Strojno učenje
7
Linearne metode
&
SVM

Tomislav Šmuc

Literatura

Linearne metode
The Elements of Statistical Learning
Hastie, Tibshirani, Friedman (1st ed - ch. 4)

SVM
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

A User's Guide to Support Vector Machines, Ben-Hur & Weston

T. Joachims, *Learning to Classify Text using Support Vector Machines*. Kluwer, 2002.
Classic' Reuters data set: http://www.daviddlewis.com /resources /testcollections/reuters21578/

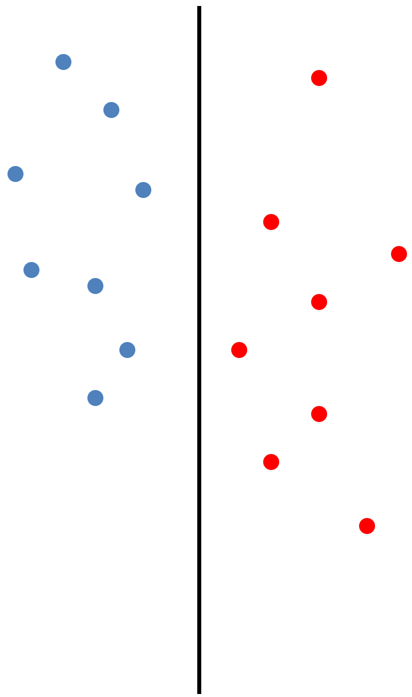
SVM Software:
http://www.kernel-machines.org/

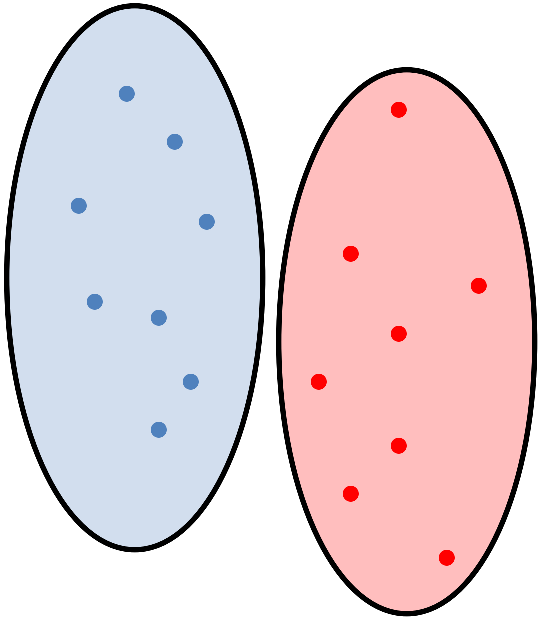
- Diskriminativni

(Učenje linije koja razdvaja klase)

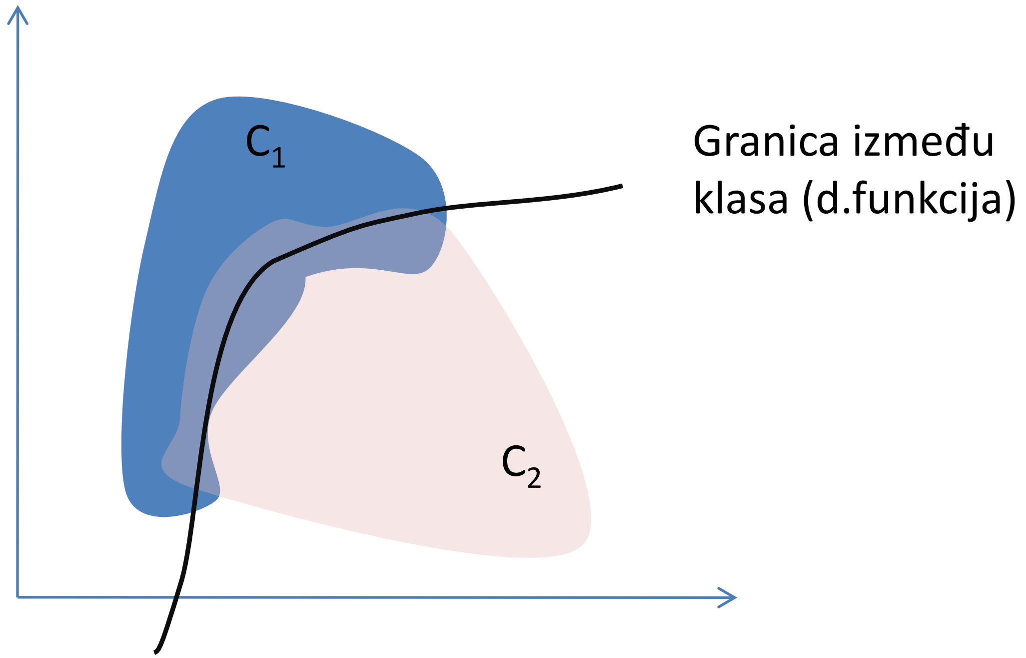
- Generativni

Učenje modela za svaku pojedinu klasu





Diskriminativna funkcije



Fisherova linearna diskriminantna metoda

Skup N d-dimenzionalnih primjera ($\mathbf{x}_1$,$\mathbf{x}_2$,$\mathbf{x}_3$,..., $\mathbf{x}_N$) – točaka

N_1 – broj primjera u podskupu D_1 klase y_1
N_2 – broj primjera u podskupu D_2 klase y_2

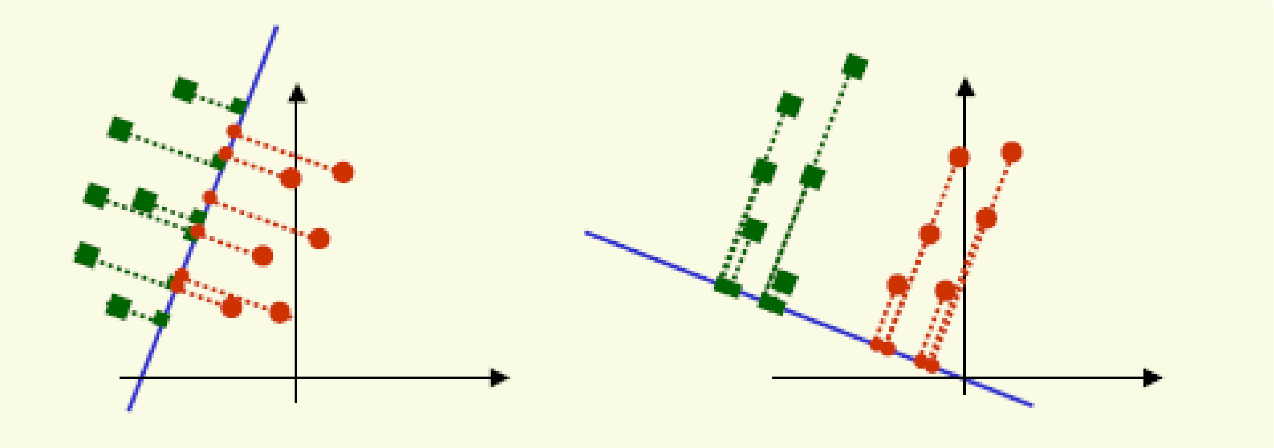
Tražimo linearnu funkciju od **x** (t.j. **w**):

$$y = \mathbf{w}^t\mathbf{x}$$

Koja najbolje razdvaja dvije klase.

Fisherova linearna diskriminativna metoda

Osnovna ideja – naći projekciju na liniju u d- dimenzionalnnom prostoru tako da se primjerci različitih klasa mogu na nkjoj lako odvojiti



loša projekcija dobra projekcija

Fisherova linearna diskriminativna metoda

Neke veličine:

$$\boldsymbol{\mu}_i = \frac{1}{N_i}\sum_{\mathbf{x}\in D_i}\mathbf{x}$$

Srednja vrijednost u d-dimenzionalnom prostoru – za klasu *i*

$$\bar{\mu}_i = \frac{1}{N_i}\sum_{y\in Y_i} y = \frac{1}{N_i}\sum_{x\in D_i}\mathbf{w}^t\mathbf{x} = \mathbf{w}^t\boldsymbol{\mu}_i$$

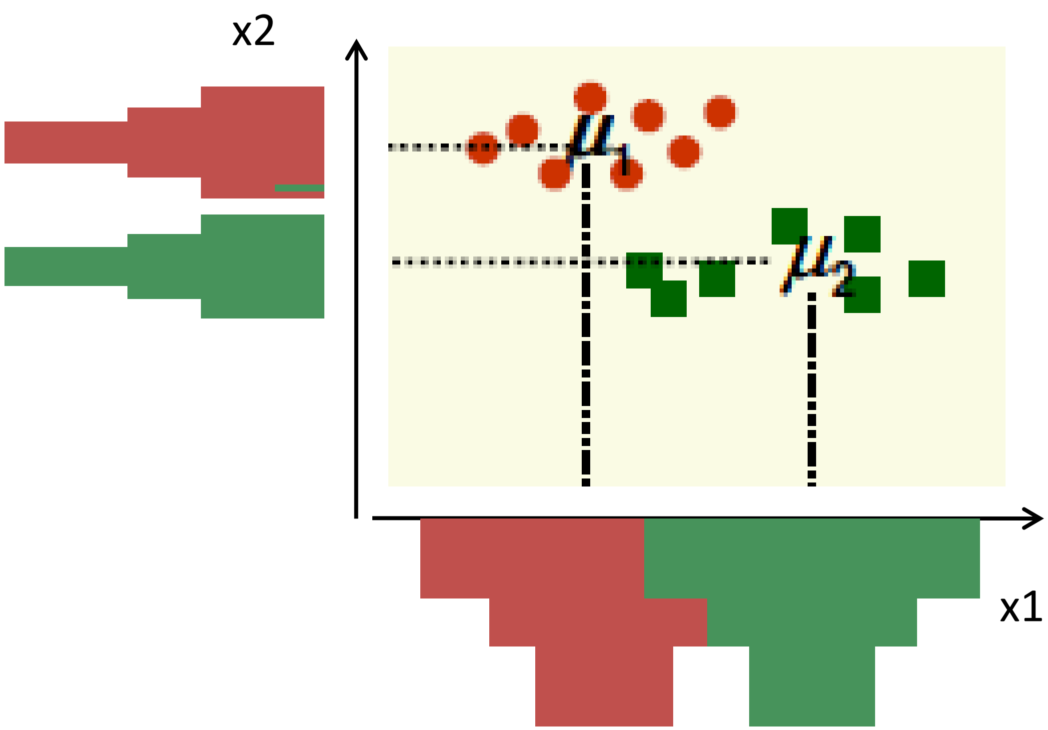
Srednja vrijednost za točke klase *i* projicirane na **w**

$$\left|\bar{\mu}_1 - \bar{\mu}_2\right| = \left|\mathbf{w}^t(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)\right|$$

Udaljenost između projiciranih srednjih vrijednosti za dvije klase

Fisherova linearna diskriminativna metoda

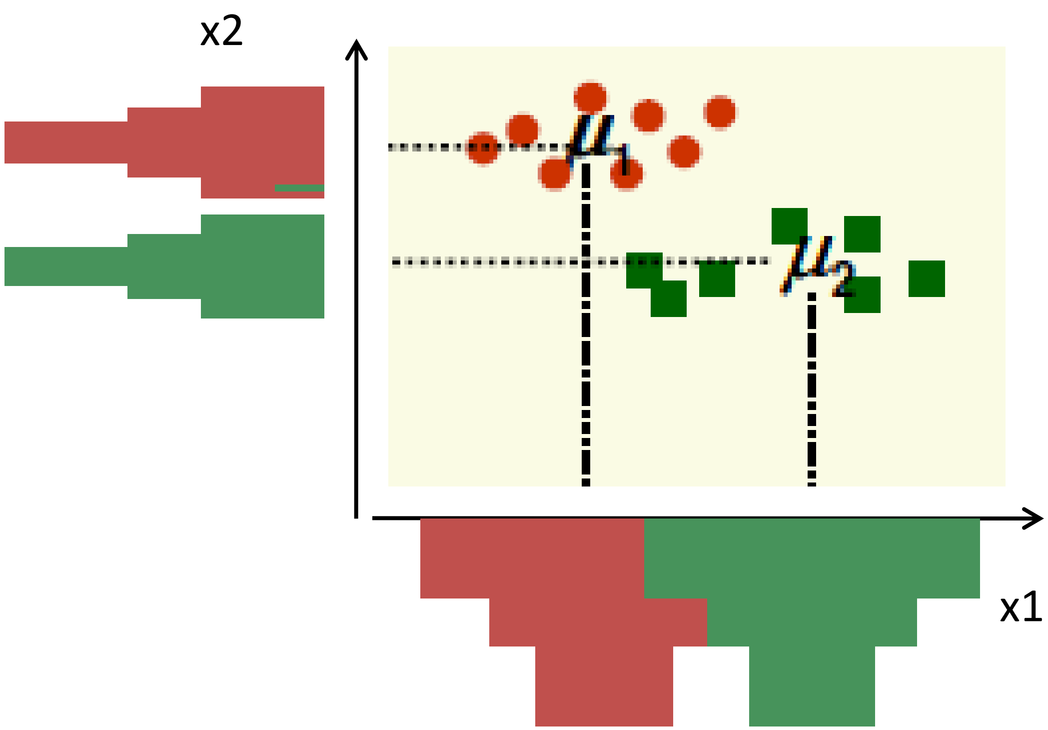
Koliko je dobra mjera separacije $\left|\bar{\mu}_1 - \bar{\mu}_2\right|$?



Koja od osi je bolja za razdvajanje klasa, x1 ili x2 ?

Fisherova linearna diskriminativna metoda

Koliko je dobra mjera separacije $|\overline{\mu}_1 - \overline{\mu}_2|$?



x2 je bolja, no:

$$|\overline{\mu}_1 - \overline{\mu}_2|_{x1} > |\overline{\mu}_1 - \overline{\mu}_2|_{x2}$$

problem je što $|\overline{\mu}_1 - \overline{\mu}_2|$

ne uzima u obzir varijancu

distribucije primjera.

Fisherova linearna diskriminantna metoda

Ako definiramo:

$$y_i = \mathbf{w}^t \mathbf{x}_i$$ projicirani primjeri

$$\tilde{s}_1^{\,2} = \sum_{y_i \in Cl_1} (|y_i - \overline{\mu}_1|)$$ "raspršenje" primjera klase1

$$s_2^2 = \sum_{y_i \in Cl_2} (|y_i - \overline{\mu}_2|)$$ "raspršenje" primjera klase2

Možemo raspršenje koristiti za normalizaciju udaljenosti projiciranih "centara" između klasa !

Fisherova linearna diskriminantivna metoda

- Moramo normalizirati koristeći raspršenje klase 1 i klase 2 !
- FLD dakle svodi se na pronalaženje projekcije na liniju **w** koja maksimizira J(**w**):

želimo da brojnik bude što veći

$$J(w) = \frac{\left|\bar{\mu}_1 - \bar{\mu}_2\right|}{\left|\tilde{s}_1^2 + \tilde{s}_2^2\right|}$$

a nazivnik što manji!

Fisherova linearna diskriminantivna metoda

$$J(w) = \frac{|\tilde{\mu}_1 - \tilde{\mu}_2|}{|\tilde{s}_1^2 + \tilde{s}_2^2|}$$

- Kako izraziti J(**w**) kao funkciju – radi se zapravo o optimizaciji J(**w**),
  - u ovisnosti o **w !**

- Treba eksplicitno prikazati J u ovisnosti o **w !**

- Definiramo matrice raspršenja za svaku klasu – prije projekcije - za originalne primjere

$$\mathbf{S}_1 = \sum_{\mathbf{x}_i \in Cl_1} (\mathbf{x}_i - \mu_1)(\mathbf{x}_i - \mu_1)^t$$

$$\mathbf{S}_2 = \sum_{\mathbf{x}_i \in Cl_2} (\mathbf{x}_i - \mu_2)(\mathbf{x}_i - \mu_2)^t$$

Fisherova linearna diskriminantivna metoda

- Definiramo matricu raspršenja *unutar* klasa za svaku klasu

$$\mathbf{S}_W = \mathbf{S}_1 + \mathbf{S}_2$$

- Uz prethodnu definiciju

$$\tilde{s}_1^2 = \sum_{y_i \in Cl_1} (y_i - \overline{\mu}_1)^2$$

- I ako koristimo:

$$y_i = \mathbf{w}^t \mathbf{x}_i \quad \mathbf{i} \qquad \overline{\mu}_1 = \mathbf{w}^t \boldsymbol{\mu}_1$$

- Dobivamo:

$$\begin{aligned}
\tilde{s}_1^2 &= \sum_{y_i \in Cl} (\mathbf{w}^t \mathbf{x}_i - \mathbf{w}^t \boldsymbol{\mu}_1)^2 \\
&= \sum_{y_i \in Cl} (\mathbf{w}^t (\mathbf{x}_i - \boldsymbol{\mu}_1))^t (\mathbf{w}^t (\mathbf{x}_i - \boldsymbol{\mu}_1)) \\
&= \sum_{y_i \in Cl} ((\mathbf{x}_i - \boldsymbol{\mu}_1)^t \mathbf{w})^t ((\mathbf{x}_i - \boldsymbol{\mu}_1)^t \mathbf{w}) \\
&= \sum_{y_i \in Cl} \mathbf{w}^t (\mathbf{x}_i - \boldsymbol{\mu}_1)(\mathbf{x}_i - \boldsymbol{\mu}_1)^t \mathbf{w} = \mathbf{w}^t \mathbf{S}_1 \mathbf{w}
\end{aligned}$$

Fisherova linearna diskriminantivna metoda

- Slično kao i za klasu 1

$$\tilde{s}_2^2 = \mathbf{w}^t \mathbf{S}_2 \mathbf{w}$$

- Dakle

$$\tilde{s}_1^2 + \tilde{s}_2^2 = \mathbf{w}^t \mathbf{S}_1 \mathbf{w} + \mathbf{w}^t \mathbf{S}_2 \mathbf{w} = \mathbf{w}^t \mathbf{S}_W \mathbf{w}$$

- Ako definiramo matricu raspršenja između klasa $\mathbf{S}_B$ kao mjeru separacije između srednjih vrijednosti između klasa prije projekcije

$$\mathbf{S}_B = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^t$$

- A razlika između projiciranih srednjih vrijednosti je:

$$\begin{aligned}(\bar{\mu}_1 - \bar{\mu}_2)^2 &= (\mathbf{w}^t \boldsymbol{\mu}_1 - \mathbf{w}^t \boldsymbol{\mu}_2)^2 \\ &= \mathbf{w}^t (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^t \mathbf{w} \\ &= \mathbf{w}^t \mathbf{S}_B \mathbf{w}\end{aligned}$$

Fisherova linearna diskriminantivna metoda

- Na kraju je naša funkcija cilja

$$J(\mathbf{w}) = \frac{(\overline{\mu}_1 - \overline{\mu}_2)^2}{\tilde{s}_1^2 + \tilde{s}_2^2} = \frac{\mathbf{w}^t \mathbf{S}_B \mathbf{w}}{\mathbf{w}^t \mathbf{S}_W \mathbf{w}}$$

- Da bi je optimirali, našli maximum – prva derivacija po **w** se mora izjednačiti s nulom

$$\frac{d}{d\mathbf{w}} J(\mathbf{w}) = 0$$

- Na koncu se to svede na problem određivanja svojstvenih vrijednosti

$$\Rightarrow \mathbf{S}_B \mathbf{w} = \lambda \mathbf{S}_W \mathbf{w}$$

$$\Rightarrow \mathbf{S}_B \mathbf{w} = \lambda \mathbf{S}_W \mathbf{w}$$

- Ako postoji inverzna matrica - nakon sređivanja:

$$\mathbf{w} = \mathbf{S}_w^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$$

- Za potrebe klasifikacije - još je potrebno odrediti graničnu vrijednost *t*:

$$y_i = \mathbf{w}^t \mathbf{x}_i < \mathrm{t} \Rightarrow \mathrm{y}_1$$
$$y_i = \mathbf{w}^t \mathbf{x}_i \geq \mathrm{t} \Rightarrow \mathrm{y}_2$$

Linearna diskriminativna analiza

- K klasa, **X** (n x p) matrica podataka

$$p(c_k \mid \mathbf{x}) \propto p(\mathbf{x} \mid c_k, \theta_k)$$

- Svaka klasa se modelira kao multi-varijantna normalna distribucija

$$p(\mathbf{x} \mid c_k) = \frac{1}{(2\pi)^{p/2} \mid \mathbf{\Sigma}_k \mid^{1/2}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_k)^T \mathbf{\Sigma}_k^{-1}(\mathbf{x}-\boldsymbol{\mu}_k)}$$

- LDA pretpostavlja da je $\Sigma_k = \Sigma$ za sve *k*, pa vrijedi:

$$\log \frac{p(c_k \mid \mathbf{x})}{p(c_l \mid \mathbf{x})} = \log \frac{p(c_k)}{p(c_l)} - \frac{1}{2}(\boldsymbol{\mu}_k - \boldsymbol{\mu}_l)^T \mathbf{\Sigma}^{-1}(\boldsymbol{\mu}_k - \boldsymbol{\mu}_l) + \mathbf{x}^T \mathbf{\Sigma}^{-1}(\boldsymbol{\mu}_k - \boldsymbol{\mu}_l)$$

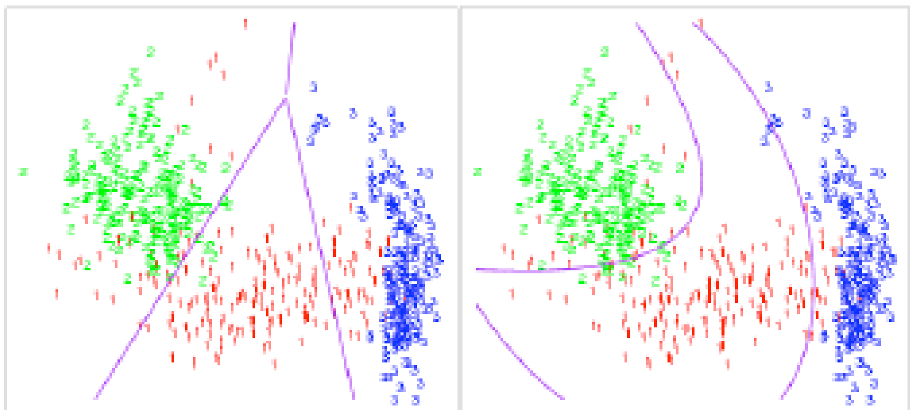
- U klasifikaciji – klasa se određuje prema – linearnoj diskriminantnoj funkciji :

$$\delta_k(\mathbf{x}) = \mathbf{x}^T \mathbf{\Sigma}^{-1} \boldsymbol{\mu}_k - \frac{1}{2} \boldsymbol{\mu}_k^T \mathbf{\Sigma}^{-1} \boldsymbol{\mu}_k + \log p(c_k)$$

i

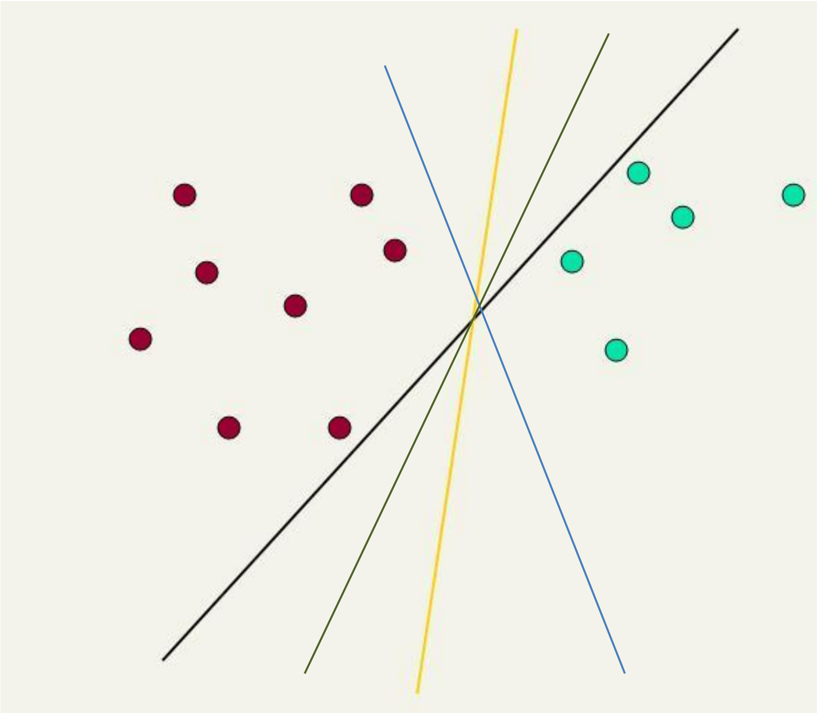
$$klasa(\mathbf{x}) = \arg\max \delta_k(\mathbf{x})$$

Linearna diskriminativna analiza



LDA QDA

Koja hiper-ravnina je najbolja ?



SVM - Support Vector Machines
– Metoda jezgrenih funkcija

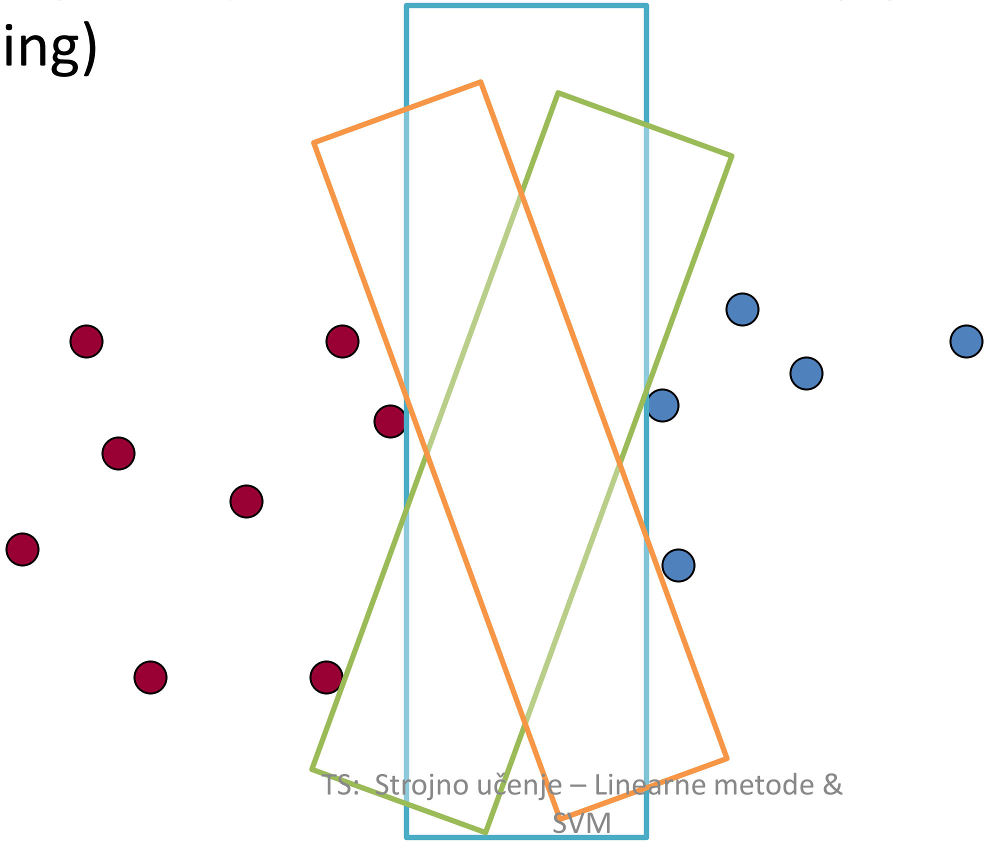
SVM – nalazi optimalnu ravninu razdvajanja

- Maksimizirajući udaljenost između hiperravnina i "kritičnih" točaka blizu ravnine razdvajanja (en. decision boundary)

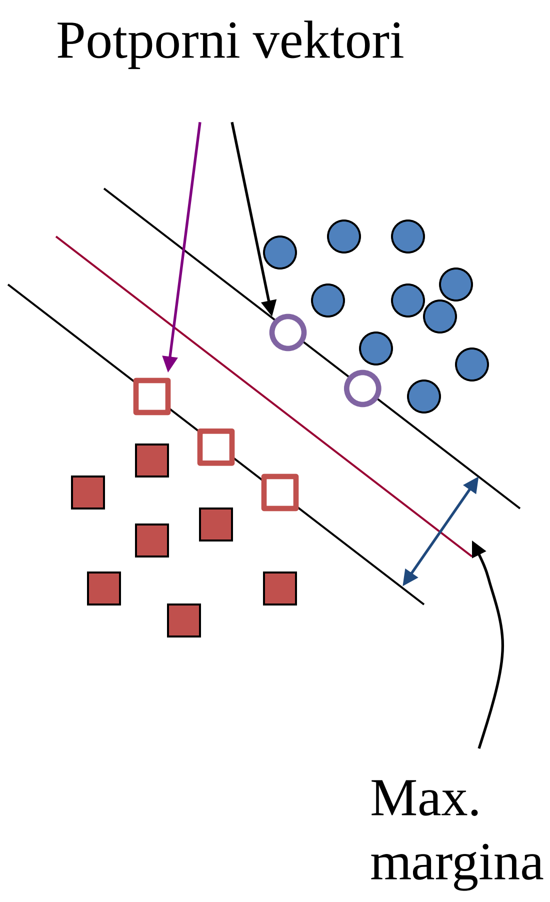
Intuitivno:
Ako ne postoje točke blizu površine razdvajanja, to znači da nemamo nesigurnih klasifikacijskih odluka !?

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)
- Ustvari je funkcija odluke definirana preko podskupa primjera iz skupa za učenje *tzv. Potpornih vektora (en. support vectors*)
- Problem određivanja SV: *kvadratni optimizacijski problem (en. quadratic programming)*



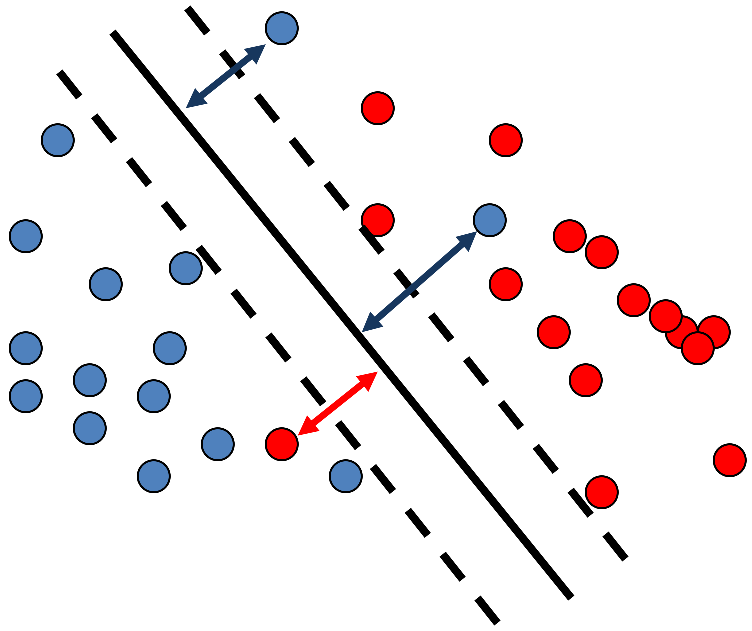
Separabilni i neseparabilni problemi

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



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 ➔ $\mathrm{sign}(\mathbf{w}^T\mathbf{x}_i + b)$
- margina $\mathbf{x}_i$ je određena sa ➔ $y_i\,(\mathbf{w}^T\mathbf{x}_i + b)$
- Funkcionalna margina cijelog skupa točaka je

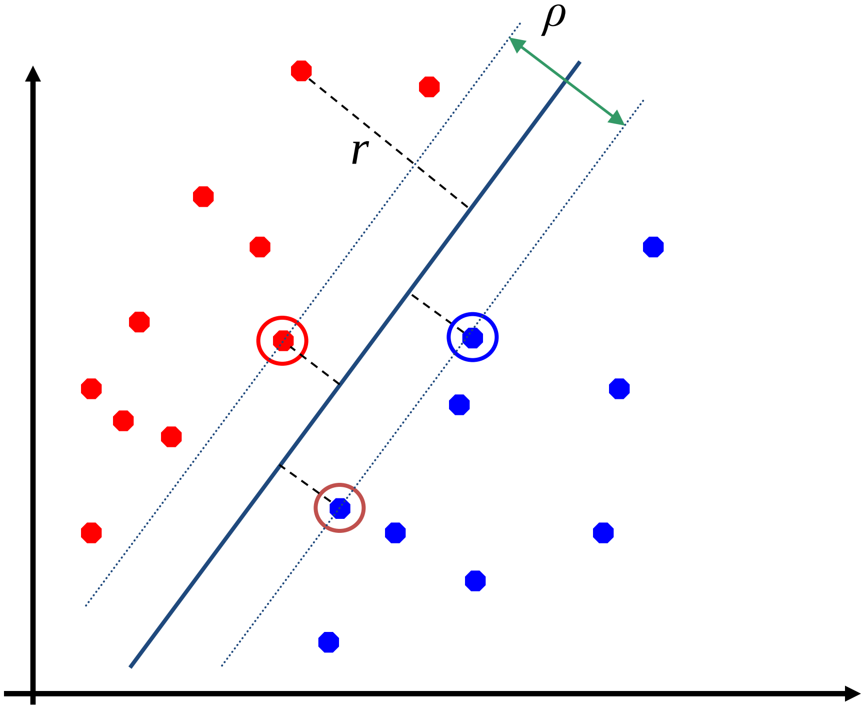
  minimum $y_i\,(\boldsymbol{w}^T\boldsymbol{x}_i + b)$

Pojam - geometrijske margine

- Udaljenost između plohe i točke x

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

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



Linearni SVM - izvod

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

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

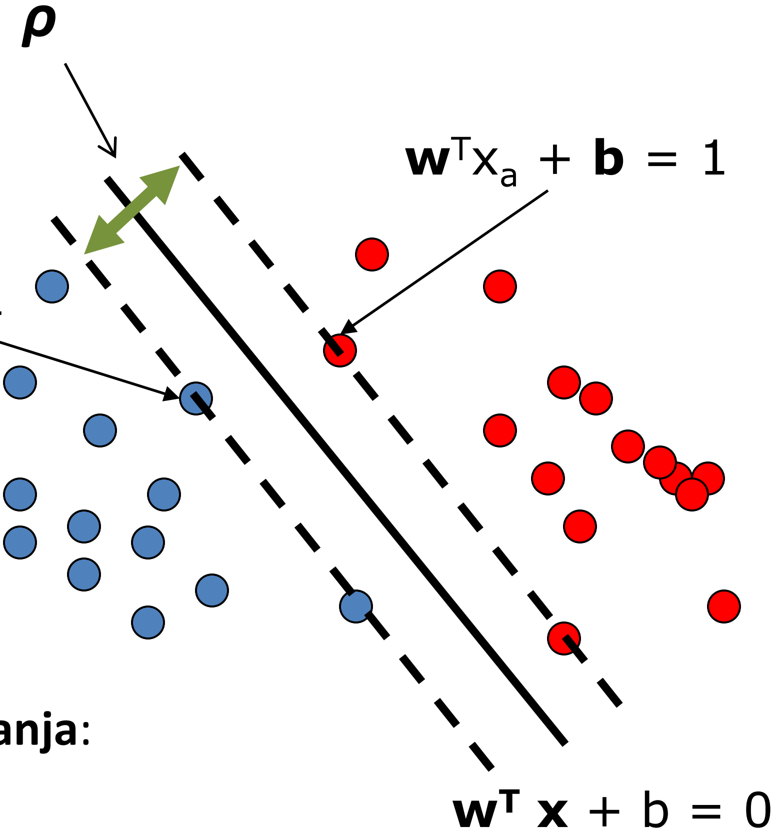
$$\mathbf{w^T x_i} + b \leq -1 \quad \text{ako je } y_i = -1$$

- Za potporne vektore (koji 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: $\rho = \dfrac{2}{\|\mathbf{w}\|}$

Linearni SVM - izvod



Ploha razdvajanja

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

Uz ograničenja:

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

Dolazimo do vrijednosti za marginu razdvajanja:

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

$\rho = ||\mathbf{x}_a - \mathbf{x}_b||_2 = 2/||\mathbf{w}||_2$

Linearni SVM - izvod

- Ovaj se problem može formulirati kao *kvadratni optimizacijski problem:*

Pronađi **w** i *b* tako da je

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

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 - izvod

- Bolja formulacija – kao minimizacijski problem:

  min $||\mathbf{w}||$ = max $(2/\ ||\mathbf{w}||)$

> Naći **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 :
>
> $y_i\ (\mathbf{w^T x_i} + b) \geq 1$

Rješavanje optimizacijskog problema

Nađi **w** i b - takve da je:

$\mathbf{\Phi(w)} = \frac{1}{2} \mathbf{w}^T\mathbf{w}$ - 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
- Dobro poznati i rješivi problem – puno (više ili manje složenih) algoritama za nihovo rješavanje (MATLAB, MATHEMATICA)
- Ustvari rješavanje se tipično radi nakon prevođenja u tzv. *dualni problem* u kjojem se tzv. *Lagrange-ovim multiplikatorima* α_i penalizira svako ograničenje iz primarnog problema

Rješavanje optimizacijskog problema

Dualna formulacija:

Naći $\alpha_1 \ldots \alpha_N$ tako da je:

$\mathbf{Q}(\boldsymbol{\alpha}) = \sum \alpha_i - \frac{1}{2} \sum \sum \alpha_i \alpha_j y_i y_j \mathbf{x_i}^T \mathbf{x_j}$ **-** maksimalno

Te da vrijedi

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

(2) $\alpha_i \geq 0 \ \forall \alpha_i$

Rješenje optimizacijskog problema

$\mathbf{w} = \Sigma\alpha_i y_i \mathbf{x_i}$ $\quad b = y_k - \mathbf{w}^T\mathbf{x_k}$ za bilo koji $\mathbf{x_k}$ za koji je $\alpha_k \neq 0$

- Svaki α_i koji je različit od 0 – zapravo znači da je taj $\mathbf{x_i}$ – potporni vektor.
- Konačno – klasifikacijska funkcija ima ovaj oblik:

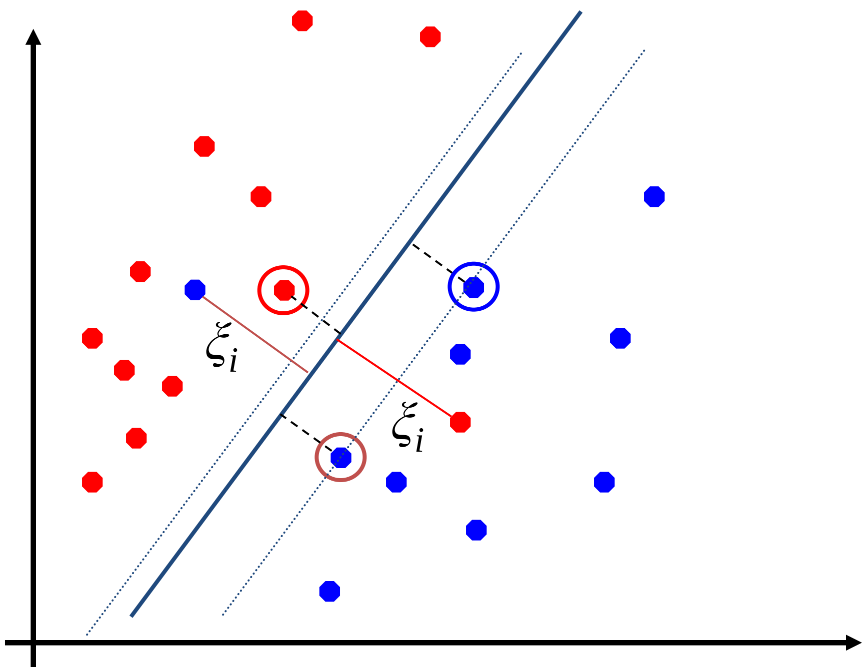
$$f(\mathbf{x}) = \Sigma\alpha_i y_i \mathbf{x_i}^T\mathbf{x} + b$$

- Rješenje je proporcionalno sumi umnožaka - *unutarnjih produkta* $\mathbf{x_i}^T\mathbf{x}$ između nove točke (primjera) $\mathbf{x}$ i svih potpornih vektora $\mathbf{x_i}$!
- Treba zapamtiti i da je u rješavanju optimizacijskog problema također korišten produkt $\mathbf{x_i}^T\mathbf{x_j}$ između svih parova točaka skupa za učenje.

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:

$\mathbf{\Phi}(\mathbf{w}) = ½\, \mathbf{w}^T\mathbf{w} + C\Sigma\xi_i$ - 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 $\alpha_1 \ldots \alpha_N$ tako da je

$$\mathbf{Q}(\boldsymbol{\alpha}) = \Sigma\alpha_i - ½\Sigma\Sigma\alpha_i\alpha_j y_i y_j \mathbf{x_i}^T\mathbf{x_j} \quad \text{- maksimalno}$$

Te da vrijedi

(1) $\Sigma\alpha_i y_i = 0$

(2) $0 \leq \alpha_i \leq C \; \forall \; \alpha_i$

- Ni ξ_i ali ni Lagrange-ovi multiplikatori se ne nalaze u dualnoj formulaciji!
- No, i dalje vrijedi: $\mathbf{x_i}$ sa $\alpha_i > 0$ *– su potporni vektori.*
- Rješenje dualnog problema (soft margin):

$$\mathbf{w} = \Sigma\alpha_i y_i \mathbf{x_i}$$

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

w nije potreban kod procesa klasifikacije – kao i prije !

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

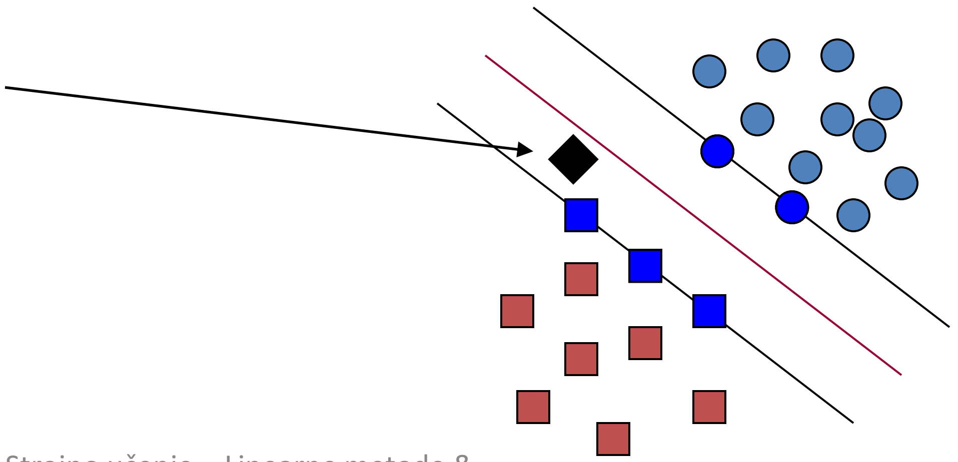
Klasifikacija

- 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 ($wx + b = \Sigma\alpha_i y_i \mathbf{x_i}^T\mathbf{x} + b$)
  - Ako stavimo neku graničnu mjeru t >0 (pouzdanost)

p > t: => 1

p < -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

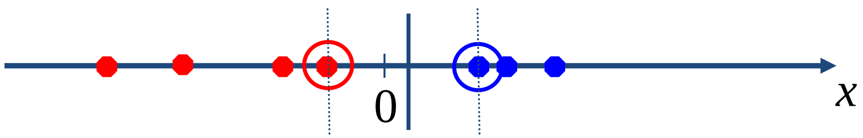
- Potporni vektori se pronalaze – optimizacijskim algoritmima
- Rješavani problem je kvadratni optimizacijski problem s linearnim ograničenjima .
- U dualnoj formulaciji problema i rješenja – potporni vektori odnosno točke iz skupa za učenje se pojavljuju u skalarnim produktima:

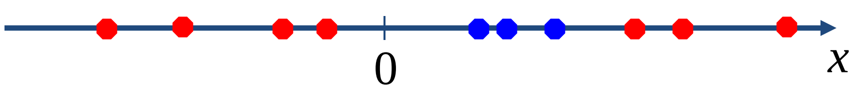
Nađi $\alpha_1 \ldots \alpha_N$ tako da je
$\mathbf{Q}(\boldsymbol{\alpha}) = \Sigma\alpha_i - ½\Sigma\Sigma\alpha_i\alpha_j y_i y_j \boxed{\mathbf{x_i}^T\mathbf{x_j}}$ - maksimalno i vrijedi:
(1) $\Sigma\alpha_i y_i = 0$
(2) $0 \leq \alpha_i \leq C \; \forall \; \alpha_i$

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

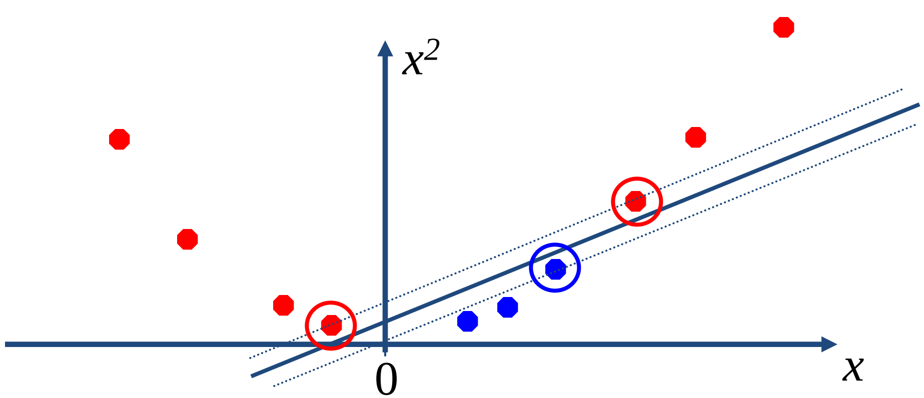
Nelinearni SVM

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



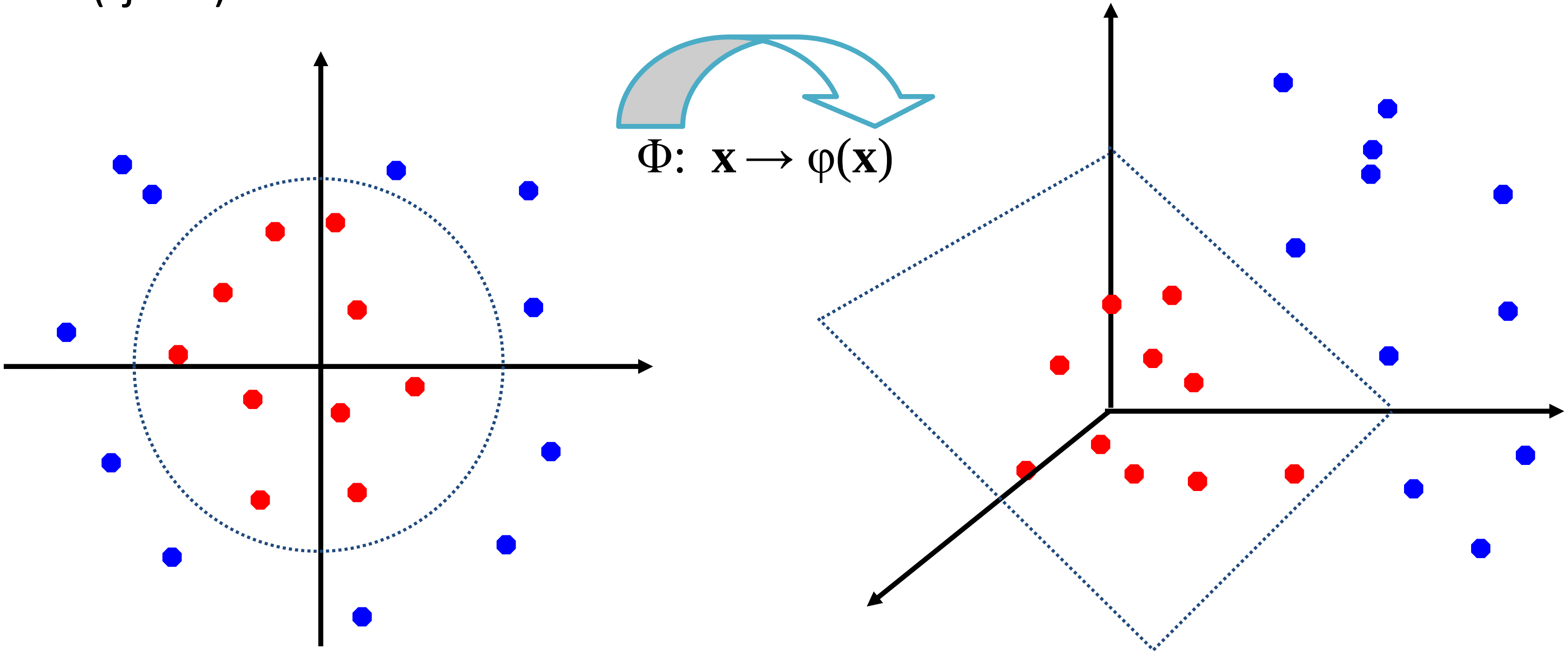
- No, što ako to ne vrijedi ?
- 

- … 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)



“Kernel trik”

- Linearni SVM - skalarni produkt između vektora $K(\mathbf{x}_i,\mathbf{x}_j)=\mathbf{x}_i^T\mathbf{x}_j$
- Transformacija kojom svaku točku mapiramo u neki novi prostor

$$\Phi: \mathbf{x} \rightarrow \phi(\mathbf{x})$$

- Sada bi skalarni produkt trebao izgledati drukčije:

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

- *Kernel funkcija* - funkcija koja korespondira nutarnjem produktu u nekom “ekspandiranom” prostoru novih varijabli

“Kernel trik”

- Primjer: originalni prostor

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

Neka je $K(\mathbf{x}_\mathbf{i},\mathbf{x}_\mathbf{j})=(1 + \mathbf{x}_\mathbf{i}^\mathbf{T}\mathbf{x}_\mathbf{j})^2$,

Pokažimo da vrijedi $K(\mathbf{x}_\mathbf{i},\mathbf{x}_\mathbf{j})= \phi(\mathbf{x}_\mathbf{i})^\mathbf{T}\phi(\mathbf{x}_\mathbf{j})$:

$$K(\mathbf{x}_\mathbf{i},\mathbf{x}_\mathbf{j})=(1 + \mathbf{x}_\mathbf{i}^\mathbf{T}\mathbf{x}_\mathbf{j})^2 = 1+ x_{i1}^2x_{j1}^2 + 2\, x_{i1}x_{j1}\, x_{i2}x_{j2}+ x_{i2}^2x_{j2}^2 + 2x_{i1}x_{j1} + 2x_{i2}x_{j2}=$$

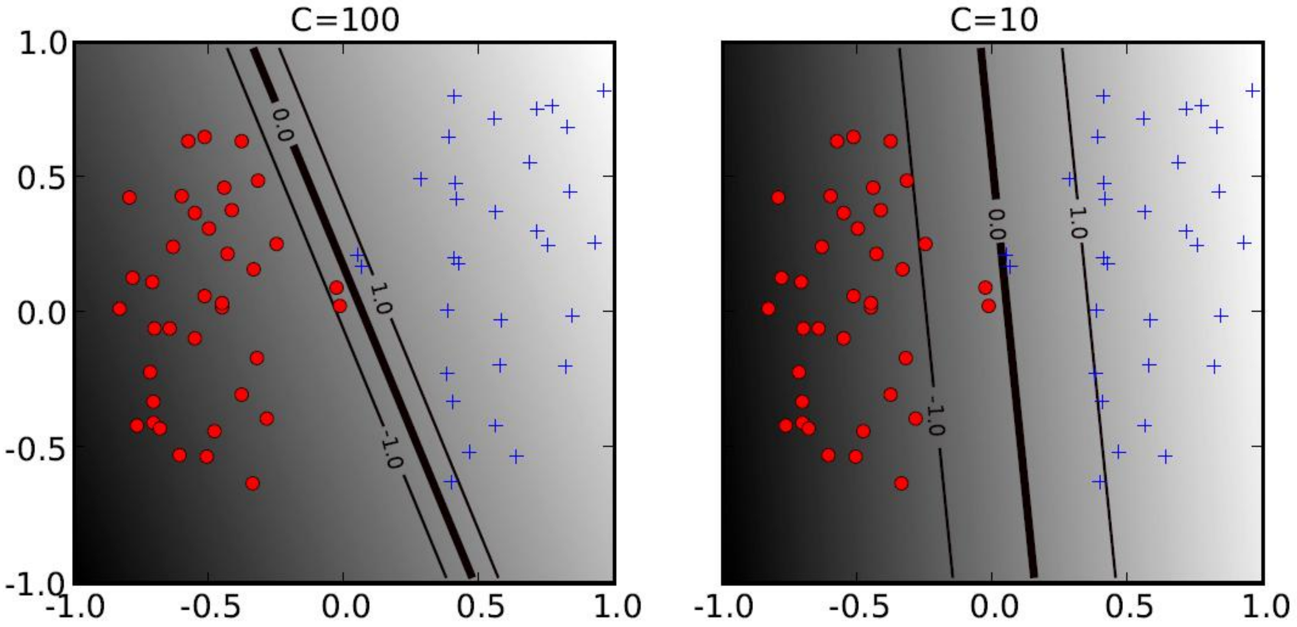
$$= [1 \;\; x_{i1}^2 \;\; \sqrt{2}\, x_{i1}x_{i2} \;\; x_{i2}^2 \;\; \sqrt{2}x_{i1} \;\; \sqrt{2}x_{i2}]^\mathbf{T}\, [1 \;\; x_{j1}^2 \;\; \sqrt{2}\, x_{j1}x_{j2} \;\; x_{j2}^2 \;\; \sqrt{2}x_{j1} \;\; \sqrt{2}x_{j2}]$$

$$= \phi(\mathbf{x}_\mathbf{i})^\mathbf{T}\phi(\mathbf{x}_\mathbf{j}) \qquad \text{gdje je } \phi(\mathbf{x}) = [1 \;\; x_1^2 \;\; \sqrt{2}\, x_1x_2 \;\; x_2^2 \;\; \sqrt{2}x_1 \;\; \sqrt{2}x_2]$$

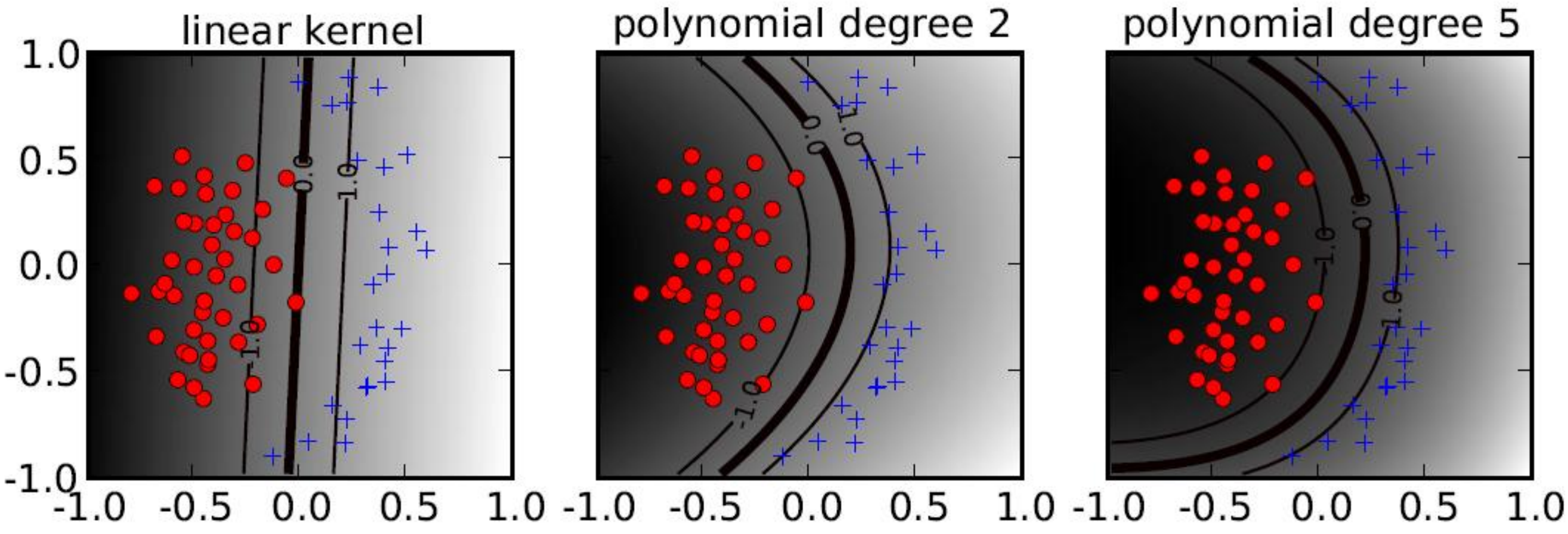
Kerneli

- Zašto?
  - Omogućavaju pretvaranje neseparabilnih problema u separabilne.
  - Mapiranje u bolje (reprezentacijski) prostore
- Uobičajeni kerneli
  - Linearni
  - Polinomijalni K(**x,z**) = (1+$\mathbf{x}^T\mathbf{z})^d$
  - RBF $K(\mathbf{x}_i,\mathbf{x}_j) = e^{\frac{-\|\mathbf{x}_i - \mathbf{x}_j\|}{2\sigma^2}}$

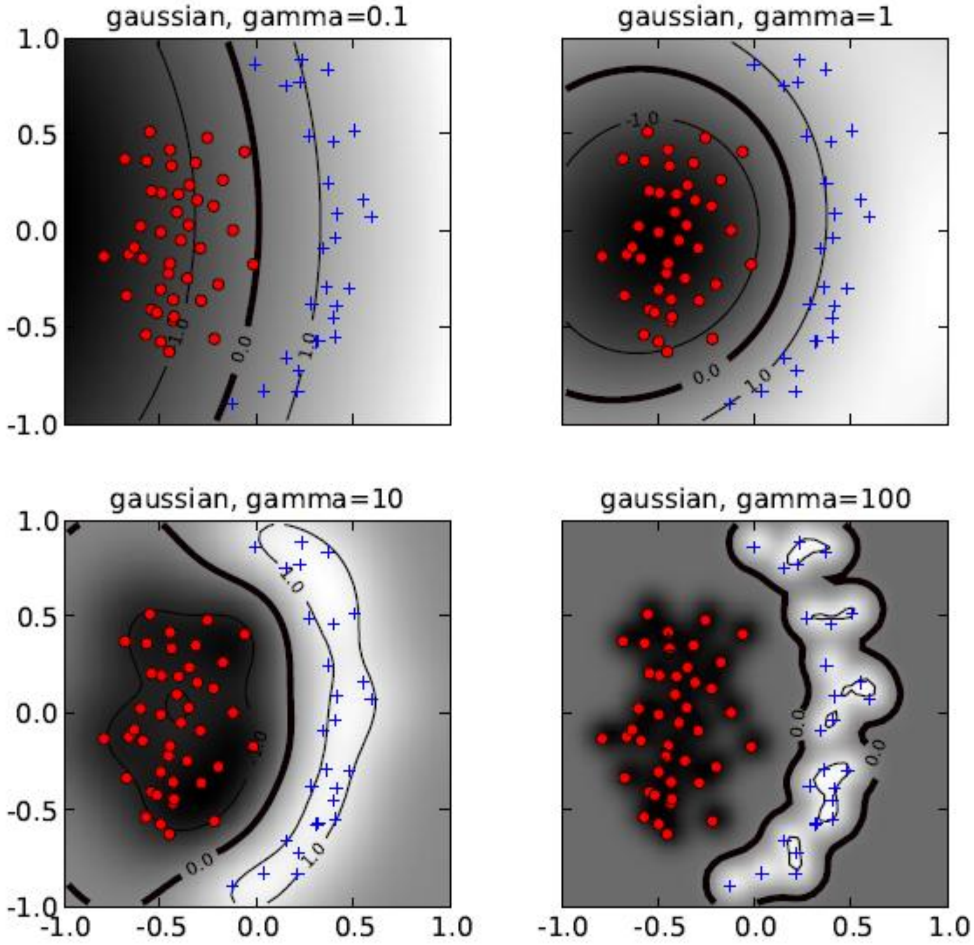
Linearni SVM – efekt različitih vrijednosti parametra C



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



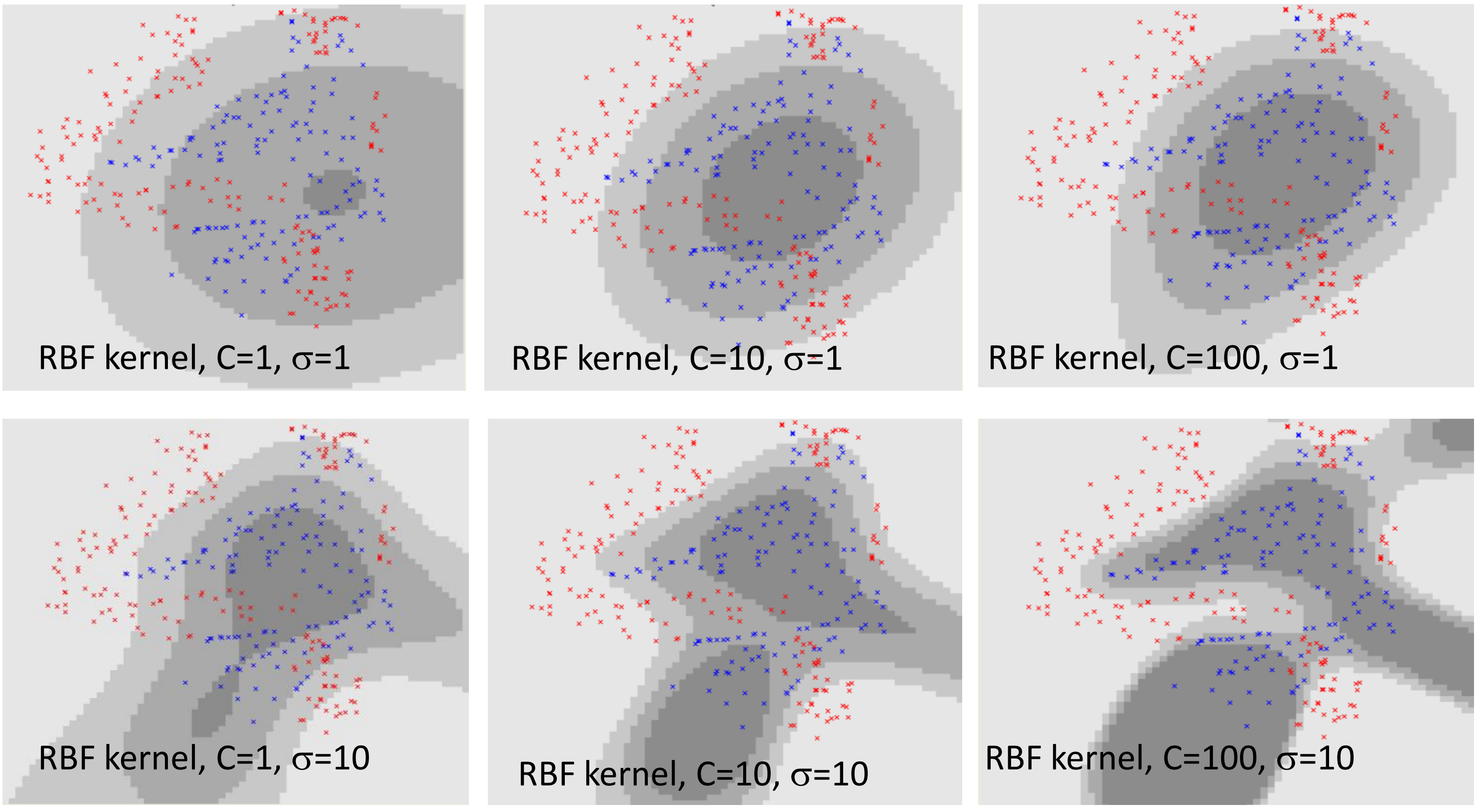
Odabir parametara kernela – efekti



Gaussian kernel, C=const.

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

Odabir parametara parametara kernela i faktora C



Karakteristike SVM-a

- Po mnogima – najbolja metoda ukupno gledajući
- Popularna i vrlo dobri rezultati – text mining
- U praksi – mnoge druge metode rade otprilike isto dobro
  - Postoji mnogo komparacija koje to pokazuju

Važno je: Razumijevanje i iskustvo za efektivno korištenju

- Odabir kernela, odabir parametara (C, parametri kernela)
- Dobro je kombinirati selekciju varijabli i SVM u prakti
- RapidMiner (više SVM varijanti + vizualizacija)

Sažetak

- Definiranje plohe razdvajanja preko potpornih vektora
  - Potporni vektor = “kritične” točke (primjeri) najbliži plohi razdvajanja
- Pojam kernela:
  - Moćan alat za mapiranje u visoko-dimenzionalne prostore, kao i (re)definiranje metrika sličnosti
- Samostalni paketi i software:
  - SVM Torch; SVMLight; LibSVM
  - RapidMiner, WEKA, MATLAB, R ....
  - www.kernel-machines.org