1 Optimalna hiperravnina za slučaj linearno neodvojivih razreda

-SVMs (Support Vector Machines)
- do sada stroj s potpornim vektorima → razredi linearno odvojivi

Teži slučaj - linearno neodvojivi uzorci

- za zadani skup uzoraka za učenje (engl. training set) NIJE MOGUĆE KONSTRUIRATI HIPERRAVNINU KOJA RAZDVAJA UZORKE BEZ POGREŠKE

ŽELIMO: Naći optimalnu hiperravninu koja minimizira vjerojatnost klasfikacijske pogreške

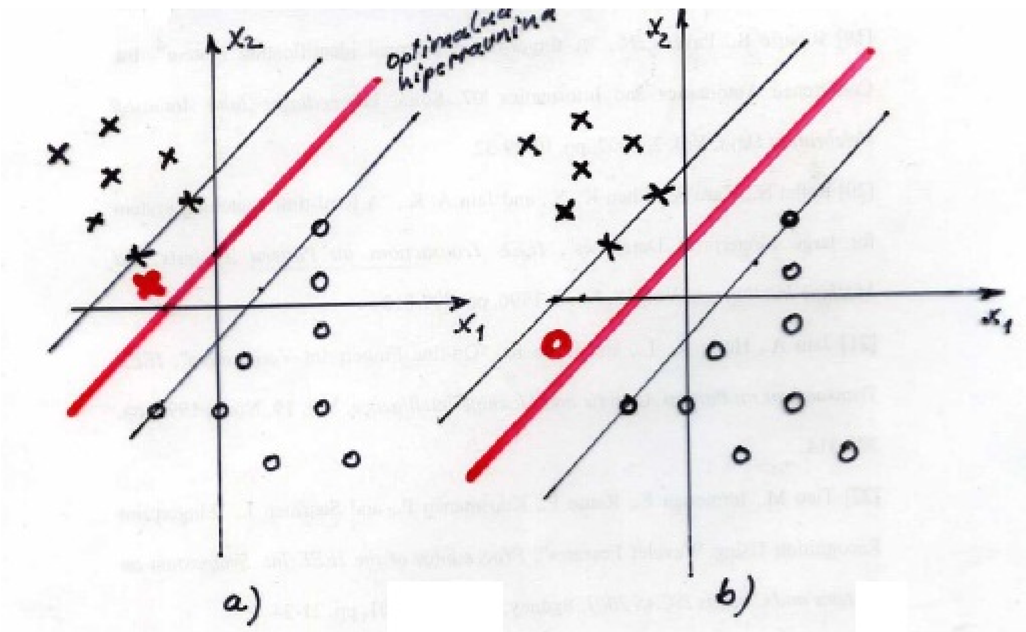
Margina razdvajanja naziva se mekom (engl. soft) ako za uzorak $(\vec{X}_i, d_i)$ ne vrijedi

$d_i \cdot (\vec{W}^T \vec{X}_i + b) \geq +1,\ i = 1, 2, \ldots, N$

Povreda uvjeta $d_i(\vec{W}^T \vec{X}_i + b) \geq 1$ može nastupiti na dva načina

(i) Uzorak $(\vec{X}_i, d_i)$ pada unutar područja odvajanja ALI na "pravu" stranu decizijske hiperravnine (na slici a)

(ii) Uzorak $(\vec{X}_i, d_i)$ pada na "krivu" stranu decizijske hiperravnine (na slici b)



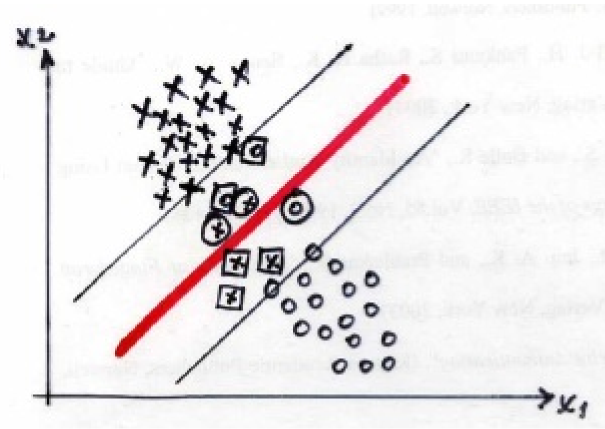
Slika 1: Prikaz slučajeva povrede uvjeta

- Margina je definirana kao udaljenost između para paralelnih hiperravnina opisanih s:

$\vec{W}^T \vec{X} + b = \pm 1$

Uzorci za učenje pripadaju jednoj od tri kategorije uzoraka:

(i) vektori koji padaju izvan pojasa i koji su ispravno klasificirani /ti vektori zadovoljavaju: $d_i \cdot (\vec{W}^T \vec{X}_i + b) \geq 1,\ i = 1, 2, \ldots, N$/



Slika 2: Drugačiji prikaz povrede

(ii) vektori koji padaju unutar pojasa (margine) ali su pravilno razvrstani:(x)(o)
za njih vrijedi: $0 \leq d_i \cdot (\vec{W}^T \vec{X}_i + b) < 1$

(iii) vektori s pogrešnom klasifikacijom: [o][x]
za njih vrijedi $d_i \cdot (\vec{W}^T \vec{X}_i + b) < 0$

Sva tri slučaja mogu se objediniti u jedinstveni oblik: $d_i \cdot (\vec{W}^T \vec{X}_i + b) \geq 1 - \zeta_i$
- novi skup varijabli:$\{\zeta_i\}_{i=1}^{N}$

(i) slučaj: odgovaraju uzorci za koje je $\zeta_i = 0$

(ii) slučaj: $0 < \zeta_i \leq 1$

(iii) slučaj: $\zeta_i > 1$

- Varijable ζ_i se nazivaju "labave" varijable (engl. slack variables) - one su mjera odstupanja vektora od idealnog uvjeta separabilnosti uzoraka.
- za $0 \leq \zeta_i \leq 1$ uzorak se nalazi unutar područja odvajanja ALI s prave strane decizijske ravnine
- za $\zeta_i > 1$ uzorak pada na krivu stranu decizijske ravnine
- potporni vektori su oni posebni uzorci koji zadovoljavaju jednadžbu
$d_i \cdot (\vec{W}^T \vec{X}_i + b) \geq 1 - \zeta_i$, $i = 1, 2, \ldots, N$ čak i za slučaj $\zeta_i > 0$.

ZADATAK: Naći decizijsku funkciju (hiperravninu razdvajanja) za koju je klasifikacijska pogreška minimalna!
Drugačije interpretirano: Učiniti marginu što je moguće većom ali uz uvjet da je broj uzoraka sa $\zeta_i > 0$ što je moguće manji!
Matematička interpretacija:
- minimizirati funkciju koštanja (funkcional)
$J(\vec{W}, b, \vec{\zeta}) = \frac{1}{2}\left\|\vec{W}\right\|^2 + C\sum_{i=1}^{N} I(\zeta_i)$, gdje je $\vec{\zeta}$ vektor s komponentama ζ_i i
$I(\zeta_i) = \begin{cases} 1 & \zeta_i > 0 \\ 0 & \zeta_i = 0 \end{cases}$

Parametar C je pozitivna konstanta kojom se upravlja relativan međuutjecaj $\frac{1}{2}\left\|\vec{W}\right\|^2$ i $\sum\limits_{i=1}^{N} I(\zeta_i)$

S. Haykin (1999.) predlaže dva moguća načina izbora C:

(i) eksperimentalno preko standardne uporabe uzoraka za učenje i validaciju

(ii) C se određuje analitički procjenom VC (Vapnik-Chervonenkis) dimenzije - koja je mjera ekspresivnosti porodice klasifikacijskih funkcija koje su ostvarene strojem za učenje:
dimenzija h, $h \leq \min\left\{\lceil \frac{D^2}{\rho^2} \rceil, m_0\right\} + 1$
D - dijametar najmanje kugle koja sadrži sve ulazne vektore $\vec{x_1}, \vec{x_2}, \ldots, \vec{X_N}$
ρ - margina $= \frac{2}{\|\vec{W}\|}$ i
m_0 - je dimenzionalnost ulaznosg prostora

Optimizacija izraza: $J(\vec{W}, b, \vec{\zeta}) = \frac{1}{2}\left\|\vec{W}\right\|^2 + C\sum\limits_{i=1}^{N} I(\zeta_i)$, teška je jer je uključena i funkcija I(.) koja je diskontinuirana!
Uobičajen pristup $\rightarrow$ izabrati optimizaciju funkcije koja je "bliska" funkciji koštanja:
$J(\vec{W}, b, \vec{\zeta}) = \frac{1}{2}\left\|\vec{W}\right\|^2 + C\sum\limits_{i=1}^{N} \zeta_i$, uz $d_i \cdot (\vec{W}^T\vec{X}_i + b) \geq 1 - \zeta_i$, $i = 1, 2, \ldots, N$ i $\zeta_i \geq 0$, $i = 1, 2, \ldots, N$

Formalno: primarni problem za neseparatibilne razrede

- Za zadani skup uzoraka za učenje $\left\{\vec{X}_i, d_i\right\}_{i=1}^{N}$ naći optimalne vrijednosti težinskog vektora $\vec{W}$ i pomaknuća b tako da su zadovoljena sljedeća ograničenja:
$d_i \cdot (\vec{W}^T\vec{X}_i + b) \geq 1 - \zeta_i$, za $i = 1, 2, \ldots, N$ i $\zeta_i \geq 0$, za sve i
i to takve da $\vec{W}$ i labave varijable ζ_i minimiziraju funkciju koštanja
$J(\vec{W}, b, \vec{\zeta}) = \frac{1}{2}\vec{W}^T\vec{W} + C\sum\limits_{i=1}^{N} \zeta_i$, pri čemu je C korisničko definiran pozitivni parametar.
Uporabom metode Lagrangeovih multiplikatora i postupkom sličnim za lin. separatibilni SM slijedi:
$\mathcal{L}(\vec{W}, b, \vec{\zeta}, \vec{\lambda}, \vec{\mu}) = \frac{1}{2}\left\|\vec{W}\right\|^2 + C\sum\limits_{i=1}^{N} \zeta_i - \sum\limits_{i=1}^{N} \mu_i\zeta_i - \sum\limits_{i=1}^{N} \lambda_i \cdot [d_i(\vec{W}^T\vec{X}_i + b) - 1 + \zeta_i]$
Odgovarajući K-K-T (Karush-Kuhn-Tucker) uvjeti su
$\frac{\partial \mathcal{L}}{\partial \vec{W}} = \vec{0}$ ili $\vec{W} = \sum\limits_{i=1}^{N} \lambda_i d_i \vec{X}_i$
$\frac{\partial \mathcal{L}}{\partial b} = 0$ ili $\sum\limits_{i=1}^{N} \lambda_i d_i = 0$
$\frac{\partial \mathcal{L}}{\partial \zeta_i} = \vec{0}$ ili $C - \mu_i - \lambda_i = 0$
$i = 1, 2, \ldots, N$
$\lambda_i[d_i(\vec{W}^T\vec{X}_i + b) - 1 + \zeta_i] = 0$, $i = 1, 2, \ldots, N$
$\mu_i\zeta_i = 0$, $i = 1, 2, \ldots, N$

$\mu_i \geq 0$, $\lambda_i \geq 0$, $i = 1, 2, \ldots, N$
Pridružena Wolfeova dualna reprezentacija postaje:
maksimiziraj $\mathcal{L}(\vec{W}, b, \vec{\lambda}, \vec{\zeta}, \vec{\mu})$ uz ograničenja :
$\vec{W} = \sum\limits_{i=1}^{N} \lambda_i d_i \vec{X}_i$
$C - \mu_i - \lambda_i = 0$, $i = 1, 2, \ldots, N$
$\lambda_i \geq 0$, $\mu_i \geq 0$, $i = 1, 2, \ldots, N$
Uvrštavanjem nejednadžbi u $\mathcal{L}$ dobiva se:
$\max\limits_{\vec{\lambda}}(\sum\limits_{i=1}^{N} \lambda_i - \frac{1}{2} \sum\limits_{i=1}^{N} \lambda_i \lambda_j d_i d_j \vec{X}_i^{\,T} \vec{X}_j)$ uz uvjete $0 \leq \lambda_i \leq C$, $i = 1, 2, \ldots, N$ i $\sum\limits_{i=1}^{N} \lambda_i d_i = 0$
Lagrangeovi multiplikatori koji odgovaraju točkama (uzorcima) koji leže unutar margine ili na krivoj stani klasifikatora, tj. $\zeta_i > 0$ su svi jednaki maksimalno dopuštenoj vrijednosti C.
- promotrimo izraz:

$$\max_{\vec{\lambda}}(\sum_{i=1}^{N} \lambda_i - \frac{1}{2} \sum_{i=1}^{N} \lambda_i \lambda_j d_i d_j \vec{X}_i^{\,T} \vec{X}_j)$$

vidimo da niti labave varijable ζ_i niti Lagrangeovi multiplikatori pridruženi labavim varijablama μ_i ne ulaze eksplicitno u problem
- oni su prisutni neizravno kroz C, odnosno kroz uvjet: $0 \leq \lambda_i \leq C$
Razlika u odnosu na linearno separatibilni slučaj:
Lagrangeovi multiplikatori su (moraju biti) ograničeni sa C. / $0 \leq \lambda_i \leq C$ /
-za linearno separatibilni slučaj $C \to \infty$ /$\lambda_i \geq 0$/
Osim ove modifikacije
$\lambda_i \geq 0 \to 0 < \lambda_i \leq C$
postupak računanja optimalnih vrijednosti vektora $\vec{W}$ i pomaknuća b je jednak kao i kod separatibilnog slučaja
Rješenje:
$\vec{W}_0 = \sum\limits_{i=1}^{N_s} \lambda_{0,i} d_i \vec{X}_i$, gdje je N_S broj vektora koji su potporni, a $0 < \lambda_{0,i} < C$
pomaknuće b_0 se može odrediti uzimajući bilo koji $(\vec{X}_i, d_i)$ iz skupa za učenje za koji vrijedi $0 < \lambda_{0,i} < C$ i zato je $\zeta_i = 0$