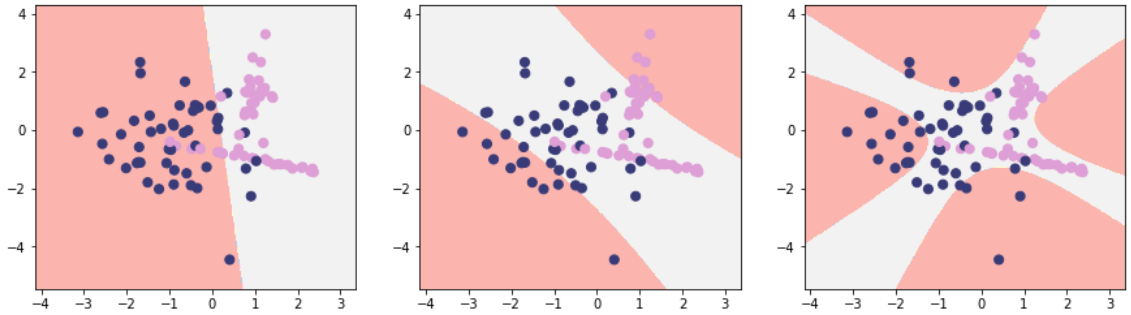
U slučaju klasifikacije ciljna varijabla y je nominalna ili diskretna, odnosno prostor $\mathcal{Y}$ je diskretni prostor oznaka klasa kojima primjeri mogu pripadati.

Zadatak klasifikacijskih algoritama je konstrukcija *hipoteze* koja određuje pripada li primjer nekoj klasi ili ne. To, na primjer, može biti indikatorska funkcija koja s obzirom na odnos primjera x prema hiperravnini u prostoru koju je konstruirao algoritam određuje pripada li primjer klasi. *Model* je skup svih ostvarivih hipoteza.

S obzirom na skup primjera za učenje odabiremo složenost hipoteze. Presloženi modeli dovode do *prenaučenosti* što rezultira lošom sposobnošću generalizacije. Drugim riječima, hipoteza se previše prilagodi skupu primjera za učenje pa loše reagira na dosad neviđene primjere. S druge pak strane, nedovoljno složene hipoteze dovode do *podnaučenosti* što rezultira velikim grešakama na skupu primjera za učenje. Takve hipoteze će i loše generalizirati.

Postoji čitav niz modela kojima rješavamo klasifikacijske probleme. Dijele se na *parametarske* (složenost modela ne ovisi o broju primjera za učenje) i *neparametarske* (broj parametara i složenost rastu s brojem primjera za učenje). Primjeri parametarskih modela su logistička regresija, perceptron, metoda potpornih vektora u primarnoj formulaciji itd. Neki od neparametarskih modela su stabla odluke, algoritam k-najbližih susjeda i metoda

potpornih vektora u dualnoj formulaciji.



Slika 1.1: Binarna klasifikacija primjera iz $\mathbb{R}^2$ metodom potpornih vektora. Korištenjem različitih jezgri postignuta je različita složenost. Na prvom grafu je korištena linearna jezgra, na drugom polinomijalna stupnja 2, a na trećem polinomijalna stupnja 4. Više o tome u petom poglavlju.

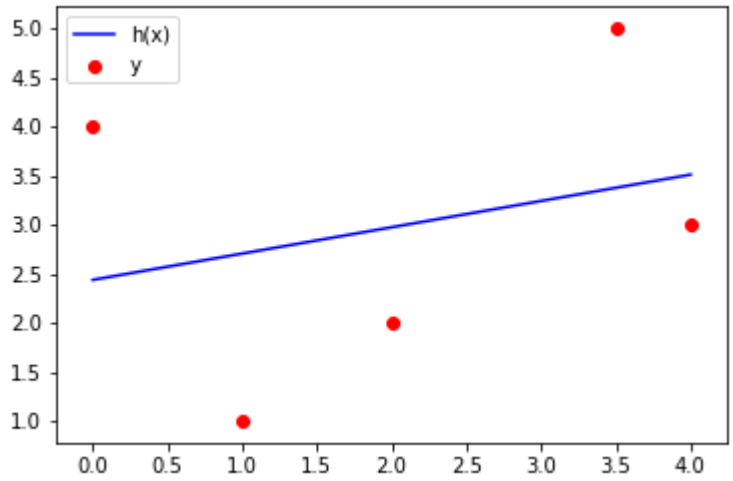
1.2 Regresija

U slučaju regresije varijabla y je kontinuirana, $y \in \mathbb{R}$. Iz skupa za učenje $\mathcal{D}$ induciramo nepoznatu funkciju $f : \mathcal{X} \to \mathbb{R}$ na način da u najboljem slučaju vrijedi $y_i = f(x_i)$. Učenje funkcije je zapravo interpolacija između točaka x_i, to jest ekstrapolacija izvan tih točaka. Kako je u podacima često prisutan šum ε, to jest neželjena anomalija koja se može javiti zbog pogrešaka u mjerenju ili označavanju, zapravo imamo $y_i = f(x_i) + \varepsilon$. Rezultat učenja je aproksimacija funkcije f koju nazivamo hipoteza.

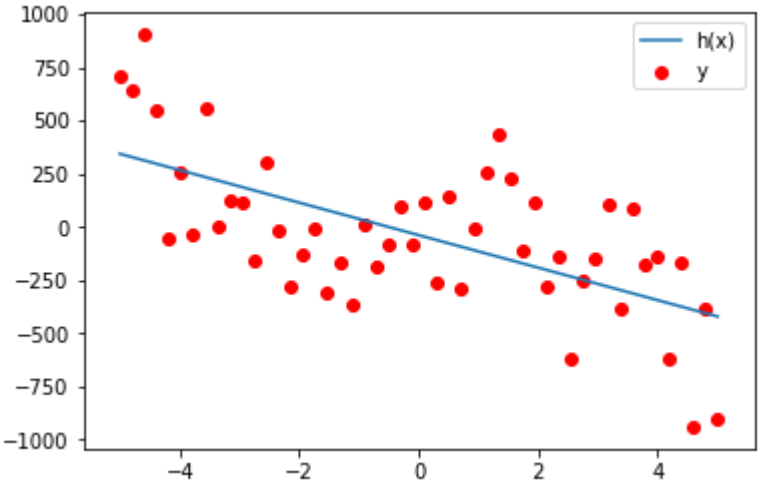
Primjer 1.2.1. *Neka je zadan skup podataka* $\mathcal{D} = \{(0,4),(1,1),(2,2),(3.5,5),(4,3)\}$. *Učenjem modela linearnom regresijom dobivamo hipotezu* $h \approx 0.268x + 2.4375$. *Srednja kvadratna pogreška* ($\frac{1}{N}\sum_{i=1}^{N}(y_i - h(x_i))^2$) *je* 1.839. *Primjer vidimo na slici 1.2.*

Kod regresije također može doći do prenaučenosti, odnosno podnaučenosti. To vidimo na sljedećem primjeru.

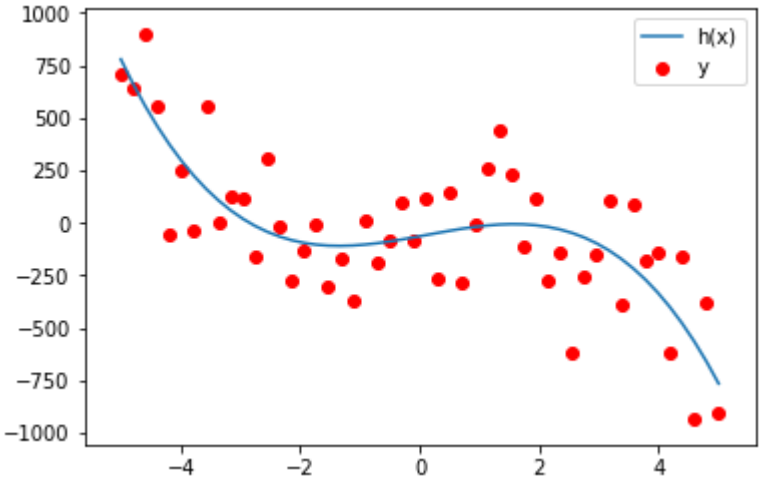
Primjer 1.2.2. *Proizvoljno generiram skup od* 50 *primjera, uniformno distribuiranih u intervalu* $[-5,5]$ *pomoću funkcije* $f(x) = 4 + x - x^2 - 6x^3$ *uz šum* $\sigma = 200$. *Na slici 1.3 prikazan je graf funkcije koju smo dobili linearnom aproksimacijom, na slici 1.4 graf koji smo dobili polinomijalnom aproksimacijom stupnja 3, a na slici 1.5 vidimo graf funkcije koju smo dobili polinomijalnom aproksimacijom stupnja 8.*



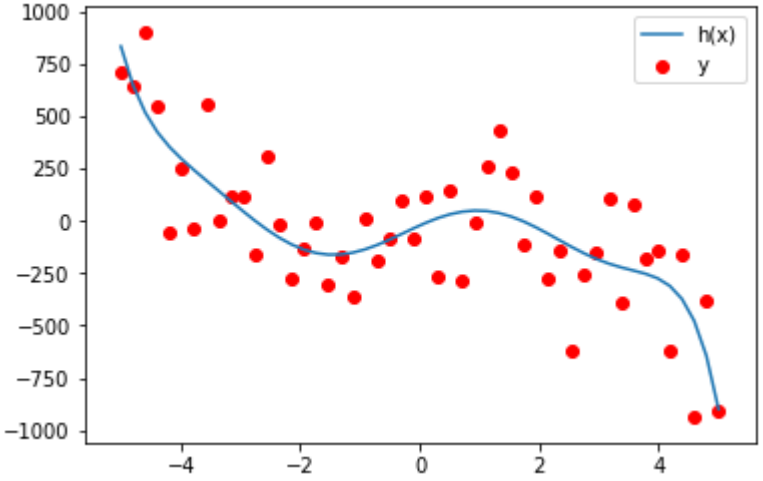
Slika 1.2: Linearna regresija.



Slika 1.3: Linearna regresija.



Slika 1.4: Polinomijalna regresija stupnja 3.



Slika 1.5: Polinomijalna regresija stupnja 8.

Poglavlje 2

Osnovni teoremi optimizacije

U ovom poglavlju obradit ćemo osnovne teoreme u teoriji optimizacije koji će nam kasnije pomoći u konstrukciji metoda za učenje. Za više informacija pogledati deveto poglavlje u knjizi [11].

2.1 Fermatov teorem

Teorem opisuje metodu za pronalazak minimuma, odnosno maksimuma, funkcije definirane na čitavom prostoru, bez ograničenja.

Funkcija $f(x)$ definirana na $\mathbb{R}^n$ ima lokalni minimum (maksimum) u točki $x^* \in \mathbb{R}^n$ ako postoji okolina točke x^* takva da za svaki x iz te okoline vrijedi $f(x^*) \leq f(x)$ ($f(x^*) \geq f(x)$). Lokalne minimume i maksimume zajednički zovemo lokalni ekstremi.

Za početak promotrimo funkcije definirane na $\mathbb{R}$.

Definicija 2.1.1. *Funkcija $f(x)$ definirana na $\mathbb{R}$ je diferencijabilna u točki x^* ako postoji α takav da vrijedi*

$$f(x^* + \lambda) = f(x^*) + \alpha\lambda + r(\lambda), \tag{2.1}$$

gdje je $r(\lambda) = o(|\lambda|)$; to jest za proizvoljni $\varepsilon > 0$ postoji $\delta > 0$ takav da za svaki $\lambda \in \mathbb{R}$

$$|\lambda| < \delta \tag{2.2}$$

povlači

$$|r(\lambda)| < \varepsilon|\lambda|. \tag{2.3}$$

α nazivamo *diferencijal funkcije f u točki x^**, a označava se sa $f'(x^*)$.
Odnosno,

$$f'(x^*) = \lim_{\lambda \to 0} \frac{f(x^* + \lambda) - f(x^*)}{\lambda} = \alpha. \tag{2.4}$$

Teorem 2.1.2. *(Fermat). Neka je $f(x)$ funkcija jedne varijable, diferencijabilna u točki x^*. Ako je x^* točka u kojoj funkcija f poprima lokalni ekstrem, onda je*

$$f'(x^*) = 0. \tag{2.5}$$

Točka x^* za koju vrijedi (2.5) naziva se *stacionarna točka*.

Definicija 2.1.3. *Funkcija $f(x)$ definirana na $\mathbb{R}^n$ je diferencijabilna u točki $x^* = (x_1^*, \ldots, x_n^*)$ ako postoji $\alpha = (\alpha_1, \ldots, \alpha_n)$ takav da vrijedi*

$$f(x^* + h) = f(x^*) + \sum_{i=1}^{n} \alpha_i h_i + r(h), \tag{2.6}$$

gdje je $r(h) = o(|h|)$; to jest za proizvoljni $\varepsilon > 0$ postoji $\delta > 0$ takav da

$$|h| = \sqrt{h_1 + \cdots + h_n} < \delta \tag{2.7}$$

povlači

$$|r(h)| \leq \varepsilon h. \tag{2.8}$$

$\alpha = (\alpha_1, \ldots, \alpha_n)$ nazivamo *gradijent funkcije f u točki x^**, a označava se sa $\nabla f(x^*)$. Vrijednost

$$\alpha_i = \lim_{\lambda \to 0} \frac{f(x^* + \lambda e_i) - f(x^*)}{\lambda} \tag{2.9}$$

gdje je $e_i = (0, \ldots, 1, \ldots, 0)$ naziva se *i-ta parcijalna derivacija* i označava se sa $f_{x_i}(x^*)$ ili $\partial f(x^*)/\partial x_i$, odnosno u skraćenom obliku $\partial_{x_i} f(x^*)$.

Korolar 2.1.4. *(Fermatov teorem za funkciju n varijabli). Neka je f funkcija n varijabli, diferencijabilna u točki x^*. Ako je x^* točka u kojoj f poprima lokalni ekstrem, onda je*

$$\nabla f(x^*) = 0, \tag{2.10}$$

odnosno,

$$f_{x_i}(x^*) = \cdots = f_{x_n}(x^*) = 0. \tag{2.11}$$

2.2 Lagrangeov teorem

Sljedeći korak u teoriji optimizacije je promatranje problema uvjetne optimizacije, odnosno traženje minimuma (ili maksimuma) funkcije f_0 s n varijabli, pod uvjetom da vrijede određene jednakosti:

$$f_1(x) = \cdots = f_m(x) = 0. \tag{2.12}$$

Za funkcije $f_k(x)$, $k = 0, 1, \ldots, m$, pretpostavljamo da posjeduju određenu glatkoću, odnosno pretpostavljamo da su u skupu X, $X \subset \mathbb{R}^n$, neprekidne, kao i njihove parcijalne derivacije (klase C^1 na X).

Kažemo da je $x^* \in X$ točka lokalnog minimuma (maksimuma) u problemu

$$f_0(x) \to \min \tag{2.13}$$

s ograničenjima (2.12) ako postoji $\varepsilon > 0$ takav da za svaki x koji zadovoljava (2.12) i uvjet

$$|x - x^*| < \varepsilon \tag{2.14}$$

vrijedi

$$\begin{gathered} f_0(x) \geq f_0(x^*) \\ \big(f_0(x) \leq f_0(x^*)\big). \end{gathered} \tag{2.15}$$

Sljedeći teorem navodim bez dokaza. Za više informacija pogledati [3].

Teorem 2.2.1. *(Teorem o implicitnoj funkciji). Neka je $F(x, y) \in C^1$ u okolini točke (x^*, y^*) takva da vrijedi $F(x^*, y^*) = 0$ i $F_y(x^*, y^*) \neq 0$. Tada postoji okolina U točke (x^*, y^*) u kojoj postoji implicitna funkcija $y = f(x)$ takva da vrijedi $f(x^*) = y^*$ i $F(x, f(x)) = 0$ za svaki $x \in U$. Također,*

$$f'(x) = -\frac{F_x}{F_y}, \quad x \in U. \tag{2.16}$$

U sljedećem teoremu pretpostavljamo da je $n = 2$ i da je zadan samo jedan uvjet jednakosti. Za više informacija pogledati [4].

Teorem 2.2.2. *Neka su funkcije $f(x, y)$ i $g(x, y)$ klase C^1 na okolini točke (x^*, y^*) i neka je (x^*, y^*) točka lokalnog ekstrema funkcije f, pod uvjetom da vrijedi jednakost $g(x, y) = 0$. Neka je $\nabla g \neq 0$. Tada postoji λ takva da vrijedi*

$$\nabla(f + \lambda g)(x^*, y^*) = 0. \tag{2.17}$$

Dokaz. Kako je, po pretpostavci, $\nabla g \neq 0$, tada vrijedi da je barem jedna od vrijednosti $g_x(x, y)$ i $g_y(x, y)$ različita od 0 u nekoj točki (x, y). Bez smanjenja općenitosti pretpostavimo da u točki (x^*, y^*) vrijedi

$$g_y(x^*, y^*) \neq 0. \tag{2.18}$$

Iz pretpostavki teorema znamo da je $g \in C^1$ i da vrijedi

$$g(x^*, y^*) = 0 \tag{2.19}$$

pa možemo primjeniti teorem 2.2.1 koji kaže da postoji funkcija $y = y(x)$ takva da vrijedi $g(x, y(x)) = 0$.
Nadalje, vrijedi

$$y'(x) = -\frac{g_x}{g_y} \tag{2.20}$$

Kako funkcija $f(x, y)$ u točki (x^*, y^*) ima lokalni ekstrem, slijedi da $f(x, y(x))$ ima lokalni ekstrem. Iz toga slijedi sustav

$$\begin{cases} f_x + f_y y'(x) = 0 & \text{u } (x^*, y^*) \\ y'(x) = -\frac{g_x}{g_y} \end{cases} \tag{2.21}$$

Nadalje, definiramo

$$\lambda = -\frac{f_y}{g_y}, \tag{2.22}$$

to jest

$$f_y + \lambda g_y = 0. \tag{2.23}$$

Iz (2.21) i (2.22) slijedi

$$f_x + \lambda g_x = 0. \tag{2.24}$$

Iz (2.23) i (2.24) imamo (2.17).

□

Definiramo funkciju

$$L(x, \lambda, \lambda_0) = \sum_{k=0}^{m} \lambda_k f_k(x) \tag{2.25}$$

koju zovemo Lagrangeova funkcija, a koeficijente λ_k, $k = 0, 1, \ldots, m$ zovemo Lagrangeovi multiplikatori.
Sada ćemo generalizirati prethodni teorem.

Teorem 2.2.3. *(Lagrange). Neka su funkcije $f_k(x)$, $k = 0, 1, \ldots, m$, neprekidne i diferencijabilne u okolini točke x^*. Ako je x^* točka lokalnog ekstrema, tada možemo pronaći Lagrangeove multiplikatore $\lambda^* = (\lambda_1^*, \ldots, \lambda_m^*)$ i λ_0 takve da nisu svi jednaki nuli i da su zadovoljeni sljedeći uvjeti (uvjeti stacionarnosti):*

$$L_{x_i}(x^*, \lambda^*, \lambda_0^*) = 0, \quad i = 1, 2, \ldots, n. \tag{2.26}$$

Da bi osigurali da je $\lambda_0 \neq 0$ dovoljno je da su vektori

$$\nabla f_1(x^*), \nabla f_2(x^*), \ldots, \nabla f_m(x^*) \tag{2.27}$$

linearno nezavisni. Stoga, da bi pronašli stacionarnu točku moramo riješiti sustav $n + m$ jednadžbi

$$\frac{\partial}{\partial x_i}\left(\sum_{k=0}^{m} \lambda_k f_k(x)\right) = 0 \quad (n \text{ jednadžbi}, i = 1, \ldots, n) \tag{2.28}$$

$$f_1(x) = \cdots = f_m(x) = 0 \quad (m \text{ jednadžbi}) \tag{2.29}$$

sa $n + m + 1$ nepoznanica. Ako je $\lambda_0 \neq 0$ možemo sve koeficijente pomnožiti konstantom kako bi dobili $\lambda_0 = 1$. Tada je broj nepoznanica jednak broju jednadžbi.

2.3 Kuhn-Tuckerov teorem

Razmatramo problem *konveksne optimizacije* u kojem minimiziramo konveksnu funkciju cilja pod određenim konveksnim uvjetima nejednakosti.

Definicija 2.3.1. *Skup A, podskup vektorskog prostora X, je konveksan ako za svaki x i y iz A vrijedi da je segment*

$$[x, y] = \{z : \ z = \alpha x + (1 - \alpha)y, \ 0 \leq \alpha \leq 1\} \tag{2.30}$$

sadržan u A.

Definicija 2.3.2. *Funkcija f na nepraznom konveksnom skupu je konveksna ako za svaki x i y iz tog skupa vrijedi*

$$f(\alpha x + (1 - \alpha)y) \leq \alpha f(x) + (1 - \alpha) f(y), \quad 0 \leq \alpha \leq 1. \tag{2.31}$$

Neka je X vektorski prostor i A, $A \subset X$, konveksan. Neka su $f_k(x)$, $k = 0, \ldots, m$, konveksne funkcije.
Sada razmatramo problem *konveksne optimizacije*:

$$f_0(x) \to \inf \tag{2.32}$$

s ograničenjima

$$x \in A, \tag{2.33}$$

$$f_k(x) \leq 0, \quad k = 1, \ldots, m. \tag{2.34}$$

U rješavanju problema koristimo Lagrangeovu funkciju

$$L = L(x, \lambda_0, \lambda) = \sum_{k=0}^{m} \lambda_k f_k(x), \tag{2.35}$$

gdje je $\lambda = (\lambda_1, \ldots, \lambda_m)$.

Teorem 2.3.3. *(Kuhn-Tucker). Ako x^* minimizira funkciju* (2.32) *s ograničenjima* (2.33) *i* (2.34)*, tada postoje Lagrangeovi multiplikatori λ_0^* i $\lambda^* = (\lambda_1^*, \ldots, \lambda_m^*)$ takvi da nisu svi nula i da vrijedi sljedeće:*

(a) Princip minimuma:

$$\min_{x \in A} L(x, \lambda_0^*, \lambda^*) = L(x^*, \lambda_0^*, \lambda^*), \tag{2.36}$$

(b) Nenegativnost multiplikatora:

$$\lambda_k^* \geq 0, \quad k = 0, 1, \ldots, m, \tag{2.37}$$

(c) Kuhn-Tuckerovi uvjeti:

$$\lambda_k^* f_k(x^*) = 0, \quad k = 1, \ldots, m. \tag{2.38}$$

Ako je $\lambda_0 \neq 0$, tada su uvjeti (a), (b) i (c) dovoljni da x^* bude rješenje optimizacijskog problema.

Da bi $\lambda_0 \neq 0$ dovoljno je da je zadovoljen tzv. *Slaterov uvjet*, to jest, dovoljno je da postoji $\overline{x}$ takav da vrijedi

$$f_k(\overline{x}) < 0, \quad k = 1, \ldots, m. \tag{2.39}$$

Korolar 2.3.4. *Ako je zadovoljen Slaterov uvjet možemo uzeti da je $\lambda_0 = 1$ i zapisati Lagrangeovu funkciju u obliku*

$$L(x, 1, \lambda) = f_0(x) + \sum_{k=1}^{m} \lambda_k f_k(x). \tag{2.40}$$

Sada su Lagrangeova funkcija, definirana na $m + n$ varijabli, i uvjeti Kuhn-Tuckerovog teorema ekvivalentni postojanju sedlaste točke (x^*, λ^*) Lagrangeove funkcije, to jest, vrijedi

$$\min_{x \in A} L(x, 1, \lambda^*) = L(x^*, 1, \lambda^*) = \max_{\lambda > 0} L(x^*, 1, \lambda). \tag{2.41}$$

Uočimo da lijeva jednakost proizlazi iz uvjeta (a) teorema, dok desna jednakost proizlazi iz uvjeta (b) i (c).

Možemo primjetiti da u Kuhn-Tuckerovom teoremu, uvjet (a) opisuje Lagrengeovu ideju: Ako je x^* rješenje minimizacijskog problema s ograničenjima (2.33) i (2.34), tada je i minimum Lagrangeove funkcije. Uvjeti (b) i (c) su specifični za ograničenja nejednakosti.

Poglavlje 3

Metoda potpornih vektora u problemima klasifikacije

Metode za konstrukciju razdvajajućih hiperravnina iznimno su važne u konstrukciji algoritama za klasifikaciju. Temelje se na pronalasku hiperravnina u prostoru koje uspješno razdvajaju primjere iz skupa za učenje na klase. U ovom poglavlju razmotrit ćemo poseban način konstrukcije koji koristi samo određene vektore iz skupa za učenje, tzv. potporne vektore. Promatramo isključivo binarnu klasifikaciju. Više detalja možete pronaći u desetom poglavlju knjige [11].

3.1 Optimalna hiperravnina

Kažemo da su dva konačna podskupa vektora x iz skupa za učenje

$$\mathcal{D} = \{(x_i, y_i) : \ x_i \in \mathbb{R}^n, \ y_i \in \{-1, 1\}, \ i = 1, \ldots, N\} \tag{3.1}$$

gdje je jedan podskup I za koji vrijedi $y = 1$, a drugi podskup II za koji je $y = -1$, razdvojivi hiperravninom

$$\langle \phi, x \rangle = c, \tag{3.2}$$

ako postoji jedinični vektor ϕ ($||\phi||$=1) i konstanta c takvi da vrijede nejednakosti

$$\begin{aligned} \langle \phi, x_i \rangle &> c, \quad \text{za } x_i \in I, \\ \langle \phi, x_j \rangle &< c, \quad \text{za } x_j \in II. \end{aligned} \tag{3.3}$$

$||x|| = \sqrt{x_1^2 + \cdots + x_n^2}$ je Euklidska norma vektora x na $\mathbb{R}^n$, a $\langle \cdot, \cdot \rangle$ skalarni produkt induciran tom normom.

Za proizvoljni jedinični vektor ϕ određujemo dvije vrijednosti:

$$\begin{aligned} c_1(\phi) &= \min_{x_i \in I} \langle \phi, x_i \rangle, \\ c_2(\phi) &= \max_{x_j \in II} \langle \phi, x_j \rangle. \end{aligned} \tag{3.4}$$

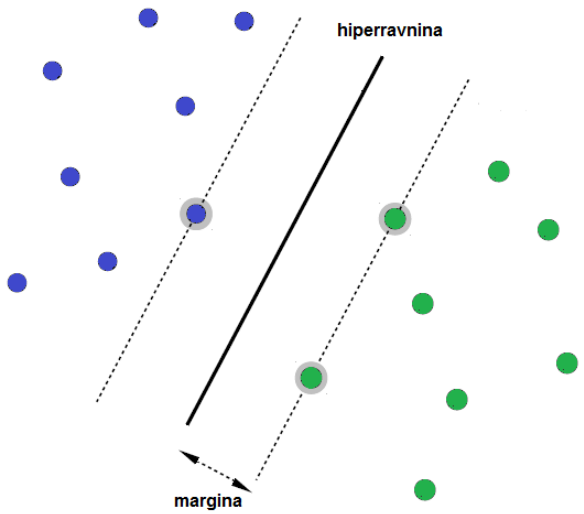
Neka je jedinični vektor ϕ_0 takav da maksimizira funkciju margine

$$\rho(\phi) = \frac{c_1(\phi) - c_2(\phi)}{2}, \quad \|\phi\| = 1, \tag{3.5}$$

pod uvjetima nejednakosti (3.3). Tada ϕ_0 i konstanta

$$c_0 = \frac{c_1(\phi_0) + c_2(\phi_0)}{2} \tag{3.6}$$

određuju hiperravninu koja razdvaja vektore iz skupa I od onih iz skupa II i posjeduje maksimalnu marginu (3.5), to jest maksimalno je udaljena od primjera iz oba podskupa. Takvu hiperravninu nazivamo *optimalna hiperravnina*.



Slika 3.1: Skica optimalne hiperravnine.

Napomena 3.1.1. *Optimalna hiperravnina je samo jedna od hiperravnina koja uspješno razdvaja linearno separabilne primjere.*

Teorem 3.1.2. *Optimalna hiperravnina je jedinstvena.*

Dokaz. Moramo pokazati da točka ϕ_0 u kojoj neprekidna funkcija $\rho(\phi)$ postiže maksimum postoji u skupu $||\phi|| \leq 1$ i da se nalazi baš na rubu $||\phi|| = 1$.

Postojanje maksimuma proizlazi iz neprekidnosti funkcije $\rho(\phi)$ na zatvorenom području $||\phi|| \leq 1$.

Pretpostavimo da se maksimum postiže u nekoj točki ϕ_0 u unutrašnjosti. Tada bi vektor

$$\phi^* = \frac{\phi_0}{||\phi_0||} \tag{3.7}$$

definirao veću marginu

$$\rho(\phi^*) = \frac{\rho(\phi_0)}{||\phi_0||}. \tag{3.8}$$

Maksimum funkcije $\rho(\phi)$ ne može se postići u dvije različite rubne točke. U protivnom, kako je funkcija $\rho(\phi)$ konveksna, maksimum bi se postigao na svakoj točki pravca koji spaja te dvije točke, a kako znamo da su to točke u unutrašnjosti za koje smo pokazali da se u njima ne mogu postizati maksimumi, dolazimo do kontradikcije. □

Kako nam je cilj pronaći učinkovitu metodu za konstrukciju optimalne hiperravnine, preformulirat ćemo problem: tražimo uređeni par vektora w_0 i konstane b_0 koji zadovoljavaju

$$\begin{aligned} \langle w_0, x_i \rangle + b_0 \geq 1, \quad &\text{za } y_i = 1, \\ \langle w_0, x_j \rangle + b_0 \leq -1, \quad &\text{za } y_j = -1, \end{aligned} \tag{3.9}$$

gdje vektor w_0 ima najmanju normu

$$||w||^2 = \langle w, w \rangle. \tag{3.10}$$

Teorem 3.1.3. *Vektor w_0 koji minimizira* (3.10) *s ograničenjima* (3.9) *povezan je s vektorom ϕ_0 koji definira optimalnu hiperravninu preko jednakosti*

$$\phi_0 = \frac{w_0}{||w_0||} \tag{3.11}$$

Margina ρ_0 između optimalne hiperravnine i razdvojenih vektora jednaka je

$$\rho(\phi_0) = \sup_{\phi_0} \frac{1}{2}\Big(\min_{i \in I} \langle \phi_0, x_i \rangle - \max_{j \in II} \langle \phi_0, x_j \rangle \Big) = \frac{1}{||w_0||}. \tag{3.12}$$

Dokaz. Pokažimo jedinstvenost vektora w_0 u kojem kvadratna funkcija (3.10) poprima minimum uz uvjete (3.9). Definirajmo jedinični vektor ϕ_0 kao u (3.11).
Uz uvjete (3.9) imamo

$$\rho\left(\frac{w_0}{\|w_0\|}\right) = \frac{1}{2}\left(c_1\left(\frac{w_0}{\|w_0\|}\right) - c_2\left(\frac{w_0}{\|w_0\|}\right)\right) \geq \frac{1}{\|w_0\|}. \tag{3.13}$$

Da bi dokazali teorem, dovoljno je pokazati da je nejednakost

$$\rho\left(\frac{w_0}{\|w_0\|}\right) > \frac{1}{\|w_0\|} \tag{3.14}$$

nemoguća. Pretpostavimo suprotno. Tada postoji jedinični vektor ϕ^* takav da vrijedi nejednakost

$$\rho(\phi^*) > \frac{1}{\|w_0\|}. \tag{3.15}$$

Definirajmo novi vektor

$$w^* = \frac{\phi^*}{\rho(\phi^*)}, \tag{3.16}$$

koji ima manju normu od $\|w_0\|$. Lako se provjeri da ovaj vektor također zadovoljava uvjete (3.9) ako je

$$b = -\frac{c_1(\phi) + c_2(\phi)}{2}. \tag{3.17}$$

To je kontradikcija s pretpostavkom da je w_0 najmanji vektor koji zadovoljava (3.9) □

Dakle, vektor w_0 s najmanjom normom koji zadovoljava uvjete (3.9) definira optimalnu hiperravninu. Ako dodatno vrijedi da je $b = 0$, optimalna hiperravnina prolazi kroz ishodište.

Radi jednostavnosti, uvjete (3.9) zapisat ćemo u ekvivalentnom obliku

$$y_i(\langle w_0, x_i\rangle + b) \geq 1, \quad i = 1, \ldots, N. \tag{3.18}$$

Stoga, da bi pronašli optimalnu hiperravninu, potrebno je minimizirati kvadratnu funkciju (3.10) tako da su zadovoljeni uvjeti (3.18).

Ovakav optimizacijski problem možemo riješiti u *primarnom prostoru* - prostor parametara w i b, ali značajnije rezultate dobit ćemo ako se prebacimo u *dualni prostor* - prostor Lagrangeovih multiplikatora koje smo definirali u prošlom poglavlju.

Kako je pokazano u prošlom poglavlju, da bi riješili problem kvadratne optimizacije, kakav imamo ovdje, moramo pronaći sedlo Lagrangeove funkcije

$$L(w, b, \alpha) = \frac{1}{2}\langle w, w\rangle - \sum_{i=1}^{l} \alpha_i(y_i[\langle w, x_i\rangle + b] - 1), \tag{3.19}$$

gdje su $\alpha_i \geq 0$ Lagrangeovi multiplikatori. Kako bi pronašli točku sedla moramo minimizirati funkciju u w i b i maksimizirati je u nenegativnim Lagrangeovim multiplikatorima $\alpha_i \geq 0$.

Prema Fermatovom teoremu, točka u kojoj se postiže minimum mora zadovoljavati

$$\partial_w L = w - \sum_{i=1}^{N} y_i \alpha_i x_i = 0, \tag{3.20}$$

$$\partial_b L = \sum_{i=1}^{N} y_i \alpha_i = 0. \tag{3.21}$$

Iz toga slijedi da za svaki vektor w koji definira optimalnu hiperravninu moraju vrijediti jednakosti

$$w = \sum_{i=1}^{N} y_i \alpha_i x_i, \tag{3.22}$$

$$\sum_{i=1}^{N} y_i \alpha_i = 0. \tag{3.23}$$

Supstitucijom (3.22) u (3.19) uz (3.23), dobivamo

$$W(\alpha) = \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{N} y_i y_j \alpha_i \alpha_j \langle x_i, x_j \rangle. \tag{3.24}$$

Promijenili smo oznaku sa $L(w, b, \alpha)$ u $W(\alpha)$ kako bi naglasili da nam je preostala jedna varijabla α takva da je

$$\alpha_i \geq 0, \quad i = 1, \ldots, N, \tag{3.25}$$

po kojoj maksimiziramo funkciju uz ograničenja (3.23). Neka je $\alpha_0 = (\alpha_1^0, \alpha_2^0, \ldots, \alpha_N^0)$ točka u kojoj se postiže maksimum. Tada dobivamo rješenje

$$w_0 = \sum_{i=1}^{N} y_i \alpha_i^0 x_i. \tag{3.26}$$

Optimalna rješenja zadovoljavaju Kuhn-Tuckerove uvjete iz prošlog poglavlja

$$\alpha_i^0 (y_i (\langle w_0, x_i \rangle + b_0) - 1) = 0, \quad i = 1, \ldots, N. \tag{3.27}$$

Iz uvjeta (3.27) zaključujemo da nenul vrijednosti α_i^0 odgovaraju onim vektorima x_i koji zadovoljaju jednakost

$$y_i (\langle w_0, x_i \rangle + b_0) = 1. \tag{3.28}$$

Geometrijski, ti vektori su najbliži optimalnoj hiperravnini. Njih nazivamo *potporni vektori*. Iz (3.26) vidimo da oni imaju ključnu ulogu u konstruiranju optimalne hiperravnine koja je oblika

$$f(x, \alpha_0) = \sum_{i=1}^{N} y_i \alpha_i^0 \langle x_i, x \rangle + b_0. \tag{3.29}$$

Neka je N_{SV} broj potpornih vektora. Tada za b_0 vrijedi

$$b_0 = \frac{1}{N_{SV}} \sum_{s=1}^{N_{SV}} (\frac{1}{y_s} - \langle w_0, x_s \rangle) = \frac{1}{N_{SV}} \sum_{s=1}^{N_{SV}} (y_s - \langle w_0, x_s \rangle) \tag{3.30}$$

Uočimo da ni razdvajajuća hiperravnina (3.29) ni funkcija cilja (3.24) ne ovise eksplicitno o dimenziji vektora x, već samo o skalarnom produktu vektora. Ova činjenica će nam kasnije omogućiti da hiperravninu konstruiramo u višedimenzionalnim prostorima (čak u beskonačnodimenzionalnim Hilbertovim prostorima).

Navedimo svojstva optimalne hiperravnine.

1. Optimalna hiperravnina je jedinstvena, to jest uređeni par vektora w_0 i konstante b_0 je jedinstven, u skupu hiperravnina za koje vrijedi $\min_{x_i \in \mathcal{X}} |\langle w, x_i \rangle + b| = 1, i = 1, \ldots, N$ (to je skup *kanonskih* hiperravnina).

2. Neka je w_0 vektor koji definira optimalnu hiperravninu. Tada je maksimum funkcionala $W(\alpha)$ dan sa

$$W(\alpha_0) = \frac{1}{2} \langle w_0, w_0 \rangle = \frac{1}{2} \sum_i \alpha_i^0. \tag{3.31}$$

To proizlazi iz jednakosti (3.23) i (3.27)

3. Norma vektora w_0 definira marginu optimalne razdvajajuće hiperravnine

$$\rho(w_0) = \frac{1}{\|w_0\|}. \tag{3.32}$$

4. Iz svojstava 2. i 3. slijedi

$$W(\alpha) < W(\alpha_0) = \frac{1}{2} \left(\frac{1}{\rho(w_0)} \right)^2, \quad \alpha \neq \alpha_0. \tag{3.33}$$

Ovaj izraz može se izabrati kao kriterij linearne neseparabilnosti dvaju skupa podataka.

Definicija 3.1.4. *Dva skupa podataka su linearno δ-neseparabilna ako je margina između hiperravnine i najbližeg vektora manja od δ.*

Stoga, ako prilikom maksimizacije vrijednost $W(\alpha)$ prijeđe $1/2\delta^2$, možemo prepostaviti da su dva skupa koja želimo razdvojiti δ-neseparabilna i da je razdvajanje s marginom δ

nemoguće.

Kada pronađemo potporne vektore možemo izračunati gornju granicu za očekivanu vjerojatnost pogreške na testnom skupu:

$$E_N(\text{vjerojatnost pogreške}) < \frac{E(\text{broj potpornih vektora})}{N},$$

gdje je E_N očekivanje za sve skupove za učenje veličine N. Uočavamo da će mali broj potpornih vektora rezultirati boljom generalizacijom. Za detalje pogledati [6].

3.2 Optimalna hiperravnina u slučaju nerazdvojivih skupova

Sada ćemo generalizirati koncept optimalne hiperravnine na slučaj nerazdvojivih skupova.

Neka je $\mathcal{D}$ iz (3.1) takav da se ne može linearno razdvojiti, odnosno ne može se razdvojiti hiperravninom na način da su svi elementi x iz skupa I na jednoj strani hiperravnine, a preostali na drugoj. Primjer linearno nerazdvojivog skupa u $\mathbb{R}^2$ vidimo na slici 3.2. Iz definicije neseparabilnosti, to znači da ne postoji par w i b takav da vrijedi

$$\langle w, w\rangle \leq \frac{1}{\rho^2} = A^2 \tag{3.34}$$

i da vrijede nejednakosti

$$y_i(\langle w, x_i\rangle + b) \geq 1, \quad i = 1, 2, \ldots, N. \tag{3.35}$$

Cilj nam je konstruirati hiperravninu sa najmanje grešaka. Kako bi to postigli, uvodimo nenegativne varijable

$$\xi_1, \ldots, \xi_N. \tag{3.36}$$

Pomoću njih formuliramo problem traženja hiperravnine sa najmanje grešaka na skupu za učenje na sljedeći način: minimiziramo funkcional

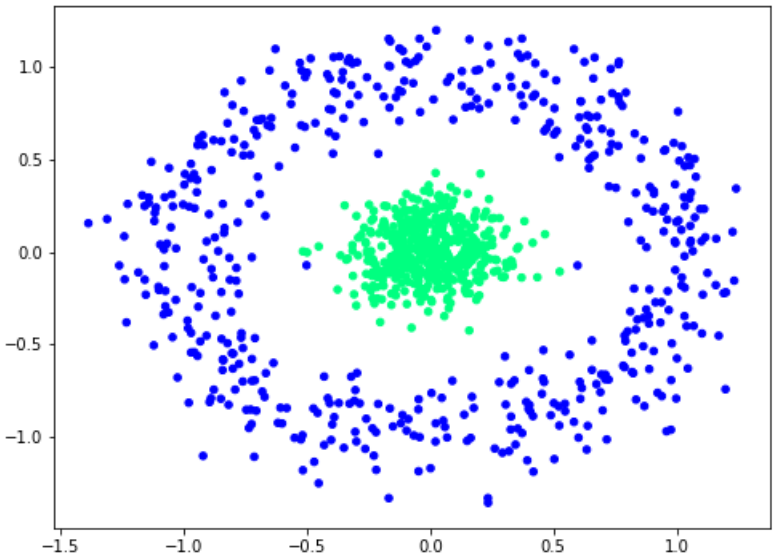
$$\Phi(\xi) = \sum_{i=1}^{N} \theta(\xi_i) \tag{3.37}$$

sa ograničenjima

$$y_i(\langle w, x_i\rangle + b) \geq 1 - \xi_i, \quad i = 1, 2, \ldots, N, \ \xi_i \geq 0 \tag{3.38}$$

i

$$\langle w, w\rangle \leq A^2, \tag{3.39}$$



Slika 3.2: Linearno nerazdvojivi skupovi u $\mathbb{R}^2$.

gdje je $\theta(\xi_i) = 0$ ako je $\xi_i = 0$, a $\theta(\xi_i) = 1$ ako je $\xi_i > 0$, $i = 1, \ldots, N$. Ovakva formulacija problema naziva se *meka margina*.
Poznato je da je za neseparabilan slučaj ovaj optimizacijski problem NP-težak, stoga razmatramo njegovu aproksimaciju: minimiziramo funkcional

$$\Phi(\xi) = \sum_{i=1}^{N} \xi_i^{\sigma} \tag{3.40}$$

s ograničenjima (3.38) i (3.39), gdje $\sigma \geq 0$ predstavlja malu vrijednost. Od sada uzimamo $\sigma = 1$, najmanji σ koji definira jednostavni optimizacijski problem.

Dakle, minimiziramo funkcional

$$\Phi(\xi, b) = \sum_{i=1}^{N} \xi_i \tag{3.41}$$

s ograničenjima (3.38) i (3.39). Hiperravninu

$$\langle w_0, x \rangle + b = 0 \tag{3.42}$$

koju smo konstruirali iz rješenja takvog optimizacijskog problema nazivamo *generalizirana optimalna hiperravnina*, ili zbog jednostavnosti, *optimalna hiperravnina*.

Kako bismo pronašli rješenje ovog optimizacijskog problema trebamo pronaći sedlo Lagrangeove funkcije

$$L(w,b,\alpha,\beta,\gamma) = \sum_{i=1}^{N}\xi_i - \frac{1}{2}\gamma(A^2 - \langle w,w\rangle) - \sum_{i=1}^{N}\alpha_i(y_i(\langle w,x_i\rangle + b) - 1 + \xi_i) - \sum_{i=1}^{N}\beta_i\xi_i \quad (3.43)$$

(minimum u varijablama w, b, ξ_i i maksimum u nenegativnim multiplikatorima α_i, β_i, γ). Parametri koji minimiziraju Lagrangeovu funkciju moraju zadovoljavati sljedeće uvjete

$$\partial_w L = \gamma w - \sum_{i=1}^{N}\alpha_i y_i x_i = 0, \quad (3.44)$$

$$\partial_b L = -\sum_{i=1}^{N} y_i\alpha_i = 0, \quad (3.45)$$

$$\partial_{\xi_i} L = 1 - \alpha_i - \beta_i = 0. \quad (3.46)$$

Iz tih uvjeta proizlazi

$$w = \frac{1}{\gamma}\sum_{i=1}^{N}\alpha_i y_i x_i, \quad (3.47)$$

$$\begin{aligned}\sum_{i=1}^{N}\alpha_i y_i &= 0,\\ \alpha_i + \beta_i &= 1.\end{aligned} \quad (3.48)$$

Supstitucijom (3.47) u Lagrangeovu funkciju uz (3.48), dolazimo do funkcionala

$$W(\alpha,\gamma) = \sum_{i=1}^{N}\alpha_i - \frac{1}{2\gamma}\sum_{i,j=1}^{N}\alpha_i\alpha_j y_i y_j\langle x_i,x_j\rangle - \frac{\gamma A^2}{2}, \quad (3.49)$$

kojeg maksimiziramo uz ograničenja

$$\sum_{i=1}^{N} y_i\alpha_i = 0, \quad (3.50)$$

$$0 \le \alpha_i \le 1, \quad (3.51)$$

$$\gamma \ge 0. \quad (3.52)$$

Maksimizacija (3.49) pod tim uvjetima može se provesti rješavajući problem kvadratne optimizacije, nekoliko puta, za fiksne vrijednosti γ, a maksimizacija s obzirom na γ preko

line search metode. Također, možemo pronaći parametar γ koji maksimizira (3.49) i supstituirati ga natrag u (3.49). Lako se provjeri da se maksimum od (3.49) postiže u

$$\gamma = \frac{\sqrt{\sum_{i,j=1}^{N} \alpha_i \alpha_j y_i y_j \langle x_i, x_j \rangle}}{A}. \tag{3.53}$$

Na taj način dolazimo do funkcionala

$$W(\alpha) = \sum_{i=1}^{N} \alpha_i - A \sqrt{\sum_{i,j=1}^{N} \alpha_i \alpha_j y_i y_j \langle x_i, x_j \rangle} \tag{3.54}$$

kojeg moramo maksimizirati s ograničenjima

$$\begin{aligned} \sum_{i=1}^{N} y_i \alpha_i &= 0, \\ 0 \leq \alpha_i &\leq 1. \end{aligned} \tag{3.55}$$

Vektor parametara $\alpha_0 = (\alpha_1^0, \ldots, \alpha_N^0)$ definira generaliziranu optimalnu hiperravninu

$$f(x) = \frac{A}{\sqrt{\sum_{i,j=1}^{N} \alpha_i^0 \alpha_j^0 y_i y_j \langle x_i, x_j \rangle}} \sum_{i=1}^{N} \alpha_i^0 y_i \langle x_i, x \rangle + b. \tag{3.56}$$

Vrijednost konstante b bira se tako da budu zadovoljeni Kuhn-Tuckerovi uvjeti

$$\alpha_t^0 \left(\frac{A}{\sqrt{\sum_{i,j=1}^{N} \alpha_i^0 \alpha_j^0 y_i y_j \langle x_i, x_j \rangle}} \sum_{i=1}^{N} \alpha_i^0 y_i \langle x_i, x_t \rangle + b \right) = 0, \quad t = 1, \ldots, N. \tag{3.57}$$

Generalizacija meke margine

Kako bismo pojednostavili računanje, možemo uvesti modificirani koncept generalizirane optimalne hiperravnine koja je određena vektorom w koji minimizira funkcional

$$\Phi(w, \xi) = \frac{1}{2} \langle w, w \rangle + C \left(\sum_{i=1}^{N} \xi_i \right) \tag{3.58}$$

pod uvjetima (3.38). Parametar $C > 0$ je unaprijed odabrana vrijednost, a određuje kompromis između veličine margine i ukupne kazne (mekoće margine). Najčešće se odabire eksperimantalnim metodama poput *unakrsne provjere*.

Koristeći istu tehniku uvođenja Lagrangeove funkcije, dolazimo do metode za rješenje koja je skoro ekvivalentna metodi za optimizacijski problem kod separabilnog slučaja. Pripadna Lagrangeova funkcija je

$$L(w,b,\xi,\alpha,\beta) = \frac{1}{2}\langle w,w\rangle + C\sum_{i=1}^{N}\xi_i - \sum_{i=1}^{N}\alpha_i(y_i(\langle w,x_i\rangle + b) - 1 + \xi_i) - \sum_{i=1}^{N}\beta_i\xi_i, \quad (3.59)$$

gdje su $\alpha_i \geq 0$ i $\beta_i \geq 0$ Lagrangeovi multiplikatori. Nakon derivacije gornjeg izraza po w, b i ξ_i dobivamo

$$w = \sum_{i=1}^{N}\alpha_i y_i x_i, \quad i = 1,\ldots,N, \quad (3.60)$$

$$\sum_{i=1}^{N}\alpha_i y_i = 0, \quad i = 1,\ldots,N, \quad (3.61)$$

$$\alpha_i = C - \beta_i, \quad i = 1,\ldots,N. \quad (3.62)$$

Iz uvjeta Kuhn-Tuckerovog teorema imamo:

$$\begin{aligned}\alpha_i(y_i(\langle w_0,x_i\rangle + b_0) - 1 + \xi_i) = 0, \quad i = 1,\ldots,N,\\ \beta_i\xi_i = (C-\alpha_i)\xi_i = 0, \quad i = 1,\ldots,N.\end{aligned} \quad (3.63)$$

Kako bismo pronašli vektor w generalizirane optimalne hiperravnine potrebno je maksimizirati isti funkcional kao u separabilnom slučaju

$$W(\alpha) = \sum_{i=1}^{N}\alpha_i - \frac{1}{2}\sum_{i,j=1}^{N}y_i y_j\alpha_i\alpha_j\langle x_i,x_j\rangle \quad (3.64)$$

uz drugačija ograničenja:

$$\begin{aligned}0 \leq \alpha_i \leq C, \quad i = 1,\ldots,N,\\ \sum_{i=1}^{N}\alpha_i y_i = 0.\end{aligned} \quad (3.65)$$

Prelaskom u dualni oblik reducirali smo broj varijabli na N.

Iz (3.63) proizlaze tri moguća rješenja za α_i:

1. $\alpha_i = \xi_i = 0$ pa su primjeri x_i točno klasificirani,

2. $C > \alpha_i > 0$, iz toga slijedi $y_i(\langle w_0, xi\rangle + b_0) - 1 + \xi_i = 0$ i $\xi_i = 0$. Stoga, $y_i(\langle w_0, xi\rangle + b_0) = 1$ i x_i je potporni vektor. Takve potporne vektore zovemo *slobodni*. Oni leže na dvijema marginama,

3. $\alpha_i = C$, iz toga slijedi $y_i(\langle w_0, xi\rangle + b_0) - 1 + \xi_i = 0$ i $\xi_i \geq 0$. x_i je potporni vektor i to *ograničeni*. Leži na "pogrešnoj" strani margine. Za $1 > \xi_i \geq 0$, x_i je i dalje točno

klasificiran, no za $\xi_i \geq 1$, x_i je pogrešno klasificiran. Detaljnije o ovome u knjizi [6].

Optimalna hiperravnina je oblika

$$\sum_{i=1}^{N} \alpha_i^0 y_i \langle x_i, x \rangle + b_0 = 0. \tag{3.66}$$

Za računanje b_0 koristimo isti izraz kao u (3.30), ali koristimo isključivo slobodne potporne vektore.

Ako je koeficijent C u funkcionalu (3.58) jednak optimalnoj vrijednosti parametra γ_0 za maksimizaciju funkcionala (3.49), tada se rješenja oba optimizacijska problema podudaraju.

Poglavlje 4

Metoda potpornih vektora u problemima regresije

U ovom poglavlju pokazat ću kakvu ulogu potporni vektori imaju u rješavanju problema regresije. Za više informacija pogledati [9].

4.1 ε-SV regresija

Neka je dan skup primjera za učenje

$$\mathcal{D} = \{(x_1, y_1), \ldots, (x_N, y_N)\} \subset \mathcal{X} \times \mathbb{R}, \tag{4.1}$$

gdje je $\mathcal{X}$ ulazni prostor, najčešće $\mathbb{R}^n$. Kod ε-SV[1] regresije cilj nam je pronaći funkciju $f(x)$ čija je udaljenost od svakog y_i, $i = 1, \ldots, N$, najviše ε i koja je najviše moguće *ravna* (u smislu najmanjeg mogućeg nagiba). Drugim riječima, greške nam nisu važne sve dok su one manje od ε.

Opći oblik linearne funkcije f dan je sa

$$f(x) = \langle w, x \rangle + b, \quad w \in \mathcal{X}, b \in \mathbb{R} \tag{4.2}$$

gdje je sa $\langle \cdot, \cdot \rangle$ dan skalarni produkt na $\mathcal{X}$. Funkcija iz (4.2) je to ravnija što joj je manji w, a jedan od načina da to postignemo je minimizacija Euklidske norme $\|w\|$. Formalno, ovaj problem možemo napisati kao problem konveksne optimizacije, slično kao u konstrukciji optimalne hiperravnine iz prethodnog poglavlja:

$$\frac{1}{2}\|w\|^2 \to \min, \tag{4.3}$$

[1] support vectors (eng. potporni vektori)

s ograničenjima

$$\begin{aligned} y_i - \langle w, x_i \rangle - b &\leq \varepsilon, \\ \langle w, x_i \rangle + b - y_i &\leq \varepsilon. \end{aligned} \tag{4.4}$$

Tu smo prepostavili da je ovaj problem konveksne optimizacije rješiv, odnosno da je moguće pronaći funkciju f koja aproksimira sve parove (x_i, y_i) sa preciznošću ε. Međutim, to nije uvijek slučaj pa je potrebno uračunati pogrešku, odnosno uvesti varijable ξ_i i ξ_i^*, $i = 1, \ldots, N$, analogno kao kod problema meke margine. Dolazimo do novog problema konveksne optimizacije:

$$\frac{1}{2}\|w\|^2 + C\sum_{i=1}^{N}(\xi_i + \xi_i^*) \to \min, \tag{4.5}$$

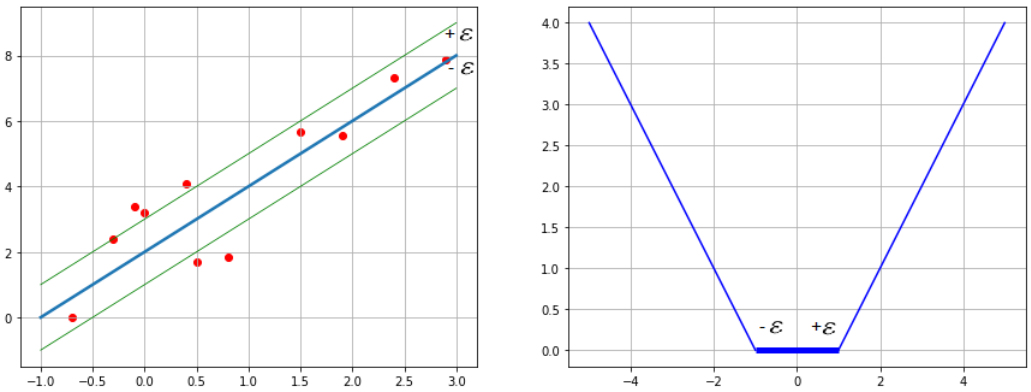
s ograničenjima

$$\begin{aligned} y_i - \langle w, x_i \rangle - b &\leq \varepsilon + \xi_i, \\ \langle w, x_i \rangle + b - y_i &\leq \varepsilon + \xi_i^*, \\ \xi_i, \xi_i^* &\geq 0. \end{aligned} \tag{4.6}$$

Parametar $C > 0$ je, kao i prije, unaprijed odabrana vrijednost (unakrsna provjera), a određuje kompromis između veličine nagiba funkcije f i veličine pogreške koju toleriramo. Primjećujemo da gornjoj formulaciji odgovara funkcija

$$|\xi|_\varepsilon := \begin{cases} 0, & \text{za } |\xi| \leq \varepsilon \\ |\xi| - \varepsilon, & \text{inače,} \end{cases} \tag{4.7}$$

koju nazivamo *ε-robusna funkcija gubitka.*



Slika 4.1: ε-SV regresija i ε-robusna funkcija gubitka

Dualna formulacija

Kao i u slučaju konstrukcije optimalne hiperravnine, problem je lakše riješiti u dualnoj formulaciji. Kasnije ćemo vidjeti da je upravo to ključ za pronalazak nelinearnih rješenja. Definiramo Lagrangeovu funkciju

$$\begin{aligned} L =& \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{N}(\xi_i + \xi_i^*) - \sum_{i=1}^{N}\alpha_i(\varepsilon + \xi_i - y_i + \langle w, x_i\rangle + b) - \\ &- \sum_{i=1}^{N}\alpha_i^*(\varepsilon + \xi_i^* + y_i - \langle w, x_i\rangle - b) - \sum_{i=1}^{N}(\beta_i\xi_i + \beta_i^*\xi_i^*). \end{aligned} \tag{4.8}$$

Dualne varijable iz (4.8) moraju biti nenegativne $\alpha_i, \alpha_i^*, \beta_i, \beta_i^* \geq 0$. Moramo pronaći točku sedla Lagrangeove funkcije u primarnim varijablama:

$$\partial_b L = \sum_{i=1}^{N}(\alpha_i^* - \alpha_i) = 0, \tag{4.9}$$

$$\partial_w L = w - \sum_{i=1}^{N}(\alpha_i - \alpha_i^*)x_i = 0, \tag{4.10}$$

$$\partial_{\xi_i^{(*)}} L = C - \alpha_i^{(*)} - \beta_i^{(*)} = 0. \tag{4.11}$$

Supstitucijom (4.9), (4.10) i (4.11) u (4.8) dobivamo dualni optimizacijski problem:

$$\begin{aligned} W =& -\frac{1}{2}\sum_{i,j=1}^{N}(\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*)\langle x_i, x_j\rangle - \\ &-\varepsilon\sum_{i=1}^{N}(\alpha_i + \alpha_i^*) + \sum_{i=1}^{N} y_i(\alpha_i - \alpha_i^*) \to \max, \end{aligned} \tag{4.12}$$

s ograničenjima

$$\begin{aligned} \sum_{i=1}^{N}(\alpha_i - \alpha_i^*) = 0, \\ \alpha_i, \alpha_i^* \in [0, C]. \end{aligned} \tag{4.13}$$

Varijable β_i i β_i^* smo eliminirali pomoću uvjeta (4.11). Iz gornje formulacije problema proizlazi da je funkcija f sada dana sa

$$f(x) = \sum_{i=1}^{N}(\alpha_i - \alpha_i^*)\langle x_i, x\rangle + b. \tag{4.14}$$

Za w vrijedi

$$w = \sum_{i=1}^{N} (\alpha_i - \alpha_i^*) x_i. \tag{4.15}$$

Iz uvjeta Kuhn-Tuckerovog teorema znamo da vrijedi

$$\begin{aligned} \alpha_i(\varepsilon + \xi_i - y_i + \langle w, x_i \rangle + b) = 0, \\ \alpha_i^*(\varepsilon + \xi_i^* + y_i - \langle w, x_i \rangle - b) = 0, \end{aligned} \tag{4.16}$$

i

$$\begin{aligned} (C - \alpha_i)\xi_i = 0, \\ (C - \alpha_i^*)\xi_i^* = 0. \end{aligned} \tag{4.17}$$

Iz gornjih uvjeta možemo donijeti nekoliko važnih zaključaka. Uočavamo da se samo oni primjeri (x_i, y_i) za koje vrijedi $\alpha_i^{(*)} = C$ nalaze izvan ε-okoline funkcije f, odnosno imaju grešku $\xi_i^{(*)}$ različitu od nule. Također, vrijedi $\alpha_i \alpha_i^* = 0$, to jest ne postoje primjeri za koje su i α_i i α_i^* različiti od nula jer bi u protivnome obje varijable greške bile nenula. Za $\alpha_i^{(*)} \in (0, C)$ imamo $\xi_i^{(*)} = 0$ i drugi faktor u (4.16) mora biti nula. Iz toga proizlazi da za b vrijedi:

$$\begin{aligned} b = y_i - \langle w, x_i \rangle - \varepsilon, \quad \text{za } \alpha_i \in (0, C), \\ b = y_i - \langle w, x_i \rangle + \varepsilon, \quad \text{za } \alpha_i^* \in (0, C). \end{aligned} \tag{4.18}$$

Kako za sve primjere strogo unutar ε-okoline funkcije f α_i i α_i^* nestaju, odnosno drugi faktor u (4.16) je različit od nule, oni ne sudjeluju u konstrukciji funkcije f. Preostale vektore, kao i u slučaju klasifikacije, nazivamo *potporni vektori*.

4.2 Funkcija rizika

Na samom početku predstavit ću osnovne pojmove iz teorije vjerojatnosti. Za više informacija o tome pogledati [8]. Za više informacija o funkciji rizika pogledati [9].

Osnovni pojmovi iz teorije vjerojatnosti

Definicija 4.2.1. *Neka je $(\Omega, \mathcal{F}, \mathbb{P})$ vjerojatnosni prostor. Funkcija $X : \Omega \to \mathbb{R}$ je diskretna slučajna varijabla na $(\Omega, \mathcal{F}, \mathbb{P})$ ako postoji prebrojiv skup $\mathcal{S} = \{a_1, a_2, \dots\} \subset \mathbb{R}$ takav da je*

(a) $X(\omega) \in \mathcal{S}$ za sve $\omega \in \Omega$;

(b) $\{X = a_j\} = \{\omega \in \Omega : X(\omega) = a_j\} \in \mathcal{F}$ za sve $j \in \mathbb{N}$.

Definicija 4.2.2. *Neka je $(\Omega, \mathcal{F}, \mathbb{P})$ vjerojatnosni prostor. Funkcija $X : \Omega \to \mathbb{R}$ je slučajna varijabla na $(\Omega, \mathcal{F}, \mathbb{P})$ ako vrijedi*

$$\{X \leq x\} = \{\omega \in \Omega : X(\omega) \leq x\} \in \mathcal{F}, \quad \forall x \in \mathbb{R}.$$

Definicija 4.2.3. *Neka je* $(\Omega, \mathcal{F}, \mathbb{P})$ *vjerojatnosni prostor i neka je* $X : \Omega \to \mathcal{S} = \{a_1, a_2, \dots\} \subset \mathbb{R}$ *diskretna slučajna varijabla. Funkcija* $p : \mathbb{R} \to [0, 1]$ *definirana sa*

$$p(x) = \mathbb{P}(X = x) = \mathbb{P}(\{\omega \in \Omega : X(\omega) = x\})$$

zove se vjerojatnosna funkcija gustoće slučajne varijable X.

Definicija 4.2.4. *Neka je X slučajna varijabla definirana na vjerojatnosnom prostoru* $(\Omega, \mathcal{F}, \mathbb{P})$*. Funkcija distribucije od X je funkcija* $P : \mathbb{R} \to [0, 1]$ *definirana sa*

$$P(x) = \mathbb{P}(X \leq x), \quad x \in \mathbb{R}.$$

Napomena 4.2.5. *Vjerojatnost da slučajna varijabla X poprimi vrijednost x i da slučajna varijabla Y poprimi vrijednost y pišemo kao* $P(X = x, Y = y)$*, odnosno kraće* $P(x, y)$*. Uvjetna vjerojatnost* $P(y|x)$*, odnosno vjerojatnost da varijabla Y poprimi vrijednost y, pod uvjetom da je varijabla X poprimila vjerojatnost x definirana je kao*

$$P(y|x) = \frac{P(x, y)}{P(x)}.$$

Napomena 4.2.6. *Dvije slučajne varijable X i Y su nezavisne akko za sve intervale A i B,* $A, B \subseteq \mathbb{R}$*, vrijedi*

$$P(X \in A, Y \in B) = P(X \in A)P(Y \in B).$$

Općenito o funkciji rizika

Postavimo regresijski problem na drugačiji način. Pretpostavimo da su svi primjeri iz skupa za učenje $\mathcal{D}$ iz (4.1) uzorkovani nezavisno i iz iste zajedničke distribucije $P(x, y)$. Ta se pretpostavka označava sa iid. Funkciju f ćemo tražiti tako da minimizira *funkciju rizika*

$$R(f) := \int c(x, y, f(x)) dp(x, y). \tag{4.19}$$

$c(x, y, f(x))$ označava *funkciju troška* koja određuje koliko ćemo kažnjavati pogreške s obzirom na dani skup podataka $\mathcal{D}$. Kako problem ne znamo riješiti egzaktno, koristit ćemo skup podataka $\mathcal{D}$, koji nam je već poznat, kako bi aproksimirali traženo rješenje. Integraciju zamijenjujemo empirijskom procjenom i dolazimo do problema minimizacije *empirijske* funkcije rizika

$$R_{\text{emp}}(f) := \frac{1}{N} \sum_{i=1}^{N} c(x_i, y_i, f(x_i)). \tag{4.20}$$

U slučaju kada imamo mali broj primjera u ulaznom prostoru sa velikom dimenzijom, minimum empirijske funkcije rizika može dovesti do prenaučenosti i loše generalizacije.

Iz tog razloga potrebno je dodati izraz koji će kontrolirati takve pojave. To nas dovodi do *regularizirane* funkcije rizika

$$R_{\text{reg}}(f) := R_{\text{emp}}(f) + \frac{\lambda}{2}\|w\|^2, \tag{4.21}$$

gdje je $\lambda > 0$ *regularizacijski* faktor.

Napomena 4.2.7. *Regularizacija je metoda kojom se sprječava prenaučenost modela, a sastoji se od toga da se u funkciju koja određuje pogrešku direktno ugradi mjera složenosti modela. Na taj se način zapravo sprječava pretjerana složenost jer s njom raste i ukupna pogreška. Regularizacijski izraz općenito je oblika*

$$\frac{\lambda}{2}\sum_{k=1}^{n}|w_k|^q,$$

gdje je λ regularizacijski faktor. Što je λ veći, to se više kažnjavaju složeni modeli. Najčešće uzimamo $q = 2$ zato što je analitički najpogodnije. To se naziva L_2-regularizacija. Za $q = 1$ imamo L_1-regularizaciju.

Postavlja se pitanje kako odabrati funkciju troška $c(x, y, f(x))$. U slučaju ε-SV regresije imali smo

$$c(x, y, f(x)) = |y - f(x)|_\varepsilon \tag{4.22}$$

pa je minimizacija (4.21) s funkcijom troška (4.22) ekvivalentna minimizaciji (4.5) ($C = 1/\lambda N$). Nastoji se, općenito, izbjeći korištenje komplicirane funkcije troška jer može dovesti do zahtjevnih optimizacijskih problema. Također, cilj nam je zadržati konveksnost kako bi mogli koristiti poznate rezultate iz teorije optimizacije koji se oslanjaju na to svojstvo. Idealno je koristiti funkciju troška koja najbolje odgovara skupu podataka s kojim radimo.

Ako pretpostavimo da za podatke iz uzorka vrijedi

$$y_i = g(x_i) + \xi_i, \quad i = 1, \ldots, N, \tag{4.23}$$

gdje je g neka funkcija, a ξ_i šum i ako je p gustoća vjerojatnosti šuma, optimalna funkcija troška u smislu maksimalne izglednosti bila bi

$$c(x, y, f(x)) = -\log p(y - f(x)). \tag{4.24}$$

Izglednost skupa podataka

$$\mathcal{D}_f := \{(x_1, f(x_1)), \ldots, (x_N, f(x_N))\} \tag{4.25}$$

	Funkcija troška	Funkcija gustoće
ε-robusna	$c(\xi) = \lvert\xi\rvert_\varepsilon$	$p(\xi) = \frac{1}{2(1+\varepsilon)} \exp(-\lvert\xi\rvert_\varepsilon)$
Laplaceova	$c(\xi) = \lvert\xi\rvert$	$p(\xi) = \frac{1}{2} \exp(-\lvert\xi\rvert)$
Gaussova	$c(\xi) = \frac{1}{2}\xi^2$	$p(\xi) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{\xi^2}{2})$
Hubertova	$c(\xi) = \begin{cases} \frac{1}{2\sigma}(\xi)^2, & \text{za } \lvert\xi\rvert \le \sigma \\ \lvert\xi\rvert - \frac{\sigma}{2}, & \text{inače} \end{cases}$	$p(\xi) \propto \begin{cases} \exp(-\frac{\xi^2}{2\sigma}), & \text{za } \lvert\xi\rvert \le \sigma \\ \exp(\frac{\sigma}{2} - \lvert\xi\rvert), & \text{inače} \end{cases}$
polinomijalna	$c(\xi) = \frac{1}{p}\lvert\xi\rvert^p$	$p(\xi) = \frac{p}{2\Gamma(1/p)} \exp(-\lvert\xi\rvert^p)$
po dijelovima polinomijalna	$c(\xi) = \begin{cases} \frac{1}{p\sigma^{p-1}}(\xi)^p, & \text{za } \lvert\xi\rvert \le \sigma \\ \lvert\xi\rvert - \sigma\frac{p-1}{p}, & \text{inače} \end{cases}$	$p(\xi) \propto \begin{cases} \exp(-\frac{\xi^p}{p\sigma^{p-1}}), & \text{za } \lvert\xi\rvert \le \sigma \\ \exp(\sigma\frac{p-1}{p} - \lvert\xi\rvert), & \text{inače} \end{cases}$

Tablica 4.1: Poznate funkcije troška i odgovarajuće gustoće

pod pretpostavkom da postoji šum i da su podaci iid je

$$P(\mathcal{D}_f|\mathcal{D}) = \prod_{i=1}^{N} P(f(x_i)|(x_i, y_i)) = \prod_{i=1}^{N} P(f(x_i)|y_i) = \prod_{i=1}^{N} p(y_i - f(x_i)). \tag{4.26}$$

Maksimizacija $P(\mathcal{D}_f|\mathcal{D})$ je ekvivalentna minimizaciji $-\log P(\mathcal{D}_f|\mathcal{D})$, pa iz (4.24) dobivamo

$$-\log P(\mathcal{D}_f|\mathcal{D}) = \sum_{i=1}^{N} c(x_i, y_i, f(x_i)). \tag{4.27}$$

No, na gore opisan način možemo doći do funkcije troška koja nije konveksna, pa joj stoga moramo naći odgovarajuću konveksnu zamjenu.

Rješavanje jednadžbi

Zbog jednostavnosti dodatno ćemo pretpostaviti da je c simetrična i da ima najviše dva prekida u $\pm\varepsilon, \varepsilon \ge 0$ u prvoj derivaciji i da je nula na intervalu $[-\varepsilon, \varepsilon]$. Sve funkcije troška iz tablice 4.1 su takvog oblika. c je, dakle, oblika

$$c(x, y, f(x)) = \begin{cases} 0, & \text{za } |y - f(x)| \le \varepsilon \\ \tilde{c}(|y - f(x)| - \varepsilon), & \text{inače.} \end{cases} \tag{4.28}$$

Uočavamo sličnost sa ε-robusnim gubitkom. Ovaj izbor možemo generalizirati na općenitije konveksne funkcije troška. Za funkcije koje nisu nula na $[-\varepsilon, \varepsilon]$ uvodimo dodatne varijable. Za svaki uzorak možemo izabrati različite funkcije troška $\tilde{c}_i, \tilde{c}_i^*$ i različite $\varepsilon_i, \varepsilon_i^*$.

Također, možemo dodati više Lagrangeovih multiplikatora u slučaju dodatnih prekida.

Analogno kao u (4.5) dolazimo do problema konveksne optimizacije. Držat ćemo se iste notacije i koristit ćemo konstantu C (umjesto normalizacije sa λ i N) radi jednostavnosti. Računamo:

$$\frac{1}{2}\|w\|^2 + C\sum_{i=1}^{N}(\tilde{c}(\xi_i) + \tilde{c}(\xi_i^*)) \to \min, \tag{4.29}$$

s ograničenjima

$$\begin{aligned} y_i - \langle w, x_i\rangle - b &\leq \varepsilon + \xi_i, \\ \langle w, x_i\rangle + b - y_i &\leq \varepsilon + \xi_i^*, \\ \xi_i, \xi_i^* &\geq 0. \end{aligned} \tag{4.30}$$

Koristimo Lagrangeovu funkciju i dolazimo do problema dualne optimizacije:

$$\begin{aligned} W = -\frac{1}{2}\sum_{i,j=1}^{N}(\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*)\langle x_i, x_j\rangle + \\ + \sum_{i=1}^{N}(y_i(\alpha_i - \alpha_i^*) - \varepsilon(\alpha_i + \alpha_i^*) + \\ + C(T(\xi_i) + T(\xi_i^*))) \to \max, \end{aligned} \tag{4.31}$$

gdje je

$$\begin{aligned} w &= \sum_{i=1}^{N}(\alpha_i - \alpha_i^*)x_i, \\ T(\xi) &:= \tilde{c}(\xi) - \xi\partial_\xi\tilde{c}(\xi), \end{aligned} \tag{4.32}$$

s ograničenjima

$$\begin{cases} \sum_{i=1}^{N}(\alpha_i - \alpha_i^*) & = 0, \\ \alpha & \leq C\partial_\xi\tilde{c}(\xi), \\ \xi & = \inf\{\xi | C\partial_\xi\tilde{c} \geq \alpha\}, \\ \alpha, \xi & \geq 0. \end{cases} \tag{4.33}$$

Primjeri

Pokažimo gornji račun na primjerima iz tablice 4.1. Konkretno, na dva primjera ćemo pokazati kako se (4.31), (4.32) i (4.33) mogu dalje pojednostaviti do oblika koji je primjenjiv u praksi. U slučaju ε-robusne funkcije, to jest $\tilde{c}(\xi) = |\xi|$ imamo

$$T(\xi) = \xi - \xi \cdot 1 = 0. \tag{4.34}$$

	ε	α	$CT(\alpha)$
ε-robusna	$\varepsilon \neq 0$	$\alpha \in [0, C]$	$CT(\alpha) = 0$
Laplaceova	$\varepsilon = 0$	$\alpha \in [0, C]$	$CT(\alpha) = 0$
Gaussova	$\varepsilon = 0$	$\alpha \in [0, \infty)$	$CT(\alpha) = -\frac{1}{2}C^{-1}\alpha^2$
Hubertova	$\varepsilon = 0$	$\alpha \in [0, C]$	$CT(\alpha) = -\frac{1}{2}\sigma C^{-1}\alpha^2$
polinomijalna	$\varepsilon = 0$	$\alpha \in [0, \infty)$	$CT(\alpha) = -\frac{p-1}{p}C^{-\frac{1}{p-1}}\alpha^{\frac{p}{p-1}}$
po dijelovima polinomijalna	$\varepsilon = 0$	$\alpha \in [0, C]$	$CT(\alpha) = -\frac{p-1}{p}\sigma C^{-\frac{1}{p-1}}\alpha^{\frac{p}{p-1}}$

Tablica 4.2: Vrijednosti izraza u konveksnoj optimizaciji u ovisnosti o izboru funkcije gubitka

Iz $\partial_\xi \tilde{c}(\xi) = 1$ zaključujemo

$$\xi = \inf\{\xi | C \geq \alpha\} = 0 \text{ i stoga } \alpha \in [0, C]. \tag{4.35}$$

U slučaju po dijelovima polinomijalne funkcije gubitka razlikujemo dva različita slučaja: $\xi \leq \sigma$ i $\xi > \sigma$. U prvom slučaju imamo

$$T(\xi) = \frac{1}{p\sigma^{p-1}}\xi^p - \frac{1}{\sigma^{p-1}}\xi^p = -\frac{p-1}{p}\sigma^{1-p}\xi^p \tag{4.36}$$

i $\xi = \{\xi | C\sigma^{1-p}\xi^{p-1} \geq \alpha\} = \sigma C^{-\frac{1}{p-1}}\alpha^{\frac{1}{p-1}}$ i stoga

$$T(\xi) = -\frac{p-1}{p}\sigma C^{-\frac{p}{p-1}}\alpha^{\frac{p}{p-1}}. \tag{4.37}$$

U drugom slučaju ($\xi \geq \sigma$) imamo

$$T(\xi) = \xi - \sigma\frac{p-1}{p} - \xi = -\sigma\frac{p-1}{p} \tag{4.38}$$

i

$$\xi = \inf\{\xi | C \geq \alpha\} = \sigma \text{ i stoga } \alpha \in [0, C]. \tag{4.39}$$

Ta dva slučaja možemo sjediniti u

$$\alpha \in [0, C] \text{ i } T(\alpha) = -\frac{p-1}{p}\sigma C^{-\frac{p}{p-1}}\alpha^{\frac{p}{p-1}}. \tag{4.40}$$

Uočavamo da maksimalni nagib funkcije $\tilde{c}$ određuje vrijednosti koje α može postići. U slučaju $s := \sup_{\xi \in \mathbb{R}^+} \partial_\xi \tilde{c}(\xi) < \infty$ dobivamo kompaktne intervale $[0, Cs]$ za α. To znači da je utjecaj jednog primjera ograničen, što vodi do robusnijih procjenitelja.

Važno je napomenuti da ćemo u slučajevima kada je $\varepsilon = 0$ izgubiti prednost rijetke dekompozicije, što je problematično u slučajevima kad imamo veliki skup podataka jer znatno usporava proces predikcije.

Poglavlje 5

Nelinearni rezultati korištenjem jezgri

U ovom poglavlju razmatramo postupak kojim metoda potpornih vektora daje nelinearne rezultate, što je pogodno u klasifikaciji kada primjeri nisu linearno odvojivi i regresiji u slučaju da funkcija koju tražimo nije linearna. Nelinearnost se, na primjer, može postići tako da se primjeri za učenje preslikaju u neki prostor veće dimenzije kojeg nazivamo *prostor značajki* i označavamo sa $\mathcal{F}$:

$$\Phi : \mathcal{X} \to \mathcal{F}.$$

Cilj nam iskoristiti dosadašnje metode pronalaska rješenja u prostoru značajki, što će rezultirati nelinearnim rješenjima u ulaznom prostoru.

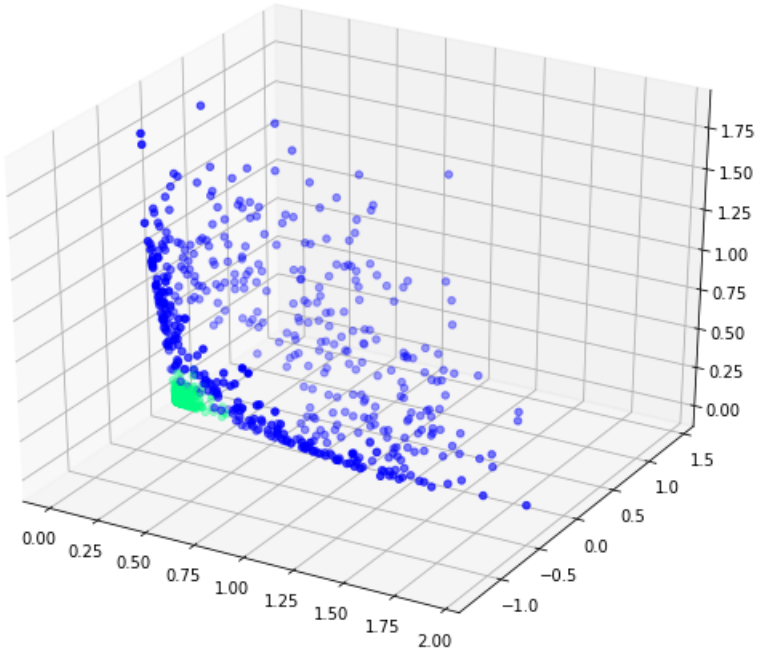
Primjer 5.0.1. *Preslikavanje* $\Phi : \mathbb{R}^2 \to \mathbb{R}^3$ *dano sa*

$$\Phi(x_1, x_2) = (x_1^2,\ \sqrt{2}x_1x_2, x_2^2) \tag{5.1}$$

je primjer preslikavanja sa ulaznog prostora R^2 *u prostor značajki* R^3*. Indeksi elemenata u ovom slučaju označavaju komponente u vektoru* $x \in \mathbb{R}^2$*. Učenje (linearnom) metodom potpornih vektora na podacima koji su prethodno prebačeni u prostor značajki na ovaj način rezultira kvadratnom funkcijom. Na slici 5.1 je prikazano preslikavanje pomoću* Φ *primjera sa slike 3.2.*

Preslikavanje poput ovoga iz primjera 5.0.1 može postati računalno zahtjevno ako se radi o preslikavanju u značajke polinomijalnog tipa velikog stupnja i u prostor značajki velike dimenzije jer je broj različitih značajki dan sa $\binom{d+p+1}{p}$, gdje je d dimenzija prostora, a p stupanj polinoma.

Za više informacija o korištenju jezgri pogledati [9].



Slika 5.1: Primjeri sa slike 3.2 preslikani pomoću preslikavanja iz primjera 5.0.1.

5.1 Implicitno preslikavanje korištenjem jezgri

Na primjeru 5.0.1 uočavamo da vrijedi:

$$\langle \Phi(x), \Phi(x') \rangle = \left\langle \left(x_1^2,\ \sqrt{2}x_1x_2, x_2^2\right), \left(x_1'^2,\ \sqrt{2}x_1'x_2', x_2'^2\right) \right\rangle = \langle x, x' \rangle^2. \tag{5.2}$$

Kako znamo da će se u dualnome problemu metode potpornih vektora preslikani vektori $\Phi(x)$ uvijek pojavljivati u obliku skalarnog produkta, to nam omogućava uvođenje tzv. *jezgrenog trika*. Umjesto eksplicitnog preslikavanja $\Phi(\cdot)$ uvodimo funkciju

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

koju nazivamo *jezgrenom funkcijom*. To nam omogućava da regresijski optimizacijski problem napišemo u obliku:

$$\begin{aligned} -\frac{1}{2}\sum_{i,j=1}^{N}(\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*)k(x_i, x_j) - \\ -\varepsilon\sum_{i=1}^{N}(\alpha_i + \alpha_i^*) + \sum_{i=1}^{N} y_i(\alpha_i - \alpha_i^*) \to \max \end{aligned} \tag{5.3}$$

s ograničenjima

$$\begin{aligned} \sum_{i=1}^{N}(\alpha_i - \alpha_i^*) &= 0, \\ \alpha_i, \alpha_i^* &\in [0, C]. \end{aligned} \tag{5.4}$$

Za w i f vrijedi

$$w = \sum_{i=1}^{N}(\alpha_i - \alpha_i^*)\Phi(x_i), \tag{5.5}$$

$$f(x) = \sum_{i=1}^{N}(\alpha_i - \alpha_i^*)k(x_i, x) + b. \tag{5.6}$$

Analogno, u klasifikaciji, u optimizacijskom problemu, skalarni produkt zamijenimo jezgrenom funkcijom k. U tom slučaju je optimalna hiperravnina iz (3.66) oblika

$$\sum_{i=1}^{N}\alpha_i^0 y_i k(x_i, x) + b_0 = 0. \tag{5.7}$$

Uočavamo da korištenjem jezgri w više nije eksplicitno dana.

U nelinearnom slučaju, optimizacijski problem odgovara pronalaženju funkcije s najmanjim mogućim nagibom u prostoru značajki, a ne u ulaznom prostoru.

5.2 Uvjeti za jezgre koje koristimo u metodi potpornih vektora

Postavlja se pitanje koje funkcije $k(x, x')$ odgovaraju skalarnom produktu u nekom prostoru značajki $\mathcal{F}$.

Neka je $\Omega \subset \mathbb{R}^n$ otvoren skup opskrbljen Lebesgueovom mjerom nasljeđenom iz $\mathbb{R}^n$.

Definicija 5.2.1. *Za $p \in \mathbb{R}$, $1 \leq p < \infty$, definiramo*

$$L^p(\Omega) = \{f : \Omega \to \mathbb{R} : \ f \text{ je izmjeriva i } \int_\Omega |f(x)|^p dx < \infty\}.$$

$L^p(\Omega)$ je normiran prostor s normom $\|f\|_{L^p(\Omega)} = \left(\int_\Omega |f(x)|^p dx\right)^{\frac{1}{p}}$.

Definicija 5.2.2.

$$L^\infty(\Omega) = \{f : \Omega \to \mathbb{R} : \ f \text{ je izmjeriva i postoji konstanta } C \text{ tdj. } |f(x)| \leq C \text{ s.s}\}$$

$L^\infty(\Omega)$ je normiran prostor s normom $\|f\|_{L^\infty(\Omega)} = \inf\{C : |f(x)| \leq C \text{ s.s}\}$.

Definicija 5.2.3. $\ell^1 = \{(x_n)_n \in \mathbb{R}^{\mathbb{N}} : \sum_{n=1}^{\infty} |x_n| < \infty\}$

ℓ^1 je normirani prostor s normom $\|x\|_1 = \sum_{n=1}^{\infty} |x_n|$.
Za više detalja o prethodnim definicijama pogledati [2].

Teorem 5.2.4. *(Mercer). Neka je $k \in L^\infty(\mathcal{X}^2)$ takav da je operator $T_k : L^2(\mathcal{X}) \to L^2(\mathcal{X})$,*

$$T_k f(\cdot) := \int_{\mathcal{X}} k(\cdot, x) f(x) d\mu(x) \tag{5.8}$$

pozitivno definitan. Neka je $\psi_j \in L^2(\mathcal{X})$ svojstvena funkcija od T_k koja odgovara svojstvenoj vrijednosti $\lambda_j \neq 0$ i neka je $\|\psi_j\|_{L^2} = 1$. Sa $\overline{\psi_j}$ je dana njena kompleksno konjugirana vrijednost. Tada

(a) $(\lambda_j(T))_j \in \ell^1$,

(b) $\psi_j \in L^\infty(\mathcal{X})$ i $\sup_j \|\psi_j\|_{L^\infty} < \infty$,

(c) $k(x, x') = \sum_{j \in \mathbb{N}} \lambda_j \overline{\psi_j(x)} \psi_j(x')$ vrijedi za skoro sve (x, x') gdje niz konvergira apsolutno i uniformno za skoro sve (x, x').

Drugim riječima, teorem znači da ako

$$\int_{\mathcal{X} \times \mathcal{X}} k(x, x') f(x) f(x') dx dx' \geq 0, \quad \forall f \in L^2(\mathcal{X}), \tag{5.9}$$

$k(x, x')$ predstavlja skalarni produkt u nekom prostoru značajki. Iz toga proizlaze neka svojstva koja imaju jezgre koje zadovoljavaju Mercerovo svojstvo. Takve jezgre nazivamo SV prikladne jezgre. Za više detalja o Mercerovom teoremu pogledati [10].

Korolar 5.2.5. *(Linearna kombinacija jezgri). Neka su $k_1(x, x')$ i $k_2(x, x')$ SV prikladne jezgre i $c_1, c_2 \geq 0$, tada je i*

$$k(x, x') := c_1 k_1(x, x') + c_2 k_2(x, x') \tag{5.10}$$

SV prikladna jezgra.

Korolar slijedi direktno iz svojstva linearnosti integrala.

Korolar 5.2.6. *(Integral jezgre). Neka je $s(x, x')$ funkcija simetrična u argumentima na $\mathcal{X} \times \mathcal{X}$, tada je*

$$k(x, x') := \int_{\mathcal{X}} s(x, z) s(x', z) dz \tag{5.11}$$

SV prikladna jezgra.

Korolar slijedi iz (5.9) i (5.11) zamjenom poretka integracije.
Za jezgre koje su tipa skalarnog produkta , to jest $k(x, x') = k(\langle x, x' \rangle)$ postoje dovoljni uvjeti da bi bile *SV prikladne jezgre.*

Teorem 5.2.7. *Svaka jezgra koja je tipa skalarnog produkta* $k(x, x') = k(\langle x, x' \rangle)$ *mora zadovoljavati*

$$k(\xi) \geq 0 \tag{5.12}$$

$$\partial_\xi k(\xi) \geq 0 \tag{5.13}$$

$$\partial_\xi k(\xi) + \xi \partial_\xi^2 k(\xi) \geq 0 \tag{5.14}$$

za svaki $\xi \geq 0$ *kako bi bila SV prikladna jezgra.*

Uvjeti iz teorema 5.2.7 su nužni ali ne i dovoljni.
Iz skupa primjera za učenje $\mathcal{D}$ iz (4.1) možemo konstruirati simetričnu $N \times N$ matricu

$$K = \begin{pmatrix} k(x_1, x_1) & k(x_1, x_2) & \dots & k(x_1, x_N) \\ k(x_2, x_1) & k(x_2, x_2) & \dots & k(x_2, x_N) \\ \vdots & \vdots & \ddots & \vdots \\ k(x_N, x_1) & k(x_N, x_2) & \dots & k(x_N, x_N) \end{pmatrix}$$

koju nazivamo *jezgrena matrica*.
Prema Mercerovom teoremu, ako je K pozitivno semidefinitna, to jest $\forall x \neq 0, x^{\mathrm{T}} K x \geq 0$ za svaki $\mathcal{D}$, tada k zadovoljava Mercerovo svojstvo.

5.3 Primjeri jezgri koje koristimo u metodi potpornih vektora

Eksplicitnim računanjem preslikavanja može se pokazati da su homogene polinomijalne jezgre

$$k(x, x') = \langle x, x' \rangle^p, \tag{5.15}$$

gdje je $p \in \mathbb{N}$, SV prikladne jezgre. Iz toga možemo zaključiti da su i nehomogene polinomijalne jezgre

$$k(x, x') = (\langle x, x' \rangle + c)^p, \tag{5.16}$$

gdje je $p \in \mathbb{N}, c > 0$, također SV prikladne jezgre. To se može vidjeti ako k napišemo kao sumu homogenih jezgri i primjenimo korolar 5.2.5.
Još jedna jezgra koja se često koristi u praksi je jezgra tangensa hiperbolnog

$$k(x, x') = \tanh(\vartheta + \phi \langle x, x' \rangle). \tag{5.17}$$

Iz teorema 5.2.7 vidi se da za $\vartheta < 0$ ili $\phi < 0$ jezgra ne zadovoljava Mercerovo svojstvo. Jedan od najvažnijih primjera su *radijalne bazne funkcije* oblika

$$k(x, x') = k(\|x - x'\|),$$

koje ovise isključivo o udaljenosti među primjerima. *Gaussova jezgra*

$$k(x, x') = \exp\Big\{-\frac{\|x - x'\|^2}{2\sigma^2}\Big\} = \exp\{-\gamma\|x - x'\|^2\}, \tag{5.18}$$

poseban je slučaj radijalne bazne funkcije. Parametar $\gamma = 1/2\sigma^2$ predstavlja preciznost. Na temelju udaljenosti primjera u ulaznome prostoru Gaussova jezgra određuje sličnost u prostoru značajki. Kako se za slične primjere $k(x, x')$ približava 1, ti primjeri su blizu i u prostoru značajki. S druge strane, velika udaljenost primjera u ulaznom prostoru rezultira ortogonalnošću u prostoru značajki. To proizlazi iz činjenice da se $k(x, x')$ približava nuli. Velika vrijednost parametra γ može dovesti do toga da su svi primjeri u prostoru značajki međusobno ortogonalni što vodi do prenaučenosti pa je stoga nužno u praksi odabrati povoljan γ. Suprotno, mala vrijednost može dovesti do podnaučenosti.

Poglavlje 6

Implementacija

U ovom poglavlju predstavit ću širu sliku metode potpornih vektora u regresiji, neke probleme prilikom implementacije i algoritam koji se najčešće koristi u rješavanju problema. Za više detalja o ovome pogledati [9] i [6].

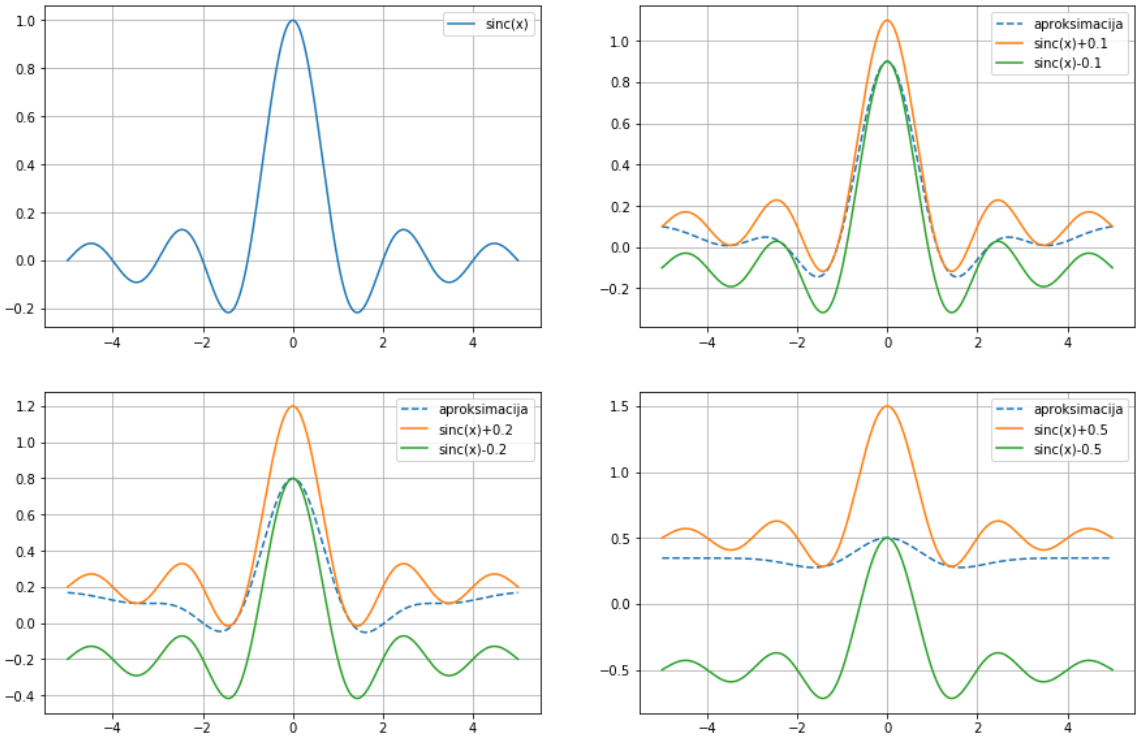
6.1 Šira slika

Ulazni podaci, za koje trebamo napraviti predikciju, preslikavaju se u prostor značajki preslikavanjem Φ. Zatim se računaju skalarni produkti sa slikama primjera iz skupa za učenje koje smo također dobili iz preslikavanja Φ. To odgovara pronalaženju vrijednosti jezgrene funkcije k u tim točkama. Konačno, skalarni produkti, odnosno vrijednosti jezgrene funkcije, sumiraju se sa težinama $\alpha_i - \alpha_i^*$. Ako dodamo konstantu b dobijemo konačni izlaz.

Slika 6.1 prikazuje kako metoda potpornih vektora izabire funkciju s najmanjim mogućim nagibom među onima koje aproksimiraju dani skup podataka sa odabranom preciznošću. Iako zahtjevamo da je funkcija najravnija u prostoru značajki, izrazito je ravna i u ulaznom prostoru, no više o tome zašto je to tako možete pronaći u [9].

Slika 6.2 pokazuje odnos između kvalitete aproksimacije i broja potpornih vektora. Manja preciznost rezultira manjim brojem potpornih vektora. Sve preostale točke skupa za učenje su redundantne, to jest bez tih primjera bi dobili istu aproksimaciju.

Na slici 6.3 možemo vidjeti kojim intenzitetom Lagrangeovi mutiplikatori α_i i α_i^* povlače i guraju aproksimaciju unutar ε-okoline. Te "sile" se aktiviraju samo u slučajevima kada aproksimacija dira ili potencijalno izlazi iz unaprijed određene okoline. To je direktna ilustracija djelovanja uvjeta Kuhn-Tuckerovog teorema: ili je regresija unutar okoline pa su Lagrangeovi multiplikatori nula ili moramo primjeniti silu u obliku $\alpha_i \neq 0$ i $\alpha_i^* \neq 0$ kako bi uvjeti bili zadovoljeni.



Slika 6.1: Gornji lijevi graf: Originalna funkcija sinc x, gornji desni graf: aproksimacija sa $\varepsilon = 0.1$ preciznošću, donji lijevi graf: $\varepsilon = 0.2$, donji desni graf: $\varepsilon = 0.5$.

6.2 Problemi u implementaciji

Klasifikacijski optimizacijski problem (3.64) s ograničenjima (3.65) možemo zapisati u matričnom obliku:

$$-0.5\alpha^T H\alpha + p^T\alpha \rightarrow \max, \tag{6.1}$$

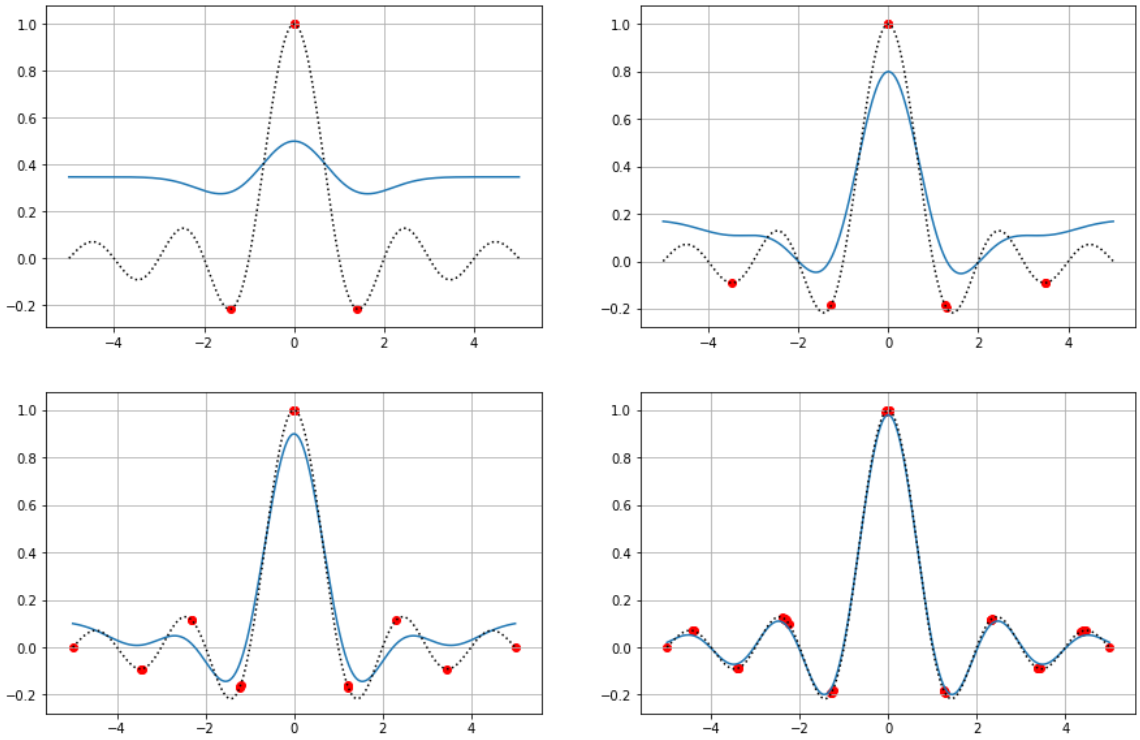
s ograničenjima

$$\begin{aligned} &y^T\alpha = 0, \\ &0 \leq \alpha_i \leq C \quad i = 1, \ldots, N, \end{aligned} \tag{6.2}$$

gdje je $\alpha = [\alpha_1, \alpha_2, \ldots, \alpha_N]^T$, H $N \times N$ Hesijan takav da je $H_{ij} = y_i y_j \langle x_i, x_j \rangle$, odnosno, u slučaju da imamo jezgrenu funkciju k umjesto skalarnog produkta, $H_{ij} = y_i y_j k(x_i, x_j)$, a p je vektor N jedinica $p = [11 \ldots 1]^T$.

Regresijski optimizacijski problem (5.5) s ograničenjima (5.6) možemo zapisati u matričnom obliku:

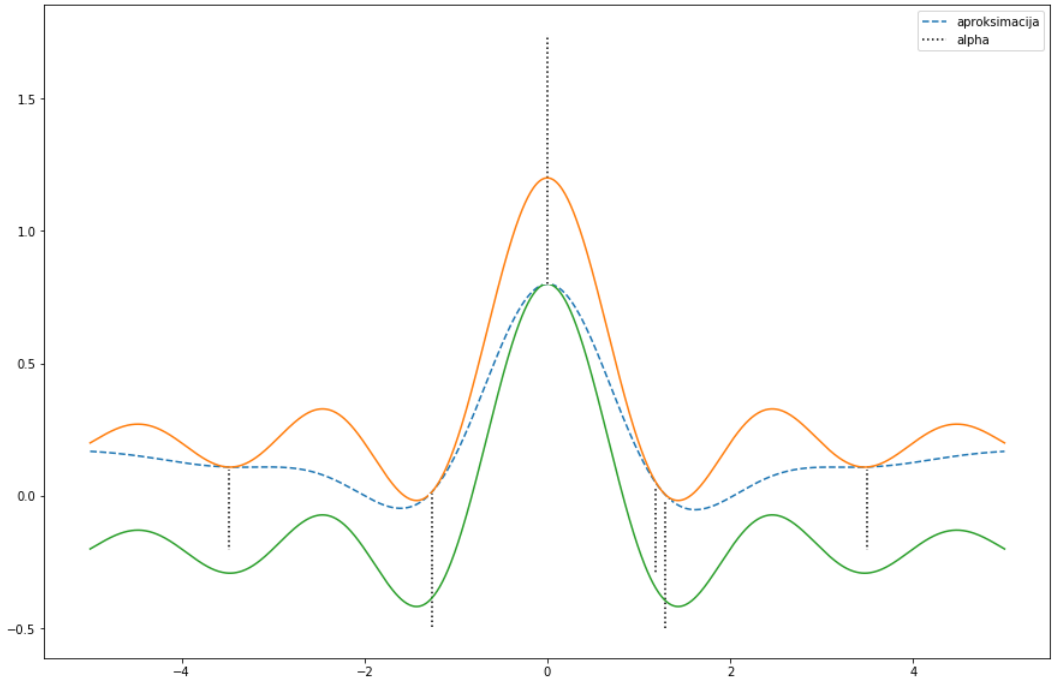
$$0.5\alpha^T H\alpha + p\alpha \rightarrow \min, \tag{6.3}$$



Slika 6.2: Gornji lijevi graf: puna linija predstavlja aprosimaciju, manje crne točke skup podataka, a veće crvene potporne vektore za $\varepsilon = 0.5$, gornji desni graf: $\varepsilon = 0.2$, donji lijevi graf: $\varepsilon = 0.1$, donji desni graf: $\varepsilon = 0.02$.

gdje je $\alpha = [\alpha_1, \alpha_2, \ldots, \alpha_N, \alpha_1^*, \alpha_2^*, \ldots, \alpha_N^*]$, H, Hesijan, $2N \times 2N$ matrica oblika H =[K $-K$;$-K$ K] gdje je K $N \times N$ matrica takva da je $K_{ij} = \langle x_i, x_j \rangle$, odnosno $K_{ij} = k(x_i, x_j)$. p je oblika $p = [\varepsilon - y_1, \varepsilon - y_2, \ldots, \varepsilon - y_N, \varepsilon + y_1, \varepsilon + y_2, \ldots, \varepsilon + y_N]$. Ograničenja (5.6) ostaju ista.

Kada je skup podataka izrazito velik (na primjer $N > 2000$) problem kvadratnog programiranja postaje zahtjevan. Ako klasificiramo skup za učenje od 50000 primjera, matrica H, koja je gusto popunjena i loše uvjetovana pa često zahtjeva regularizaciju (dodavanje malih vrijednosti na dijagonalu), ima $2,5 \cdot 10^9$ elemenata. U slučaju da koristimo 8-bajtnu floating point reprezentaciju potrebno je 20000 megabajta memorije, što je izrazito puno. To je glavni nedostatak metode potpornih vektora.



Slika 6.3: Sile koje zadržavaju aproksimaciju unutar ε-okoline

6.3 Komadanje i SMO algoritam

Komadanje

Postupak se oslanja na činjenici da su jedino potporni vektori ključni za formiranje rezultata, odnosno kada bi bili dani samo potporni vektori dobili bi identično rješenje kao u slučaju da smo imali na raspolaganju cijeli skup za učenje. Kako nam prije rješavanja problema nije poznato koji su vektori iz skupa za učenje potporni, počinjemo sa proizvoljnim podskupom skupa za učenje, to jest prvim *komadom* koji je dovoljno mali s obzirom na memoriju računala. Učimo model na tim podacima, zadržavamo samo potporne vektore a ostatak nadopunjavamo primjerima na kojima bi trenutni procjenitelj dobivao greške (npr. podaci izvan ε-okoline u slučaju regresije). Model učimo na novim podacima i ponavljamo iteracije sve dok nisu zadovoljeni uvjeti Kuhn-Tuckerovog teorema za sve primjere.

Neka je $S_w \subset \{1, \ldots, N\}$ skup oznaka primjera koje koristimo za učenje modela, a $S_f \subset \{1, \ldots, N\}$ skup oznaka fiksiranih primjera. Vrijedi da je $S_w \cup S_f = \{1, \ldots, N\}$ i $S_w \cap S_f = \emptyset$. Uzmimo primjer regresije s konveksnom funkcijom troška. Jedini nekvadratni dio pojavljuje se u $\sum_i T(\alpha_i) + T(\alpha_i^*)$. Bez smanjenja općenitosti pretpostavit ćemo

da je $\varepsilon \neq 0$ i $\alpha \in [0, C]$. Problem (4.31) s ograničenjima (4.33) možemo zapisati kao

$$\begin{aligned} &-\frac{1}{2}\sum_{i,j\in S_w}(\alpha_i-\alpha_i^*)(\alpha_j-\alpha_j^*)\langle x_i,x_j\rangle+\\ &+\sum_{i\in S_w}(\alpha_i-\alpha_i^*)\left(y_i-\sum_{j\in S_f}(\alpha_j-\alpha_j^*)\langle x_i,x_j\rangle\right)+\\ &+\sum_{i\in S_w}(-\varepsilon(\alpha_i+\alpha_i^*)+C(T(\alpha_i)+T(\alpha_i^*)))\to\max, \end{aligned} \tag{6.4}$$

s ograničenjima

$$\begin{gathered} \sum_{i\in S_w}(\alpha_i-\alpha_i^*)=-\sum_{i\in S_f}(\alpha_i-\alpha_i^*),\\ \alpha_i\in[0,C]. \end{gathered} \tag{6.5}$$

SMO algoritam

John C. Platt predložio je 1999. godine SMO[1] algoritam koji komadanje dovodi do ekstrema iterativno birajući podskupove veličine 2. Pokazalo se da je taj način i do nekoliko puta brži od klasičnog komadanja. Ključ je u tome da se za skup veličine 2 optimizacijski problem može riješiti analitički, bez korištenja kvadratnih optimizatora.

U slučaju regresije algoritam se koristi samo za ε-robusnu funkciju gubitka jer je za većinu ostalih konveksnih funkcija gubitka nemoguće dobiti eksplicitno rješenje.

Promatramo problem (6.4) u slučaju dva indeksa (i, j) u S_w skupu. C_i može biti različit za svaki primjer (može se čak razlikovati za α_i i α_i^*). Definiramo pomoćnu varijablu $s := y_i y_j$ za klasifikaciju ($y_i \in \{-1, 1\}$). Za regresiju razlikujemo četiri slučaja: $(\alpha_i, \alpha_j), (\alpha_i, \alpha_j^*), (\alpha_i^*, \alpha_j), (\alpha_i^*, \alpha_j^*)$. Neka je $s = 1$ u prvom i zadnjem slučaju, a u ostalima $s = -1$. Iz ograničenja sumacije za klasifikaciju imamo

$$s\alpha_i + \alpha_j = s\alpha_i^{stari} + \alpha_j^{stari} =: \gamma, \tag{6.6}$$

a za regresiju

$$(\alpha_i - \alpha_i^*) + (\alpha_j - \alpha_j^*) = (\alpha_i^{stari} - \alpha_i^{*stari}) + (\alpha_j^{stari} - \alpha_j^{*stari}) := \gamma. \tag{6.7}$$

Kako je $\alpha_j^{(*)} \in [0, C_j^{(*)}]$ tada $\alpha_i^{(*)} \in [L, H]$ gdje su L i H dani u tablicama 6.1 i 6.2. Sada je potrebno riješiti optimizacijski problem analitički za dvije (odnosno četiri u slučaju regresije) varijable. Definiramo φ_i kao grešku trenutne hipoteze na primjeru x_i

$$\varphi_i := y_i - f(x_i) = y_i - \left[\sum_{j=1}^{m} k(x_i, x_j)(\alpha_i - \alpha_i^*) + b\right]. \tag{6.8}$$

[1] Sequential Minimal Optimization (eng. sekvencijalna minimalna optimizacija)

	$y_i = y_j$	$y_i \neq y_j$
α_i	$\mathrm{L} = \max(0, \gamma - C_j)$ $\mathrm{H} = \min(C_i, \gamma)$	$\mathrm{L} = \max(0, \gamma)$ $\mathrm{H} = \min(C_i, \gamma + C_j)$

Tablica 6.1: Ograničenja domene za klasifikaciju

	α_j	α_j^*
α_i	$\mathrm{L} = \max(0, \gamma - C_j)$ $\mathrm{H} = \min(C_i, \gamma)$	$\mathrm{L} = \max(0, \gamma)$ $\mathrm{H} = \min(C_i, C_j^* + \gamma)$
α_i^*	$\mathrm{L} = \max(0, -\gamma)$ $\mathrm{H} = \min(C_i^*, -\gamma + C_j)$	$\mathrm{L} = \max(0, -\gamma - C_j^*)$ $\mathrm{H} = \min(C_i^*, -\gamma)$

Tablica 6.2: Ograničenja domene za regresiju

Pomoću toga definiramo

$$\begin{aligned} v_i :&= y_i - \sum_{a \neq i,j} (\alpha_a - \alpha_a^*) K_{ia} + b \\ &= \varphi_i + (\alpha_i^{stari} - \alpha_i^{*stari}) K_{ii} + (\alpha_j^{stari} - \alpha_j^{*stari}) K_{ij} \end{aligned} \tag{6.9}$$

pa vrijedi

$$v_i - v_j - \gamma(K_{ij} - K_{jj}) = \gamma_i - \gamma_j + (\alpha_i^{stari} - \alpha_i^{*stari})(K_{ii} + K_{jj} - 2K_{ij}). \tag{6.10}$$

Ako (6.4) ograničimo na (i, j) imamo

$$\begin{aligned} -\frac{1}{2} \begin{pmatrix} \alpha_i - \alpha_i^* \\ \alpha_j - \alpha_j^* \end{pmatrix}^T \begin{pmatrix} K_{ii} & K_{ij} \\ K_{ji} & K_{jj} \end{pmatrix} \begin{pmatrix} \alpha_i - \alpha_i^* \\ \alpha_j - \alpha_j^* \end{pmatrix} + v_i(\alpha_i - \alpha_i^*) + \\ + v_j(\alpha_j - \alpha_j^*) - \varepsilon(\alpha_i + \alpha_i^* + \alpha_j + \alpha_j^*) \to \max, \end{aligned} \tag{6.11}$$

s ograničenjima

$$\begin{aligned} (\alpha_i - \alpha_i^*) + (\alpha_j - \alpha_j^*) &= \gamma \\ \alpha_i, \alpha_i^*, \alpha_j, \alpha_j^* &\in [0, C] \end{aligned} \tag{6.12}$$

Nadalje, potrebno je eliminirati α_j, α_j^* koristeći ograničenje sumacije. Zanemarujući dijelove neovisne o $\alpha_i^{(*)}$ dolazimo do problema

$$\begin{aligned} -\frac{1}{2}(\alpha_i - \alpha_i^*)^2 (K_{ii} + K_{jj} - 2K_{ij}) - \varepsilon(\alpha_i + \alpha_i^*)(1 - s) + \\ + (\alpha_i - \alpha_i^*)(v_i - v_j - \gamma(K_{ij} - K_{jj})) \to \max, \end{aligned} \tag{6.13}$$

α_i, α_j	$\frac{v_i - v_j - \gamma(K_{ij} - K_{jj})}{\eta} = \alpha_i^{stari} + \frac{\varphi_i - \varphi_j}{\eta}$
α_i, α_j^*	$\frac{v_i - v_j - \gamma(K_{ij} - K_{jj}) - 2\varepsilon}{\eta} = \alpha_i^{stari} + \frac{\varphi_i - \varphi_j - 2\varepsilon}{\eta}$
α_i^*, α_j	$\frac{v_j - v_i + \gamma(K_{ij} - K_{jj}) - 2\varepsilon}{\eta} = \alpha_i^{*stari} - \frac{\varphi_i - \varphi_j + 2\varepsilon}{\eta}$
α_i^*, α_j^*	$\frac{v_j - v_i + \gamma(K_{ij} - K_{jj})}{\eta} = \alpha_i^{*stari} - \frac{\varphi_i - \varphi_j}{\eta}$

Tablica 6.3: Maksimum problema kvadratnog programiranja

s ograničenjima

$$\alpha_i^{(*)} \in [L^{(*)}, H^{(*)}]. \tag{6.14}$$

Maksimum (6.13) u ovisnosti o α_i, α_i^* vidimo u tablici 6.3. Definiramo $\eta := K_{ii} + K_{jj} - 2K_{ij}$. Sve što treba ponovo izračunati je $\varphi_i^{novi} - \varphi_j^{novi}$ i to na način

$$\begin{aligned} \varphi_i^{novi} - \varphi_j^{novi} &= \varphi_i^{stari} - ((\alpha_i^{novi} - \alpha_i^{*novi}) - (\alpha_i^{stari} - \alpha_i^{*stari}))(K_{ii} - K_{ij}) - \\ &\quad - \varphi_j^{stari} - ((\alpha_j^{novi} - \alpha_j^{*novi}) - (\alpha_j^{stari} - \alpha_j^{*stari}))(K_{ij} - K_{jj}) \\ &= \varphi_i^{stari} - \varphi_j^{novi} - \eta((\alpha_i^{novi} - \alpha_i^{*novi}) - (\alpha_i^{stari} - \alpha_i^{*stari})). \end{aligned} \tag{6.15}$$

Zbog numeričke nestabilnosti može se dogoditi da je $\eta < 0$. Kako k mora zadovoljavati Mercerovo svojstvo, u takvom slučaju $\eta = 0$. Optimalna vrijednost α_i leži na granicama H ili L. Odgovarajuću granicu možemo odabrati računajući gradijent ili jednostavno uvrštavajući vrijednosti u funkciju cilja.

Postavlja se pitanje kako pravilno odabrati indekse (i, j) kako bi maksimizirali funkciju cilja. Koristimo se dvijema petljama. Za indeks i vanjska ide po svim primjerima koji ne zadovoljavaju uvjete Kuhn-Tuckerovog teorema, prvo samo po onima gdje Lagrangeovi multiplikatori nisu ni na donjoj ni na gornjoj granici. Kada su ti zadovoljeni ide po svim primjerima koji ne zadovoljavaju Kuhn-Tuckerove uvjete. Prebacujemo se na indeks j. Za velike korake prema minimumu potrebno je tražiti velike korake u α_i. Kako je računalno skupo računati η za sve moguće parove (i, j) maksimiziramo apsolutnu vrijednost brojnika u tablici 6.3. Biramo onaj indeks j koji odgovara maksimumu apsolutne vrijednosti.

SMO algoritam ne daje automatski vrijednosti za b, no ako je barem jedna od varijabli $\alpha_i^{(*)}$ i $\alpha_j^{(*)}$ unutar granica možemo koristiti (4.18). Ako to nije slučaj postoji cijeli interval (npr. $[b_i, b_j]$) unutar kojeg možemo tražiti odgovarajuće vrijednosti za b. Tada jednostavno uzimamo $b = \frac{b_i + b_j}{2}$.

U dodatku A u [9] nalazi se pseudokod za ovaj algoritam u regresiji.

Poglavlje 7

Primjeri u analizi teksta

U ovom poglavlju ću na dva primjera prikazati korištenje metode potpornih vektora. Primjeri su pisani u programskom jeziku *Python*, a metoda je implementirana preko *scikit-learn* biblioteke o kojoj više možete pronaći na [1].

Prije svega potrebno je definirati mjere točnosti. Pretpostavimo da imamo binarnu klasifikaciju na pozitivne i negativne primjere. Primjeri koji su ispravno klasificirani su točno pozitivni (TP) i točno negativni (TN). Neispravno klasificirani primjeri su lažno pozitivni (FP) i lažno negativni (FN).
Točnost je udio točno klasificiranih u skupu svih primjera:

$$\mathrm{T} = \frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{FN}}.$$

Preciznost je udio točno pozitivnih u skupu pozitivno klasificiranih primjera:

$$\mathrm{P} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}}.$$

Odziv je udio točno pozitivnih primjera u skupu svih pozitivnih primjera:

$$R = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}.$$

F_1 *mjera* je harmonijska sredina preciznosti i odziva:

$$\mathrm{F}_1 = \frac{2}{\frac{1}{\mathrm{P}} + \frac{1}{\mathrm{R}}} = \frac{2\mathrm{PR}}{\mathrm{P}+\mathrm{R}}.$$

F_1 mjera je dobar kompromis između preciznosti i odziva i često se koristi kao mjera točnosti.
$0-1$ *pogreška* pokazuje udio pogrešno klasificiranih primjera.

Analiza filmskih kritika sa IMDb web stranice

IMDb je web stranica na kojoj korisnici mogu pisati filmske kritike i ocjenjivati filmove ocjenama od 1 do 10.

Sa [7] preuzeo sam skup podataka koji se sastoji od 25000 kritika. One s ocjenom <= 4 označene su kao negativne, dok su one s ocjenom >= 7 označene kao pozitivne. Za pojedini film nije dozvoljeno više od 30 kritika.

Primjer pozitivne kritike je:

"Bromwell High is nothing short of brilliant. Expertly scripted and perfectly delivered, this searing parody of a students and teachers at a South London Public School leaves you literally rolling with laughter. It's vulgar, provocative, witty and sharp. The characters are a superbly caricatured cross section of British society (or to be more accurate, of any society). Following the escapades of Keisha, Latrina and Natella, our three "protagonists" for want of a better term, the show doesn't shy away from parodying every imaginable subject. Political correctness flies out the window in every episode. If you enjoy shows that aren't afraid to poke fun of every taboo subject imaginable, then Bromwell High will not disappoint!"

Primjer negativne kritike je:

"Ned aKelly is such an important story to Australians but this movie is awful. It's an Australian story yet it seems like it was set in America. Also Ned was an Australian yet he has an Irish accent...it is the worst film I have seen in a long time"

Kritike sam vektorizirao pomoću *TfidfVectorizer* metode koja se nalazi u *scikit-learn* biblioteci. Ona mjeri frekvenciju pojavljivanja svake riječi i skalira ih na način da smanji utjecaj onih riječi koje se pojavljuju češće (poput riječi *the*). Svaka komponenta vektora predstavlja jednu riječ. Tako su kritike postale točke u prostoru dimenzije 74849.

Podijelio sam skup podataka na skup podataka za učenje i skup podataka za testiranje na način da sam nasumično odabrao 80% primjera za učenje iz cijelog skupa.

Učenje sam proveo pomoću SVC metode u kojoj je implementirana metoda potpornih vektora za klasifikaciju. Za parametar C odabrao sam vrijednost 1. Kako je broj značajki znantno veći od broja primjera odlučio sam se za linearnu klasifikaciju.

Na skupu primjera za testiranje izračunao sam točnost i F_1 mjeru. Točnost je 0.895, a F_1 mjera 0.896. Usporedbe radi, u slučaju polinomijalne klasifikacije stupnja 2 klasifikator je sve primjere iz skupa za testiranje označio kao negativne pa je točnost 0.497.

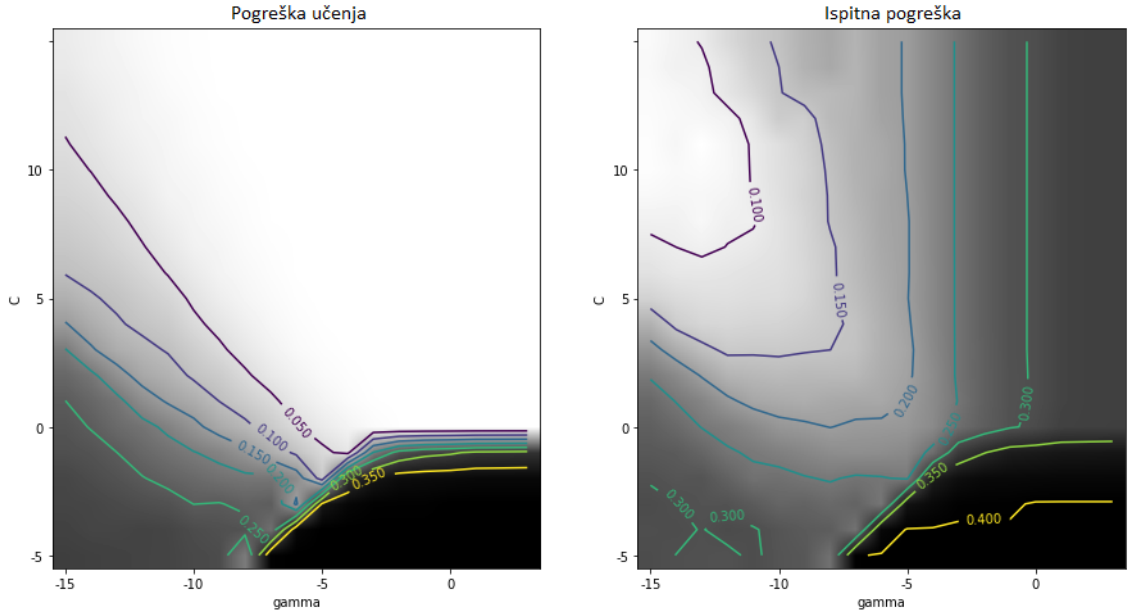
Identifikacija neželjene elektroničke pošte (spam)

Neželjena pošta podrazumijeva elektroničke poruke najčešće komercijalnog karaktera koje nerijetko nastoje obmanuti primatelja. Zbog toga je potrebno neželjenu poštu odvojiti od korisnih poruka, odnosno klasificirati svu poštu na onu koju primatelj želi otvoriti i onu koju ne želi.

Sa [5] preuzeo sam skup podataka koji se sastoji od 4601 vektora sa 58 komponenata koji su nastali vektoriziranjem elektoničkih poruka. Prvih 48 komponenata predstavljaju postotak pojavljivanja određenih riječi u poruci. Sljedećih 6 označavaju postotak pojavljivanja određenih znakova. Komponenta nakon predstavlja prosječnu duljinu neprekidnog niza velikih slova, zatim sljedeća pokazuje najveću duljinu neprekidnog niza velikih slova. Komponenta nakon predstavlja ukupnu količinu velikih slova. Posljednja komponenta je nominalna i pokazuje je li poruka neželjena ili ne.

Skup podataka sam opet nasumično podijelio na skup za učenje i skup za testiranje na način da je 60% primjera u skupu za učenje.

Kako je broj značajki u ovom primjeru znatno manji od broja primjera odlučio sam se za Gaussovu jezgru. Na temelju $0-1$ pogreške proveo sam unakrsnu provjeru za parametre C i γ, to jest za svaki $C \in \{2^{-5}, 2^{-4}, \ldots, 2^{15}\}$ i za svaku $\gamma \in \{2^{-15}, 2^{-14}, \ldots, 2^{3}\}$ učio sam model pomoću ugrađene SVC metode i računao $0-1$ pogrešku na skupu za učenje i skupu za testiranje. Iz najmanje pogreške na skupu za testiranje koja iznosi 0.077 zaključio sam da je optimalni $C = 2^{11}$, a optimalni $\gamma = 2^{-13}$. Rezultate unakrsne provjere možemo vidjeti na slici 7.1. Na njoj je prikazana ovisnost pogreške o parametrima C i γ. Što je područje svjetlije to je pogreška manja.



Slika 7.1: Ovisnost 0-1 pogreške učenja i 0-1 ispitne pogreške o parametrima C i γ. Na osima su ispisani eksponenti broja 2.

Bibliografija

[1] *scikit learn*, https://scikit-learn.org/stable/, (Kolovoz 2019.).

[2] D. Bakić, *Normirani prostori*, https://web.math.pmf.unizg.hr/~bakic/np/NP_17_18.pdf, (Srpanj 2019.).

[3] A. Censor, *The implicit function theorem*, https://www.youtube.com/watch?v=bk9IKHS5KbY, (Srpanj 2019.).

[4] _____, *Proof of the Lagrange multipliers theorem*, https://www.youtube.com/watch?v=FTkIAk0g3_g, (Srpanj 2019.).

[5] Dheeru Dua i Casey Graff, *UCI Machine Learning Repository*, 2017, http://archive.ics.uci.edu/ml.

[6] T. Huang, V. Kecman i I. Kopriva, *Kernel Based Algorithms for Mining Huge Data Sets*, Studies in Computational Intelligence, sv. 17, Springer, 2006.

[7] Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng i Christopher Potts, *Learning Word Vectors for Sentiment Analysis*, Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (Portland, Oregon, USA), Association for Computational Linguistics, lipanj 2011, str. 142–150, http://www.aclweb.org/anthology/P11-1015.

[8] N. Sarapa, *Teorija vjerojatnosti*, Školska Knjiga, 2002.

[9] A. Smola i B. Schölkopf, *A tutorial on Support Vector Regression*, (1998).

[10] J. Thickstun, *Mercer's Theorem*, https://homes.cs.washington.edu/~thickstn/docs/mercer.pdf, (Rujan 2019.).

[11] V. N. Vapnik, *Statistical Learning Theory*, John Wiley Sons, Inc., 1998.

Sažetak

Metoda potpornih vektora jedna je od tehnika nadziranog učenja koja probleme klasifikacije i regresije transformira u probleme kvadratnog programiranja. Osnovna ideja tih dviju glavnih zadaća nadziranog učenja nalazi se u prvom poglavlju. Koristeći teoreme iz teorije optimizacije, koje sam predstavio u drugom poglavlju, metoda potpornih vektora formulira hipotezu pomoću dijela primjera iz skupa za učenje koje nazivamo potporni vektori. U trećem i četvrtom poglavlju opisao sam kako se dolazi do tih rezultata.

Dualna formulacija problema kvadratnog programiranja omogućila nam je da za rješavanje problema koji zahtjevaju nelinearna rješenja uvedemo jezgrene funkcije. Svojstva i primjeri jezgrenih funkcija koje se koriste u metodi potpornih vektora opisani su u petom poglavlju. U pronalasku korisnih jezgara ključan se pokazao Mercerov teorem. U šestom poglavlju pokazao sam neke probleme na koje nailazimo u implementaciji metode na računalu i kako se to uspješno rješava primjenom iterativnih postupaka poput SMO algoritma.

Napokon, u posljednjem poglavlju pokazao sam kako se metoda potpornih vektora koristi u analizi teksta. U prvom primjeru sam pomoću IMDb kritika filmova napravio klasifikacijski model koji predviđa je li kritika pozitivna ili negativna. U drugom primjeru napravio sam klasifikacijski model koji nam govori je li elektronička poruka neželjena ili ne. Također, pokazao sam kako parametri C i γ i odabir prave jezgrene funkcije utječu na točnost modela. Na ovim primjerima lako se vidi da je metoda potpornih vektora jednostavan ali moćan alat za rješavanje klasifikacijskih zadataka.

Summary

Support vector machine is a supervised learning technique which transforms classification and regression problems in quadratic programming ones. Basic idea of these two main tasks of supervised learning is written in chapter one. Using theorems from optimization theory, which I presented in chapter two, support vector machine formulates hypothesis using only part of training dataset which we call support vectors. In chapters three and four I described a way to get to these results.

Dual formulation of quadratic programming problem gave us the ability to use kernel functions in solving problems that require nonlinear solutions. Characteristics and examples of kernel functions used in support vector machine are described in chapter five. Mercer theorem was crucial in finding useful kernels. In chapter six I showed some problems in implementing method on computer and how that can be successfully solved by using iterative algorithms, such as SMO.

Finally, in the last chapter I showed how to use support vector machine in text analysis. In the first example I used IMDb movie reviews to make classification model that predicts whether a review is good or bad. In the second example I made classification model that tells us if the e-mail is spam or not. I also showed how parameters C and γ and choosing right kernel function affects model accuracy. On these two examples it's easy to see that support vector machine is simple yet powerful tool in solving classification tasks.

Životopis

Rođen sam 24. siječnja 1996. godine u Splitu. U Marini završavam osnovnu školu, a nakon nje u Trogiru Srednju školu Ivana Lucića, smjer opće gimnazije. Godine 2014. upisujem preddiplomski sveučilišni studij Matematika na Prirodoslovno-matematičkom fakultetu Sveučilišta u Zagrebu, koji završavam 2017. godine te stičem naziv sveučilišni prvostupnik matematike. Iste godine upisujem diplomski studij Primijenjena matematika, kojeg završavam ovim radom.