Strojno učenje
7
Metoda potpornih vektora
(SVM – Support Vector Machines)

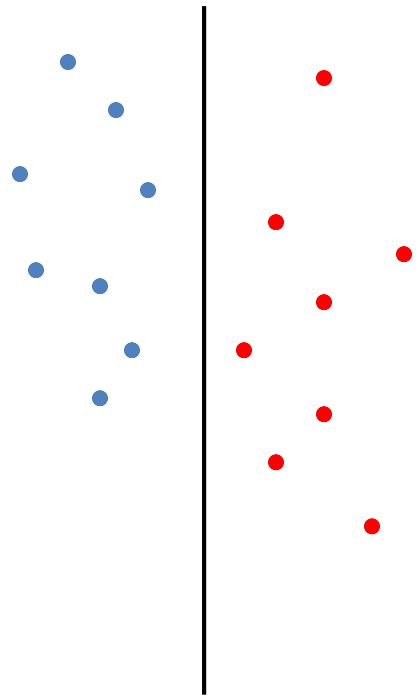
Tomislav Šmuc

PMF, Zagreb, 2020

Generativni i diskriminativni modeli

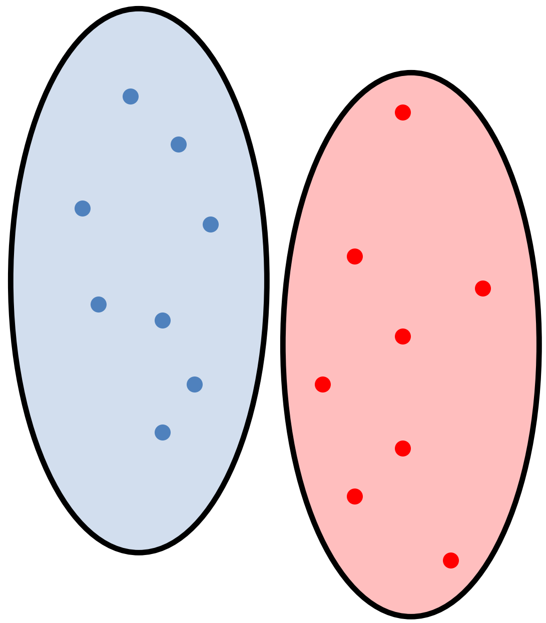
- Diskriminativni

(Učenje linije koja razdvaja klase)



- Generativni

Učenje modela za svaku pojedinu klasu



Diskriminativni algoritmi – do sada

Klasifikacija

- Klasifikator baziran na linearnoj regresiji
- Logistička regresija
- LDA, QDA, FDA

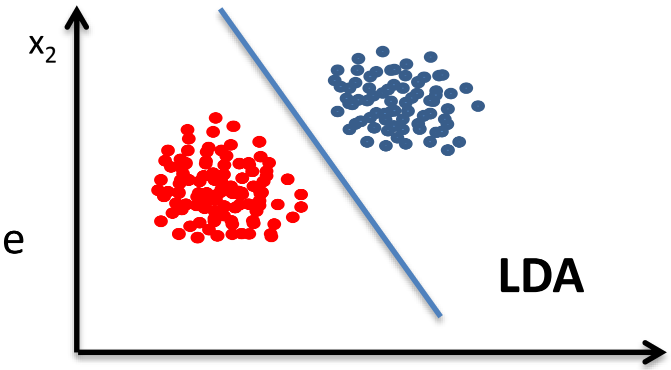
- Diskriminativne funkcije - $\delta_k(x)$- funkcije koje određuju pripadnost nekoj klasi *k*

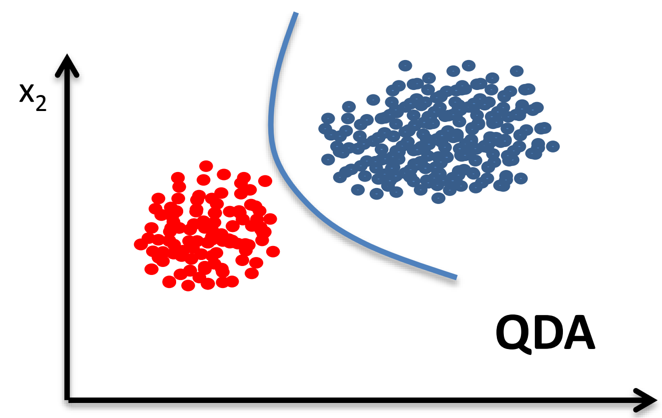
$$C(x) = \arg\max_{k \in C} \delta_k(x)$$

– određeneplohe/hiperravnine koje razdvajaju klase

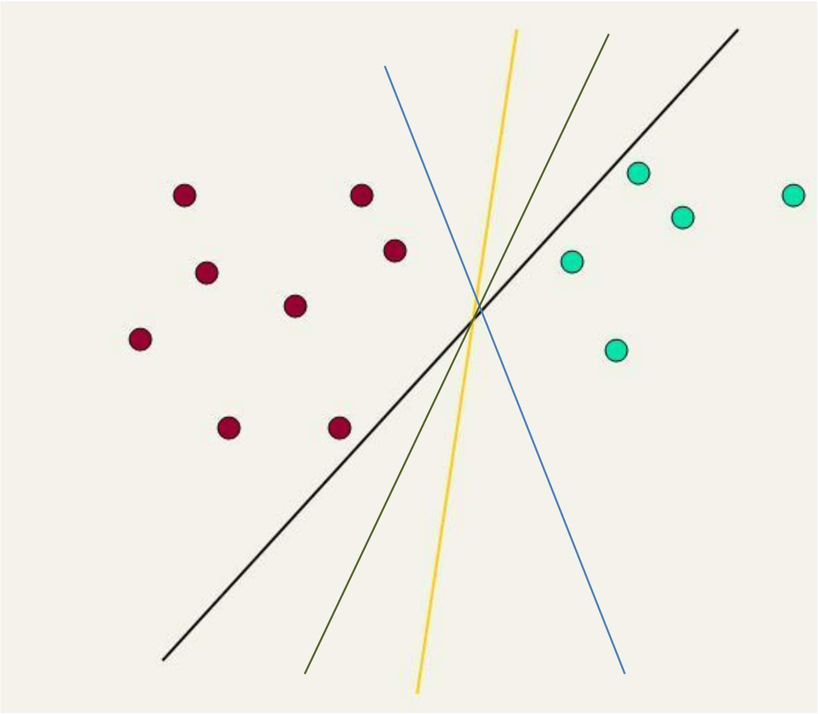
$$\{\mathbf{x} : \delta_k(\mathbf{x}) = \delta_l(\mathbf{x})\}$$

- intuicija: što je $\delta_{k(x)}$ veća to je vjerojatnost pripadanja klasi *k* veća ~ udaljenosti od plohe/hiperravnine razdvajanja





Koja hiper-ravnina je najbolja ?



Separabilni problem: Između točaka možemo provući velik broj (beskonačan) ploha koje dobro razdvajaju primjere dvije klase

Metoda potpornih vektora (metoda jezgrenih funkcija)
SVM - Support Vector Machines

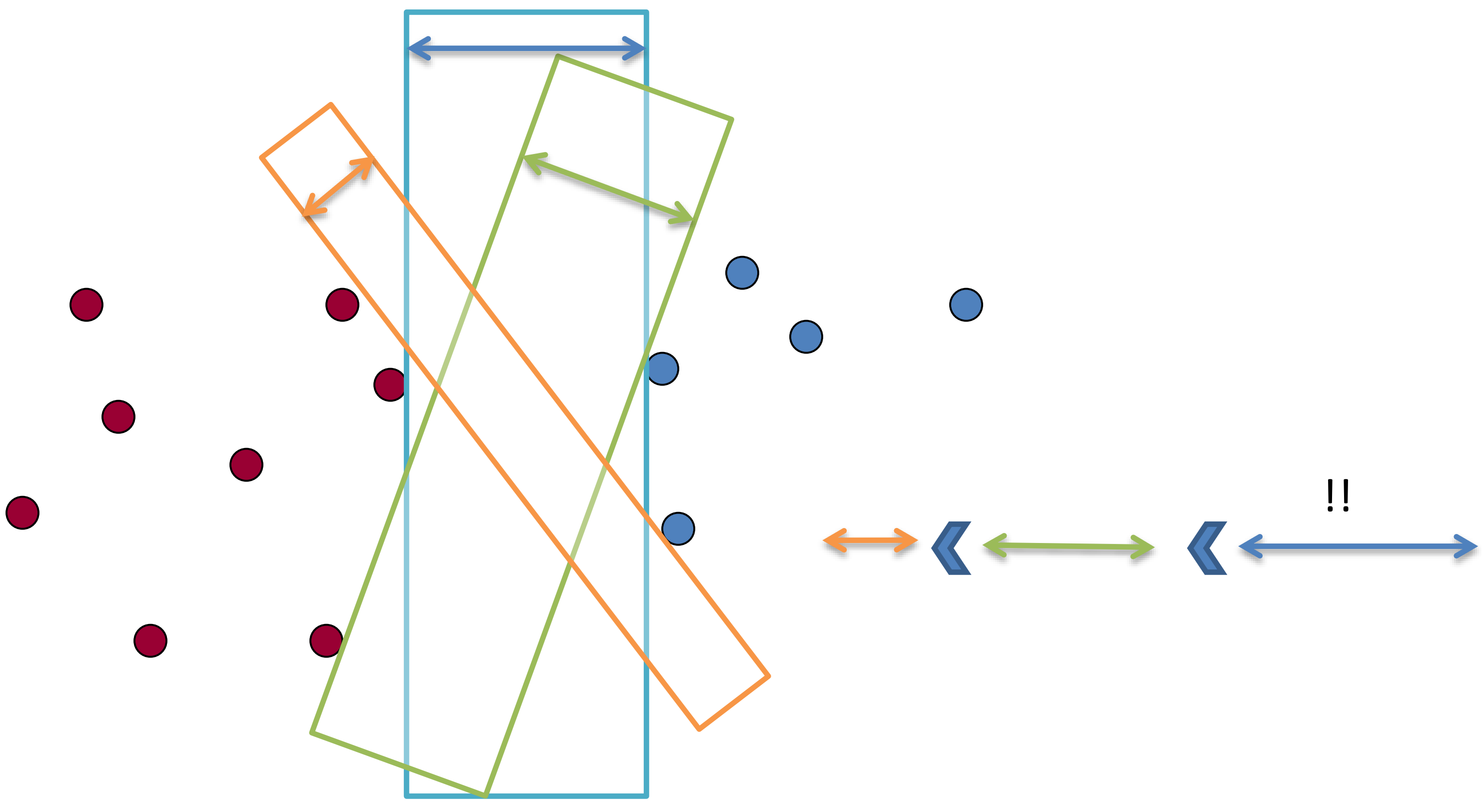
SVM – nalazi hiper-ravninu koja ima najveću **marginu** razdvajanja klasa

- margina - **udaljenost između “kritičnih” točaka** najbliže blizu plohi razdvajanja (en. decision boundary)

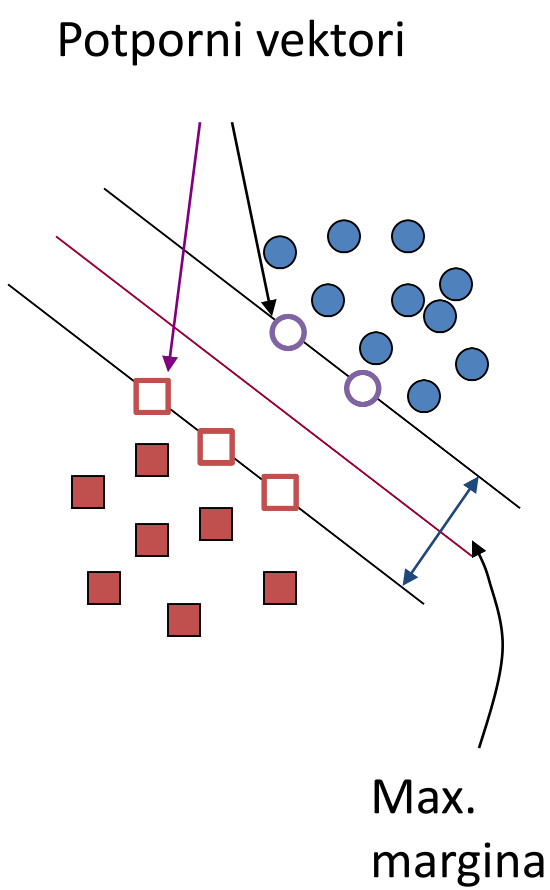
Intuicija:
Ako ne postoje točke blizu plohe razdvajanja, to znači da nemamo nesigurnih klasifikacijskih odluka !?

Druga intuicija

- Ako postavimo što je moguće veći “razmak” (marginu) između dviju klasa, postoji manje mogućnosti izbora za funkciju razdvajanja – kapacitet modela se smanjuje (manje šanse za overfitting?)



- SVM dakle **maksimiziraju *marginu m*** oko hiperravnine (plohe) razdvajanja.
    - (en. large margin classifiers)



- Kod SVM - funkcija odluke definirana je preko podskupa primjera iz skupa za učenje =>

  tzv. Potpornih vektora (en. support vectors => SV)

- Problem određivanja SV:

  kvadratni optimizacijski problem (en. quadratic programming)

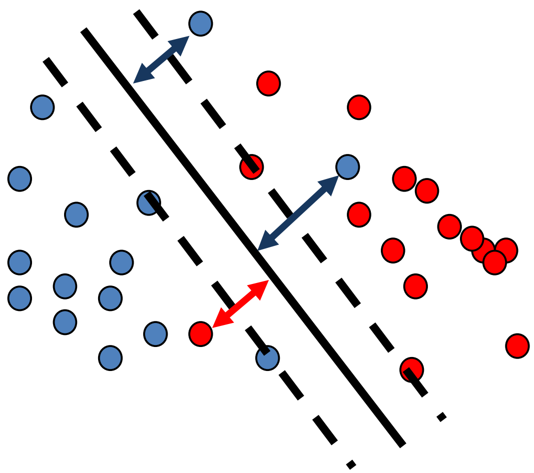
Separabilni i neseparabilni problemi

Ako se pokaže da problem nije linearno separabilan:

- Dozvoljene su greške – uz penaliziranje

Osnovni princip ostaje:

- Ploha razdvajanja u principu mora što bolje razdvajati klase (max margina)



SVM – linearno separabilni podaci (2 klase)

Geometrijska intuicija - **konveksna ljuska (Convex hull)**

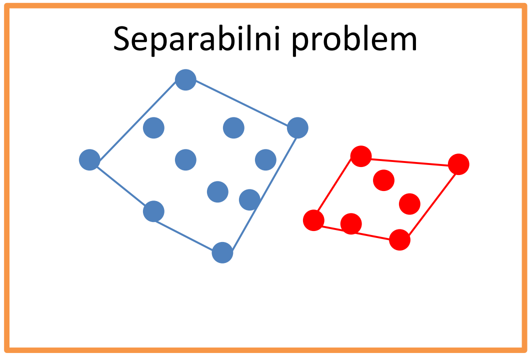
- K.LJ. – najmanji konveksni skup koji sadrži sve točke skupa točaka {$\mathbf{x}_i$}
- Za dani skup točaka {$\mathbf{x}_i$} K.LJ. je skup točaka definiran sa:

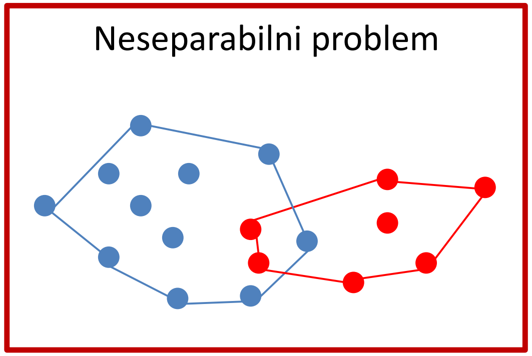
$$\boldsymbol{x} = \sum_i \alpha_i \boldsymbol{x}_i$$

Uz:

$$\alpha_i \geq 0$$

$$\sum_i \alpha_i = 1$$





- za dva skupa {$\mathbf{x}_i$} i {$\mathbf{y}_i$} vrijedi ako se njihove K.LJ. ne preklapaju – onda se radi o linearno separabilnom problemu !

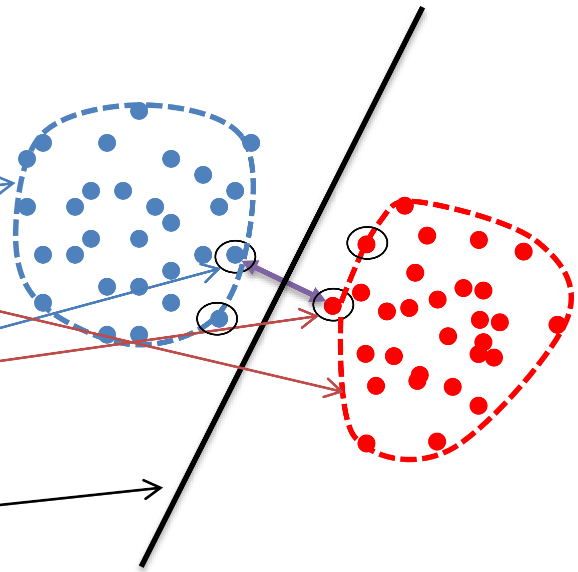
SVM – linearno separabilni podaci (2 klase)

Geometrijska intuicija - Konveksne ljuske

- K.LJ. – najmanji konveksni skup koji sadrži sve točke

Zadatak

- naći konveksnu ljusku za svaku od klasa
- Naći dvije najbliže točke u dvije konveksne ljuske
- Odrediti ravninu koja prolazi između te dvije točke

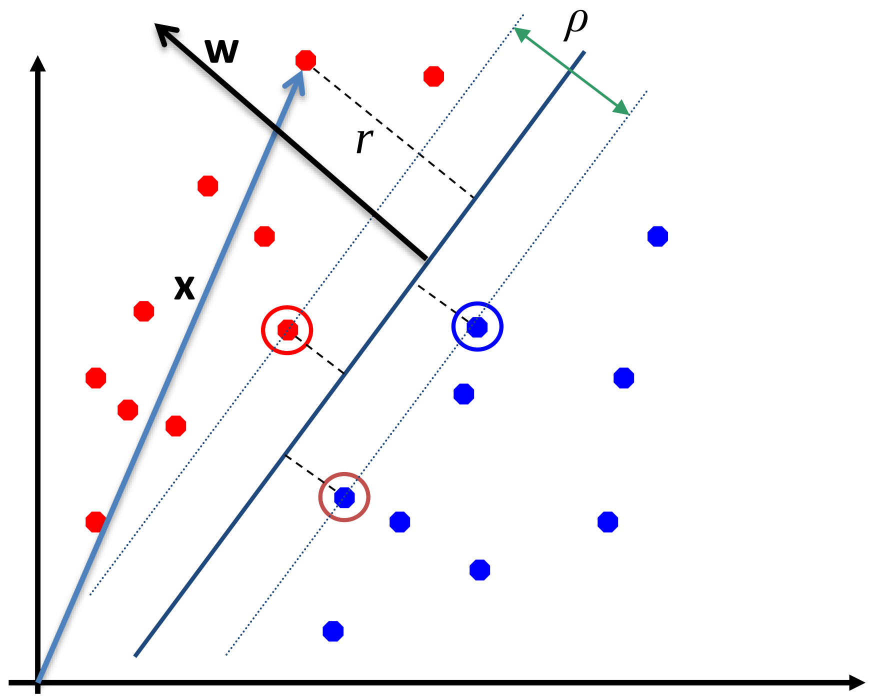


Matematička formalizacija - određivanje maksimalne margine

- **w**: normala na hiperravninu razdvajanja
- $\mathbf{x}_i$: primjer (točka) i
- y_i: klasa primjera *i* (+1/-1)
- Klasifikator je određen funkcijom ➔ $\text{sign}(\mathbf{w}^T\mathbf{x}_i + b)$
- margina $\mathbf{x}_i$ je određena sa ➔ $y_i\,(\mathbf{w}^T\mathbf{x}_i + b)$

Pojam - geometrijske margine

- Primjeri najbliže plohi su ***potporni vektori***.
- ***Margina*** *ρ* plohe razdvajanja je širina razdvajanja između potpornih vektora suprotnih klasa



Udaljenost između plohe i točke x

$$r = \frac{\mathbf{w}^T \mathbf{x} + b}{\|\mathbf{w}\|}$$

Linearni SVM - izvod (binarni klasifikacijski problem)

- Pretpostavimo da vrijedi da se sve točke nalaze najmanje na distanci 1 od plohe razdvajanja
- Tada vrijede dva ograničenja na skupu primjera za učenje $\{(\mathbf{x_i}, y_i)\}$

$$\mathbf{w}^{\mathbf{T}}\mathbf{x_i} + b \geq 1 \quad \text{ako je } y_i = 1$$

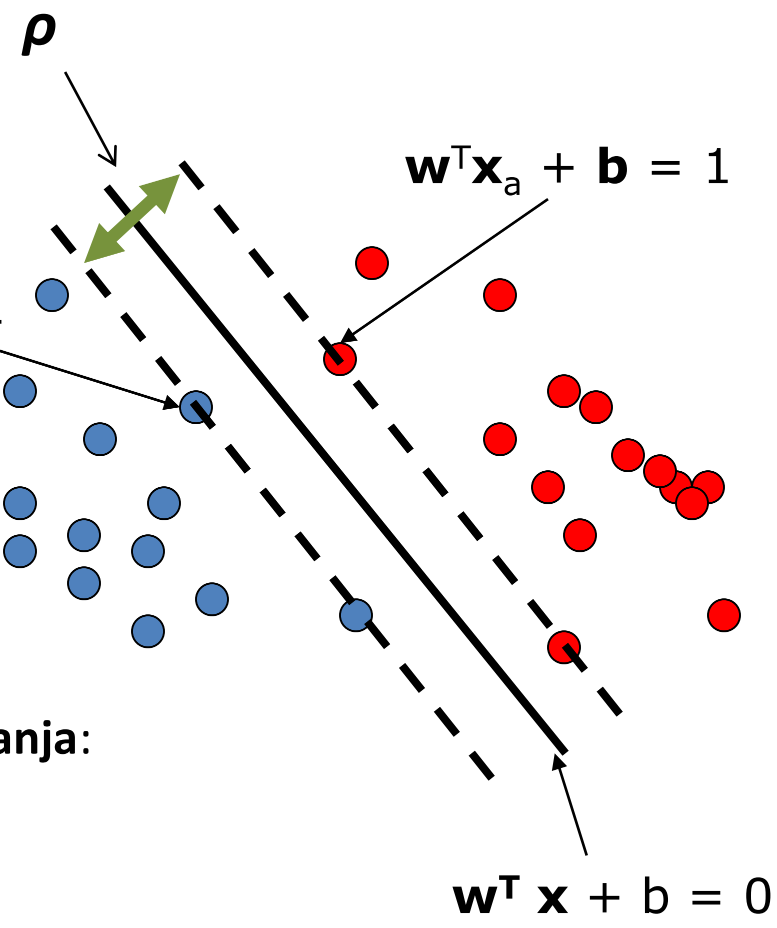
$$\mathbf{w}^{\mathbf{T}}\mathbf{x_i} + b \leq -1 \quad \text{ako je } y_i = -1$$

- Za **potporne vektore** (točke koje se nalaze na margini), ove nejednakosti su jednakosti;
- S obzirom da je udaljenost svakog primjera od plohe razdvajanja:

$$r = \frac{\mathbf{w}^T\mathbf{x} + b}{\|\mathbf{w}\|}$$

- Margina je: $\rho = \frac{2}{\|\mathbf{w}\|}$

Linearni SVM



Ploha razdvajanja

$\mathbf{w}^T \mathbf{x} + b = 0$

Uz ograničenja:

$\min_{i=1,\ldots,n} |\mathbf{w}^T\mathbf{x}_i + b| = 1$

Dolazimo do vrijednosti za marginu razdvajanja:

$\mathbf{w}^T(\mathbf{x}_a - \mathbf{x}_b) = 2$

$\rho = ||\mathbf{x}_a - \mathbf{x}_b|| = 2/||\mathbf{w}||$

Linearni SVM - formulacija optimizacijskog problema

Pronađi **w** i *b* tako da je

$$\rho = \frac{2}{\|\mathbf{w}\|} \quad \text{=> maksimalna}$$

a za sve zadane $\{(\mathbf{x}_i, y_i)\}$ mora vrijediti:

$\mathbf{w}^T\mathbf{x}_i + b \geq 1$ ako je $y_i = 1$;

$\mathbf{w}^T\mathbf{x}_i + b \leq -1$ ako je $y_i = -1$

Linearni SVM - formulacija optimizacijskog problema

- Bolja formulacija – kao minimizacijski problem
  - min $\|\mathbf{w}\|$ = max $(2/\|\mathbf{w}\|)$
- Povoljniji oblik za rješavanje
  - kvadratni optimizacijski problem uz ograničenja:

    umjesto min $\|\mathbf{w}\|$ => min $(\mathbf{w}^T\mathbf{w})$

Naći **w** i *b* - takve da je:

$\boldsymbol{\Phi}(\mathbf{w}) = \frac{1}{2}\,\mathbf{w}^T\mathbf{w}$ je minimalno;

A za sve $\{(\mathbf{x}_i, y_i)\}$ vrijedi :

$y_i\,(\mathbf{w}^T\mathbf{x}_i + b) \geq 1$

Rješavanje optimizacijskog problema => optimizacija uz ograničenja

Nađi **w** i b - takve da je:

$\boldsymbol{\Phi}(\mathbf{w}) = ½\, \mathbf{w}^T\mathbf{w}$ - minimalno;

A za sve $\{(\mathbf{x_i}, y_i)\}$ vrijedi :

$y_i\,(\mathbf{w}^T\mathbf{x_i} + b) \geq 1$

- Ova formulacija – minimizacija kvadratne funkcije uz linearna ograničenja (nejednakosti) => **primalna formulacija**
- Dobro poznati i rješivi problem – puno (više ili manje složenih) algoritama za njihovo rješavanje (MATLAB, MATHEMATICA)
- Standardno za SVM - prevođenje u ***dualni problem* u kojem se tzv. *Lagrange-ovim multiplikatorima*** optimiraju težinski faktori koji penaliziraju svako ograničenje iz primalnog problema.

Lagrange-ovi multiplikatori => optimizacija uz ograničenja

Minimiziraj funkciju f(**x**) uz <u>ograničenja u obliku jednakosti g(**x**)=0</u>

- Naći minimum funkcije $f(\boldsymbol{x}), \boldsymbol{x} \in R^d$, uz ograničenja $\mathrm{g}(\boldsymbol{x}) = 0, i = 1 \ldots m$
- Tada postoje:

  $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, .., \lambda_m)$ (Lagr. multiplikatori) za koje u minimumu $\boldsymbol{x}_*$, koji zadovoljava sva ograničenja, vrijedi:

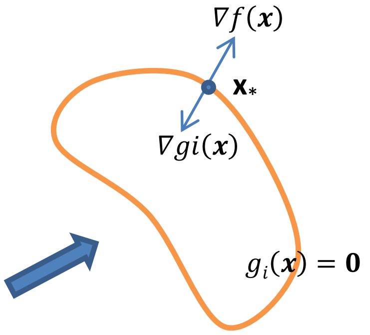
$$\nabla f(\boldsymbol{x}) = \lambda_1 \nabla g_1(\boldsymbol{x}_*) + \lambda_2 \nabla g_2(\boldsymbol{x}_*) + \cdots + \lambda_m \nabla gm(\boldsymbol{x}_*)$$

Ako uvedemo tzv. Lagrangeovu funkciju

$L(\boldsymbol{x}, \boldsymbol{\lambda}) = f(\boldsymbol{x}) + \sum_i \lambda_i g_i(\boldsymbol{x})$, proizlazi da vrijedi u $\mathbf{x}_*$:

$$\frac{\partial L}{\partial \lambda_i} = g_{i(\boldsymbol{x})} = 0$$

$$\nabla L_x = \nabla f(\boldsymbol{x}) + \lambda_1 \nabla g_1(\boldsymbol{x}) + \lambda_2 \nabla g_2(\boldsymbol{x}) + \cdots + \lambda_m \nabla gm(\boldsymbol{x})$$

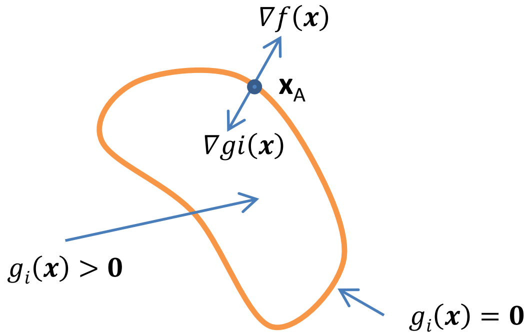


Moramo pronaći stacionarnu točku $L(\boldsymbol{x}, \boldsymbol{\lambda})$ - s obzirom na **x** i $\boldsymbol{\lambda}$!

Lagrange-ovi multiplikatori => optimizacija uz ograničenja

Minimiziraj funkciju f(**x**) uz ograničenja tipa <u>nejednakosti g(**x**)≥0</u>

- Rješenje $\mathbf{x}_*$ može biti:
  - unutar plohe $g_i(\mathbf{x})$ – tada vrijedi $g_i(\mathbf{x}) > 0$. U tom slučaju ograničenje *i* nije aktivno => $\lambda_i = 0$!
  - Na $g_i(\mathbf{x})=0$ – tada je ograničenje aktivno => $\lambda_i \geq 0$!



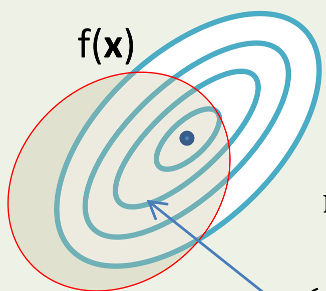
Lagrange-ovi multiplikatori => optimizacija uz ograničenja

Slučaj sa ograničenjem u obliku nejednakosti

$g_i(\boldsymbol{x}) \geq 0$

a) Aktivno ograničenje

$\min f(x) \neq \min(L(\boldsymbol{x}))$

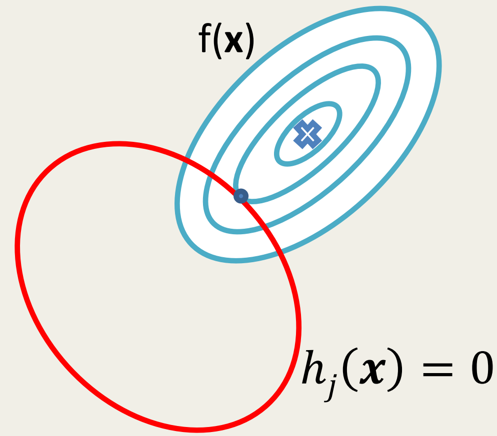


$g_i(\boldsymbol{x}) \geq 0$

b) Ograničenje nije aktivno

$\min f(x) = \min(L(\boldsymbol{x}))$

Slučaj sa ograničenjem u obliku jednakosti



✖ $\Rightarrow$ Min f(**x**) (neograničeni minimum)

• $\Rightarrow$ Min L(**x**) (ograničeni minimum f(**x**))

Lagrange-ovi multiplikatori => optimizacija uz ograničenja

Karush-Kuhn-Tucker (KKT) uvjeti
(generalizacija metode LM na ograničenja u obliku nejednakosti)

- Rješenje za nelinearne optimizacijski problem tipa:

$$min_{\boldsymbol{x}}\, f(\boldsymbol{x})$$

$$s.t. \quad g_i(\boldsymbol{x}) \geq 0$$

- Dobije se optimizacijom s obzirom na **x** i $\boldsymbol{\lambda}$:

$$L(\boldsymbol{x}, \boldsymbol{\lambda}) = f(\boldsymbol{x}) + \sum_i \lambda_i g_i(\boldsymbol{x})$$

  - Uz ograničenja u (ograničenom) minimumu $\boldsymbol{x}^*$ vrijedi:

$g_i(\boldsymbol{x}^*) \geq 0;$
$\lambda_i \geq 0;$
$\lambda_i g_i(\boldsymbol{x}^*) = 0;$

Lagrange-ova funkcija i min-max dualnost

Primalna formulacija:

$$\min_{w} \max_{\lambda} L(w, \lambda)$$

Dualna formulacija:

$$\max_{\lambda} \min_{w} L(w, \lambda)$$

Za konveksni problem:

$$\min_{w} \max_{\lambda} L(w, \lambda) = \max_{\lambda} \min_{w} L(w, \lambda)$$

Wolfe dual:

$$\max_{\lambda \geq 0} L(w, \lambda)$$

uz $\quad \frac{\partial}{\partial w} L(w, \lambda) = 0$

- Primalna formulacija Lagr. Multiplikatori + KKT uvjeti => dualna formulacija

$$L(w, \lambda, b) = \frac{1}{2}||\boldsymbol{w}||^{\boldsymbol{2}} - \sum_{\boldsymbol{j=1}}^{\boldsymbol{N}} \lambda_j \left\{ y_j \left(\boldsymbol{w}^T \boldsymbol{x}_j + b \right) - \boldsymbol{1} \right\}$$

$$\frac{\partial L}{\partial w} = 0 \Leftrightarrow \boldsymbol{w} - \sum_{j=1}^{N} \lambda_j y_j \boldsymbol{x}_j = 0 \Leftrightarrow \boldsymbol{w} = \sum_{j=1}^{N} \lambda_j y_j \boldsymbol{x}_j$$

$$\frac{\partial L}{\partial b} = 0 \Leftrightarrow \sum_{j=1}^{N} \lambda_j y_j = 0$$

Uz

$$\lambda_i \geq= 0, i = 1, \dots, N$$
$$\lambda_i(y_i(\boldsymbol{w}^T \boldsymbol{x}_i + b) - 1) = 0, i = 1, \dots N$$
$$(y_i(\boldsymbol{w}^T \boldsymbol{x}_i + b) - 1) \geq 0, i = 1, \dots N$$

Prevođenje u ***dualni problem***

- Grace Wahba - Representer teorem

$$w = \sum_{j=1}^{N} \lambda_j \, y_j x_j$$

- Ako zamijenimo **w** u primalnom problemu gornjim izrazom

$$\boldsymbol{w^T x} + \boldsymbol{b} = \left(\sum_{j=1}^{N} \lambda_j \, y_j \boldsymbol{x}_j \right)^T \boldsymbol{x} + b = \sum_{j=1}^{N} \lambda_j \, y_j \left(\boldsymbol{x}_j{}^T \boldsymbol{x} \right) + b$$

$$\boldsymbol{w^T w} = \left(\sum_{j=1}^{N} \lambda_j \, y_j \boldsymbol{x}_j \right)^T \left(\sum_{k=1}^{N} \lambda_k \, y_k \boldsymbol{x}_k \right) = \sum_{jk} \lambda_j \, \lambda_k y_j y_k (\boldsymbol{x}_j{}^T \boldsymbol{x}_k)$$

SVM - optimizacijski problem

Dualna formulacija:

Naći $\lambda_1 \ldots \lambda_N$ tako da je:

Max Q(λ) **=** $\sum \lambda_i - \frac{1}{2} \sum \sum \lambda_i \lambda_j y_i y_j \mathbf{x}_\mathbf{i}^\mathbf{T} \mathbf{x}_\mathbf{j}$

Te da vrijedi

(1) $\sum \lambda_i y_i = 0 \quad j = 1,N$

(2) $\lambda_i \geq 0 \ \forall \ i = 1,\ldots,N$

Rješenje optimizacijskog problema

$$\mathbf{w} = \sum \lambda_i y_i \mathbf{x_i} \qquad b = y_k - \mathbf{w}^T\mathbf{x_k}, \qquad \text{za } \mathbf{x_k} \text{ za koji je } \lambda_k \neq 0 \text{ (t.j. SV)}$$

- **Svaki λ_i koji je različit od 0** – zapravo znači da je taj $\mathbf{x_i}$ – potporni vektor.
- Rješenje je proporcionalno sumi umnožaka - *unutarnjih produkta* $\mathbf{x_i}^T\mathbf{x}$ između nove točke (primjera) **x** i svih potpornih vektora $\mathbf{x_i}$!
- Treba zapamtiti i da je u rješavanju optimizacijskog problema korišten produkt $\mathbf{x_i}^T\mathbf{x_j}$ između svih parova točaka skupa za učenje.

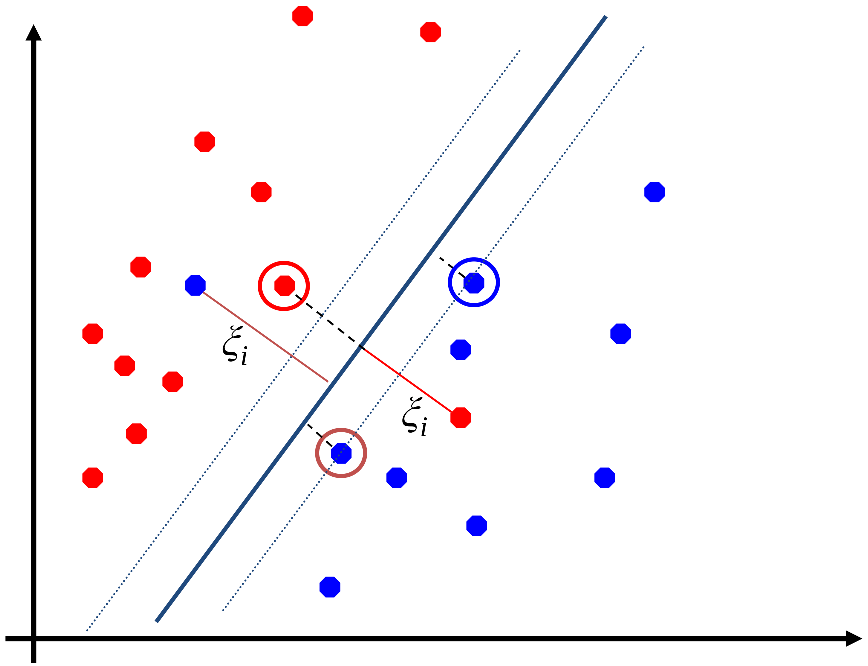
- Konačno – klasifikacijska funkcija ima ovaj oblik:

$$f(\mathbf{x}) = \sum_i (\lambda_i y_i \mathbf{x_i}^T\mathbf{x} + b)$$

- U **linearnim regresijskim problemima** učimo i koristimo ***w*** *(~ broj varijabli)*
- ***K-nn algoritmi*** **-** koristimo *sve trening točke (~N)* da bi radili predikcije
- **SVM** – koristi točke iz trening skupa ($\lambda_i \geq 0$) da bi radili predikcije – nešto između LR i k-nn ?!

Što ako problem nije linearno separabilan?

- Dodajemo nove varijable ξ_i *(en. slack variables)*
  - dozvoljavanje krivih klasifikacija za „teške” ili “šumovite” primjere



Izvod – SVM s “mekanom” marginom (en. Soft margin)

- Stara formulacija:

Nađi **w** i b - takve da je:

$\mathbf{\Phi}(\mathbf{w}) = ½\, \mathbf{w}^T\mathbf{w}$ is minimalno;

A za sve $\{(\mathbf{x_i}, y_i)\}$ vrijedi :

$y_i\, (\mathbf{w}^T\mathbf{x_i} + b) \geq 1$

- Nova formulacija sa ξ_i :

Nađi **w** i b - takve da je:

$\mathbf{\Phi}(\mathbf{w}) = ½\, \mathbf{w}^T\mathbf{w} + C \cdot \Sigma \cdot \xi_i$ - minimalno;

A za sve $\{(\mathbf{x_i}, y_i)\}$ vrijedi:

$y_i\, (\mathbf{w}^T\mathbf{x_i} + b) \geq 1 - \xi_i$ – te da vrijedi $\xi_i \geq 0$ za sve i

Rješenje - SVM s “mekanom” marginom (en. Soft margin)

- Dualni problem za SVM s “mekanom” marginom :

Nađi $\lambda_1 \ldots \lambda_N$ tako da je

$$\mathbf{max}\ \mathbf{Q}(\lambda) = \sum \lambda_i - \tfrac{1}{2}\sum\sum \lambda_i \lambda_j y_i y_j \mathbf{x_i}^T \mathbf{x_j}$$

Te da vrijedi

(1) $\sum \lambda_i y_i = 0$

(2) $0 \leq \lambda_i \leq C \ \forall \lambda_i$

- I dalje vrijedi: $\mathbf{x_i}$ sa $\lambda_i > 0$ – *su potporni vektori*.
- U dualnoj formulaciji – nema **slack v**arijabli, samo ograničenje na λ (λ < C)
- Rješenje dualnog problema (soft margin):

$$\mathbf{w} = \sum \lambda_i y_i \mathbf{x_i}$$

$$b = y_k(1 - \xi_k) - \mathbf{w}^T \mathbf{x_k},\ k = \underset{k}{\operatorname{argmax}}\ \lambda_k$$

w nije potreban kod procesa klasifikacije – kao i prije !

$$f(\mathbf{x}) = \lambda_i y_i \mathbf{x_i}^T \mathbf{x} + b$$

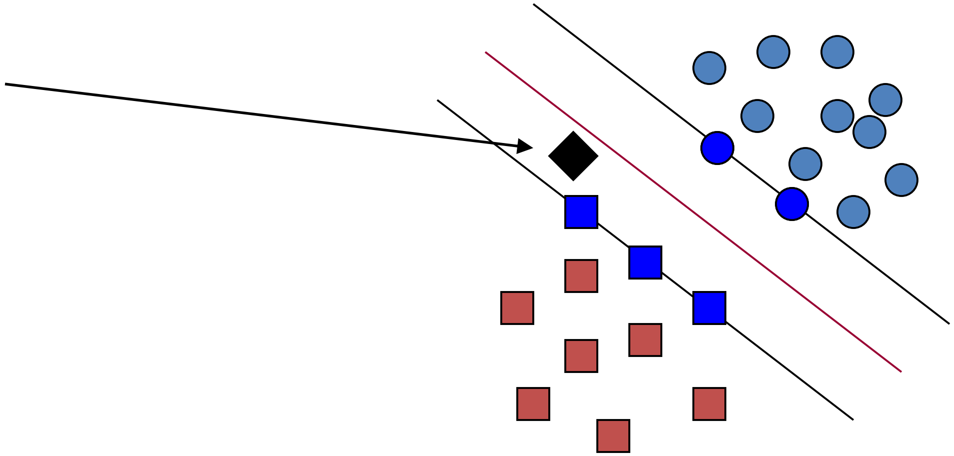
Klasifikacija

- Uz novu točku (x_1,x_2),
  - Odrediti projekciju na normalu plohe razdvajanja:
  - 2 dimenzije: p = $w_1x_1+w_2x_2+b$.
  - To ustvari znači ($\boldsymbol{wx} + b = \Sigma\ \lambda\ _i y_i \mathbf{x_i}^\mathbf{T}\mathbf{x} + b$)
  - Moguće proširenje => ako stavimo neku graničnu mjeru t >0 (pouzdanost)

f > t: => 1

f < -t: => -1

Inače: suzdržan



Linearni SVM - sažetak

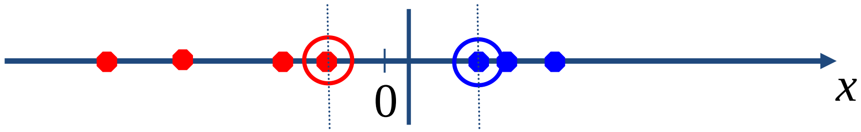
- Klasifikator je posebno određena *ploha razdvajanja* (en. *separating hyperplane)*
- Najvažnije točke - točke iz skupa za učenje koje definiraju plohu razdvajanja
  => potporni vektori
- Potporni vektori se pronalaze – optimizacijskim algoritmom
- Rješavani problem je kvadratni optimizacijski problem s linearnim ograničenjima (=konveksni o.p.).
- U dualnoj formulaciji problema i rješenja – potporni vektori odnosno točke iz skupa za učenje se pojavljuju u skalarnim produktima:

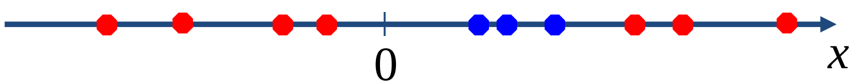
Nađi $\lambda_1 \ldots \lambda_N$ tako da je
$\mathbf{Q}(\lambda) = \Sigma\, \lambda_i - ½\Sigma\Sigma\, \lambda_i\, \lambda_j y_i y_j \mathbf{x_i}^T\mathbf{x_j}$ - maksimalno i vrijedi:
(1) $\Sigma\, \lambda_i y_i = 0$
(2) $0 \leq \lambda_i \leq C \;\forall\; \lambda_i$

$f(\mathbf{x}) = \Sigma\, \lambda_i y_i \mathbf{x_i}^T\mathbf{x} + b$

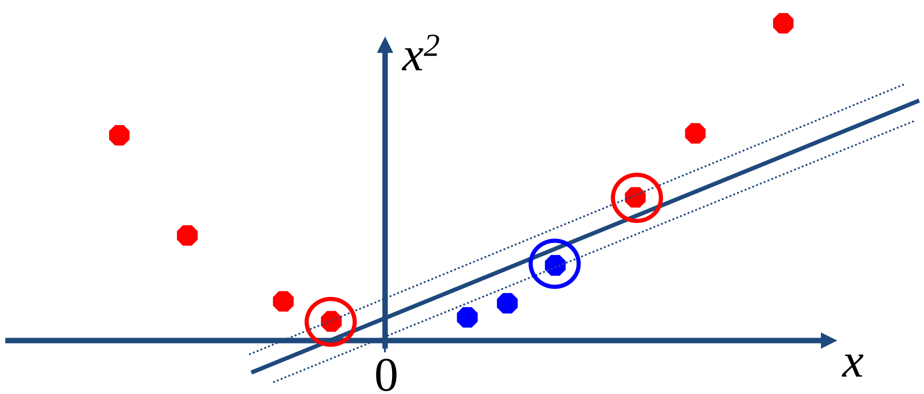
Nelinearni SVM

- Za probleme koji su linearno separabilni (i uz razumni šum) – Linearni SVM je OK :



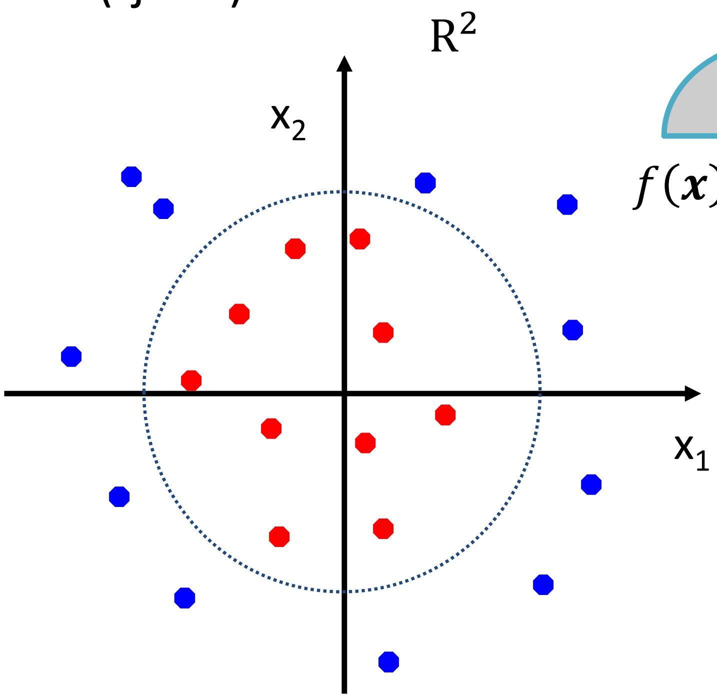
- No, što ako to ne vrijedi ?
- 

- … mapiranje podataka u (još) više-dimenzionalni prostor !

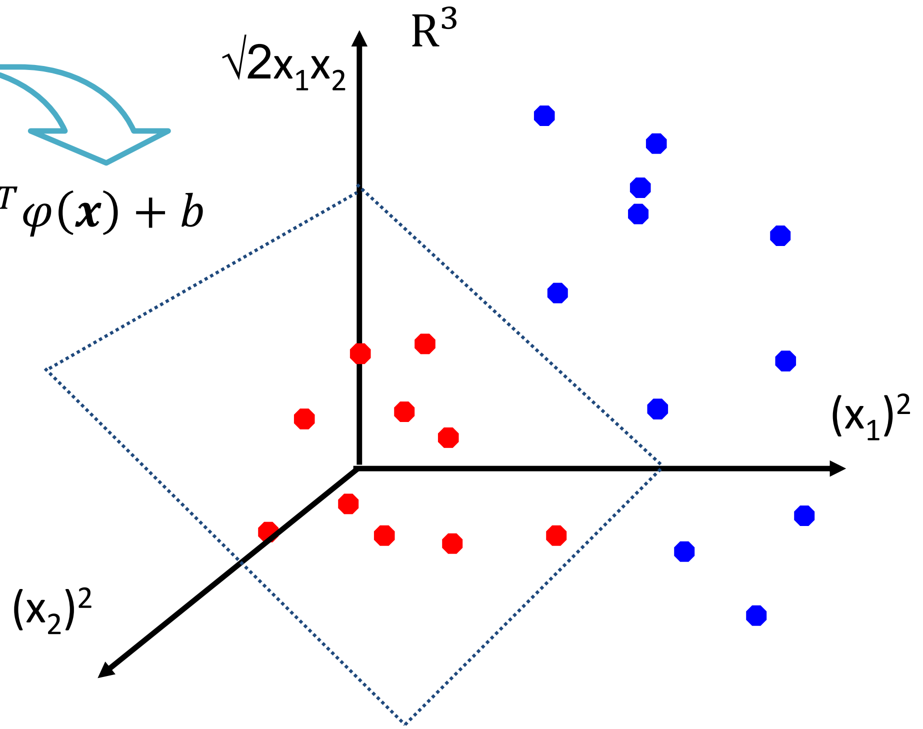


Nelinearni SVM: prostori novih dimenzija (atributa)

- Osnovna ideja: originalni prostor varijabli – možemo mapirati u novi-više-dimenzionalni prostor - gdje će naš problem biti (linearno) separabilan (rješiv)



$$f(\boldsymbol{x}) = w^T \varphi(\boldsymbol{x}) + b$$



Nelinearni SVM: prostori novih dimenzija (atributa)

- Primalni problem u transformiranom prostoru

$$f(\boldsymbol{x}) = \boldsymbol{w}^T \varphi(\boldsymbol{x}) + b$$

- Učenje $\boldsymbol{w}$ u R^D - transformiranom prostoru

$$\min_{\boldsymbol{w} \in \boldsymbol{R}^{\boldsymbol{D}}} \|\boldsymbol{w}\|^2 + C \sum_{i}^{N} (1 - y_i f(\boldsymbol{x}_i))$$

- Problem – ako je D>>d !

Nelinearni SVM: prostori novih dimenzija (atributa)

- Dualni problem u transformiranom prostoru

$$f(\boldsymbol{x}) = \alpha_i y_i \varphi(\boldsymbol{x_i})^T \varphi(\boldsymbol{x_i}) + b$$

- Učenje u transformiranom prostoru

$$\max_{\alpha_i \geq 0} \sum_i \alpha_i - \frac{1}{2} \sum_{jk} \alpha_j \alpha_k y_j y_k \varphi(\boldsymbol{x_i})^T \varphi(\boldsymbol{x_i})$$

s.t.

$$0 \leq \alpha_i \leq C \; za \; \forall i, \qquad uz \; \sum_i \alpha_i \, y_i = 0$$

- Ako izračunamo skalarne produkte $\varphi(\boldsymbol{x_i})^T \varphi(\boldsymbol{x_i})$ - učimo samo N α_i !
- Možemo li $\varphi(\boldsymbol{x_i})^T \varphi(\boldsymbol{x_i})$ pojednostavniti ?

"Kernel trik"

- Skalarni produkt bi trebao izgledati drukčije:

$$K(\mathbf{x}_i,\mathbf{x}_j)= \phi(\mathbf{x}_i)^T\phi(\mathbf{x}_j)$$

- *Kernel funkcija* - funkcija koja korespondira unutarnjem produktu u nekom "ekspandiranom" prostoru novih varijabli
  - Ne moramo uopće znati $\phi(\mathbf{x})$ da bi „operirali" u novom prostoru !!

“Kernel trik”

Primjer: originalni prostor

2-dimenzionalni vektori $\mathbf{x}=[x_1 \;\; x_2]$;

Neka je $K(\mathbf{x_i},\mathbf{x_j})=(1 + \mathbf{x_i}^{\mathbf{T}}\mathbf{x_j})^2$,

Pokažimo da vrijedi $K(\mathbf{x_i},\mathbf{x_j})= \phi(\mathbf{x_i})^{\mathbf{T}}\phi(\mathbf{x_j})$,

ako je $\phi(\mathbf{x}) = [1, \; x_1^2, \sqrt{2}\, x_1x_2, \; x_2^2, \; \sqrt{2}x_1 \; \sqrt{2}x_2]$

$$K(\mathbf{x_i},\mathbf{x_j})=(1 + \mathbf{x_i}^{\mathbf{T}}\mathbf{x_j})^2 = 1+ x_{i1}^2x_{j1}^2 + 2\, x_{i1}x_{j1}\, x_{i2}x_{j2}+ x_{i2}^2x_{j2}^2 + 2x_{i1}x_{j1} + 2x_{i2}x_{j2}=$$

$$= [1 \;\; x_{i1}^2 \;\; \sqrt{2}\, x_{i1}x_{i2} \;\; x_{i2}^2 \;\; \sqrt{2}x_{i1} \;\; \sqrt{2}x_{i2}]^{\mathbf{T}} [1 \;\; x_{j1}^2 \;\; \sqrt{2}\, x_{j1}x_{j2} \;\; x_{j2}^2 \;\; \sqrt{2}x_{j1} \;\; \sqrt{2}x_{j2}]$$

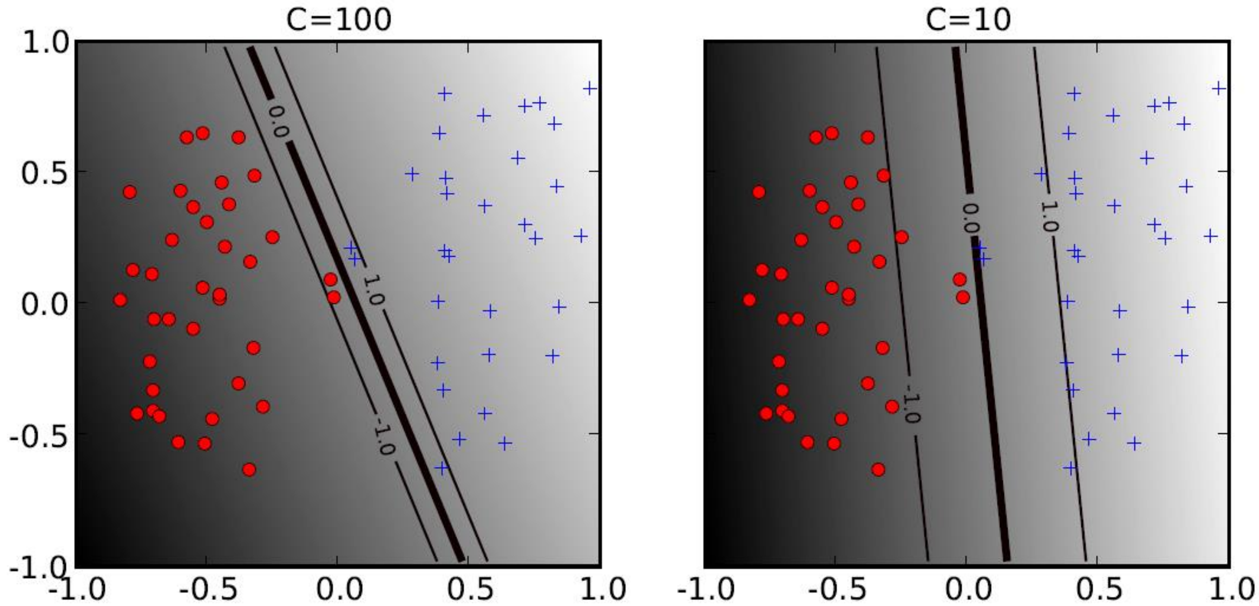
$$= \phi(\mathbf{x_i})^{\mathbf{T}}\phi(\mathbf{x_j}) \;!$$

Kerneli

- Zašto kerneli ?
  - Efikasno učenje
  - Mapiranje u bolje (reprezentacijski) – višedimenzionalne prostore
    - Omogućava pretvaranje neseparabilnih problema u separabilne.

- Uobičajeni kerneli
  - Linearni $K(\mathbf{x}_i,\mathbf{x}_j) = (\mathbf{x}_i^T\mathbf{x}_j)$
  - Polinomijalni $K(\mathbf{x}_i,\mathbf{x}_j) = (1+\mathbf{x}_i^T\mathbf{x}_j)^d$
  - RBF $K(\mathbf{x}_i,\mathbf{x}_j) = e^{\frac{-\|\mathbf{x}_i-\mathbf{x}_j\|}{2\sigma^2}}$
  - Sigmoid $K(\mathbf{x}_i,\mathbf{x}_j) = \tanh(\gamma\mathbf{x}_i^T\mathbf{x}_j + r)$

Linearni SVM – efekt različitih vrijednosti parametra C

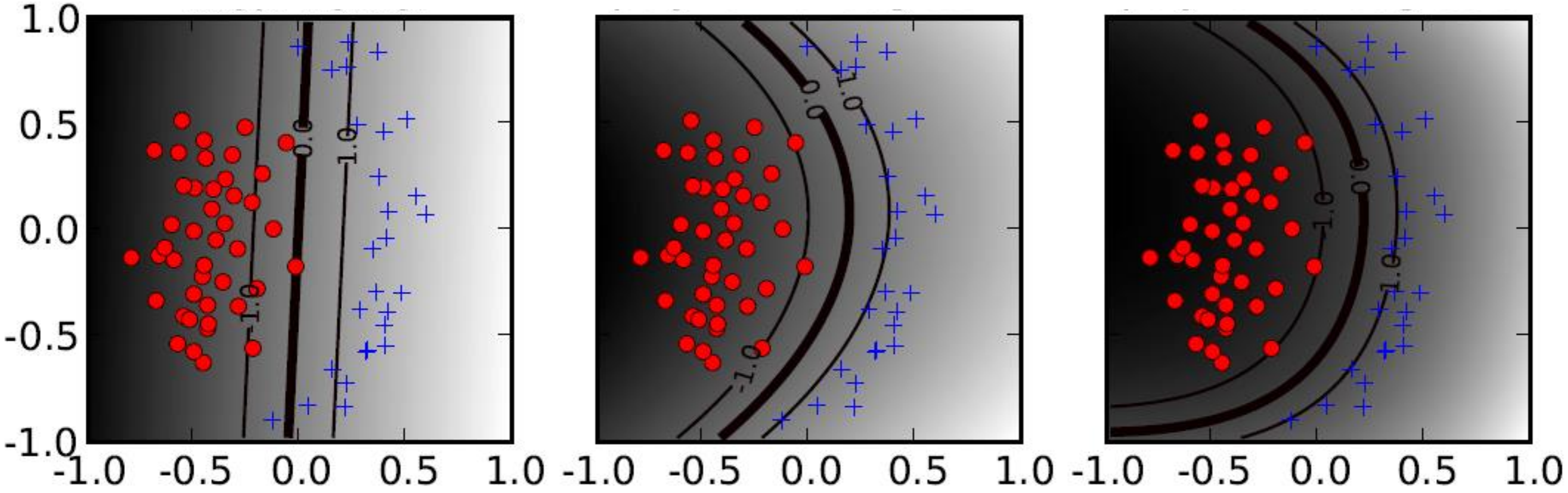


Prisjetimo se gdje smo vidjeli C

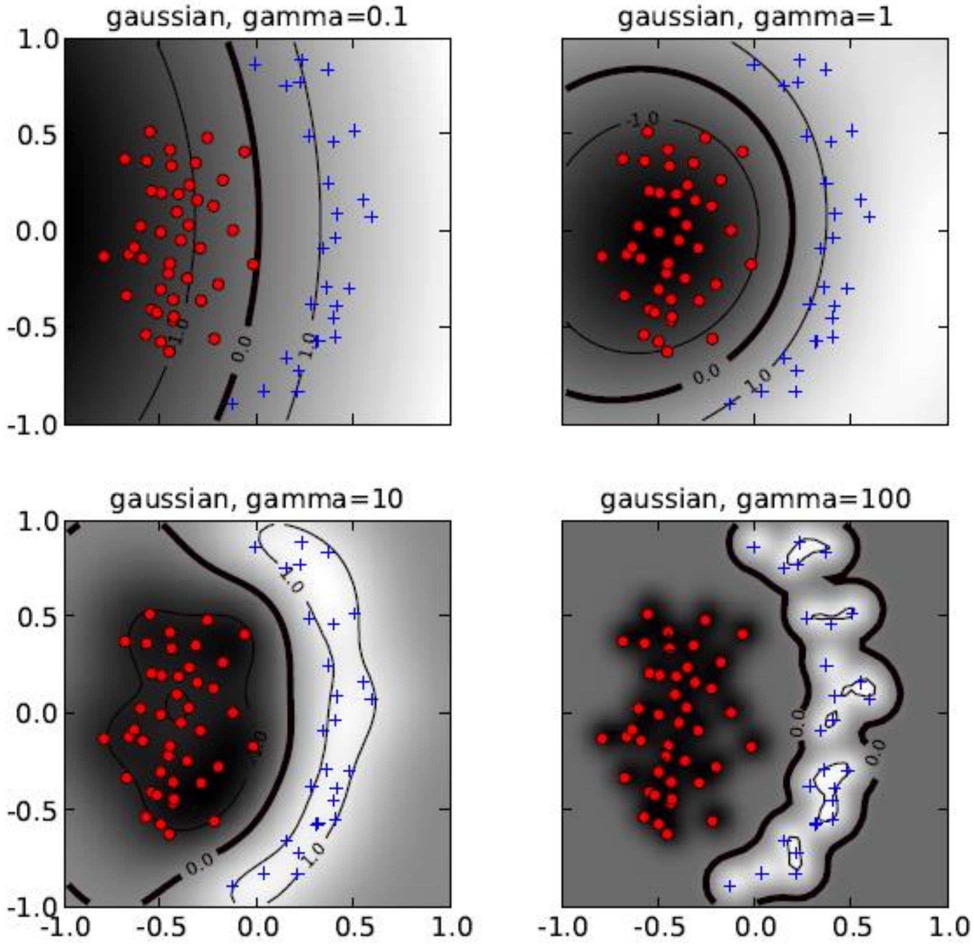
$$\boldsymbol{\Phi}(\mathbf{w}) = \tfrac{1}{2}\,\mathbf{w}^{T}\mathbf{w} + C \cdot \Sigma \cdot \xi_i$$

Odabir kernela – efekti
Polinomni kerneli (C=const.)

Linearni kernel polinom 2 stupnja polinom 5 stupnja



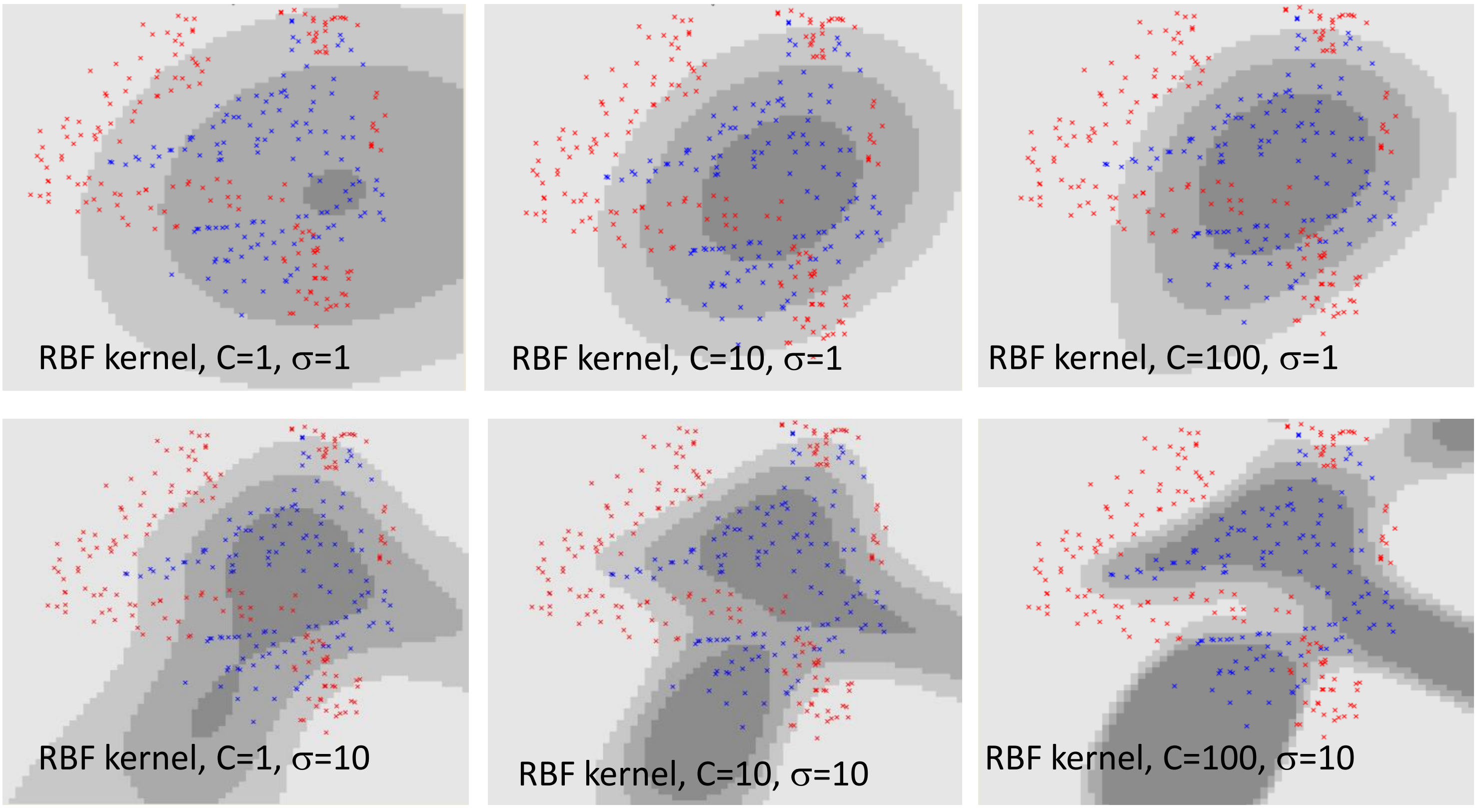
Odabir parametara kernela – efekti



Gaussian kernel, C=const.

$$k(\boldsymbol{x}, \boldsymbol{x}') = e^{(-\gamma\|\boldsymbol{x}-\boldsymbol{x}'\|^2)}$$

Odabir parametara parametara kernela i faktora C



$$\gamma = 1/\sigma$$

SVM - Sa dvije klase na više klasa

Većina SVM metoda je za binarne klasifikacijske probleme

Uobičajen pristup za višeklasne probleme je preko redukcije problema na više binarnih klasifikacijskih problema, npr. za K-klasni problem:

- Jedan-nasuprot-ostalih => (K – binarnih klasifikatora)
- Svaki-protiv-svakog => (K*(K-1)/2 – binarnih klasifikatora)
- ECOC – Error Correcting Output Codes (… u jednom od slijedećih predavanja)

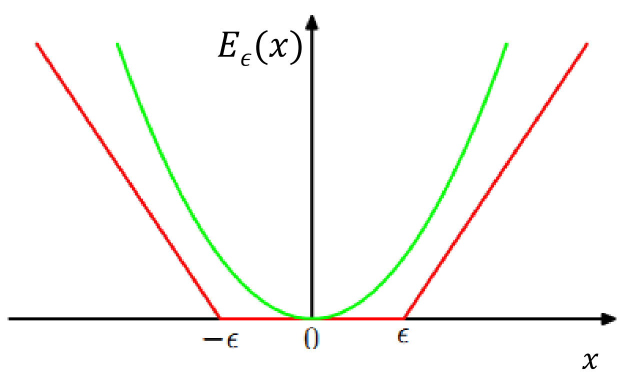
SVM – Regresija

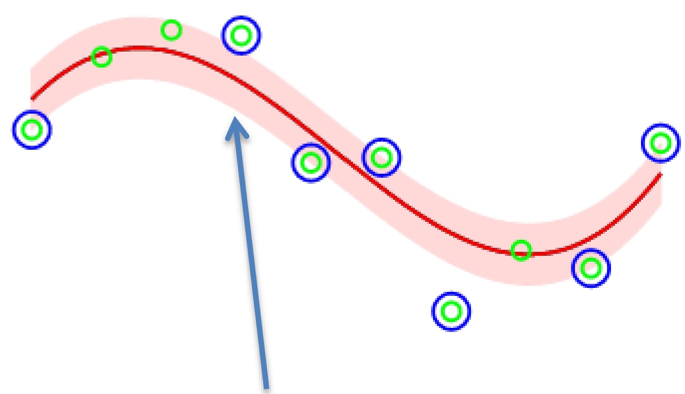
Problem

$$\min_{\boldsymbol{w}} C \sum_{n} E_{\epsilon}\left(y_n - f(x_n)\right) + \lambda \|\boldsymbol{w}\|$$

Gdje je

$$E_{\epsilon}\left(y_n - f(x_n)\right) = \begin{cases} 0 \quad za \quad |y_n - f(x_n)| < \epsilon \\ |y_n - f(x_n)| - \epsilon, \qquad inače \end{cases}$$





ϵ – insensitive tube

Karakteristike SVM-a

- Jedna od najuspješnijih metoda strojnog učenja ukupno gledajući
- Popularna – linearni SVM + text mining
- U praksi – mnoge druge metode rade otprilike isto dobro
  - Postoji mnogo usporedbi koje to pokazuju

Važno je: Razumijevanje i iskustvo za efektivno korištenje

- Odabir kernela, odabir parametara (C, parametri kernela)
  - Minimalno potrebno - napraviti tzv. „grid search” eksperiment preko intervala parametara
- kombinirati sa selekcijom varijabli (postoje SVM varijante koje implicite uključuju selekciju varijabli – min ∞ norme **w**)

Sažetak

- Definiranje plohe razdvajanja preko potpornih vektora
  - Potporni vektor = “kritične” točke (primjeri) najbliži plohi razdvajanja
- Optimizacija margine razdvajanja primjera različitih klasa – kvadratni optimizacijski problem s linearnim ograničenjima
- Pojam kernela:
  - Moćan alat za mapiranje u visoko-dimenzionalne prostore, kao i (re)definiranje metrika sličnosti

SVM literatura i materijali

The Elements of Statistical Learning
Hastie, Tibshirani, Friedman (1st ed - ch. 11)

A Tutorial on Support Vector Machines for Pattern Recognition (1998) Christopher J. C. Burges

Christopher M. Bishop. Pattern recognition and machine learning, Springer, 2006 .

Kristin P. Bennet and Campbell. Support Vector Machines: Hype or Hallelujah? SIGKDD Explorations, vol. 2, 2000

Michael E. Mavroforakis and Sergios Theodoridis. A Geometric Approach to Support Vector Machine (SVM) Classification IEEE Transactions on Neural Networks, 2006

Literatura

SVM Software preporuke

WEKA – SMO (Sequential Minimization Opt. - SVM)
R – *e1071, kernlab* – paketi

SVM-Light (T Joachims)
LibSVM (Chang &Lin)