Neuronske mreže: Stroj s potpornim vektorima (SVM)

Prof. dr. sc. Sven Lončarić
Doc. dr. sc. Marko Subašić

Fakultet elektrotehnike i računarstva
Sveučilište u Zagrebu

http://www.fer.hr/predmet/neumre_b

Pregled predavanja

- Problem kalsifikacije linearno separabilnih klasa
- Margina razdvajanja
- Vektori potpore
- Klasifikacija linearno neseparabilnih klasa
- Nelinearno preslikavanje u prostor značajki

Uvod

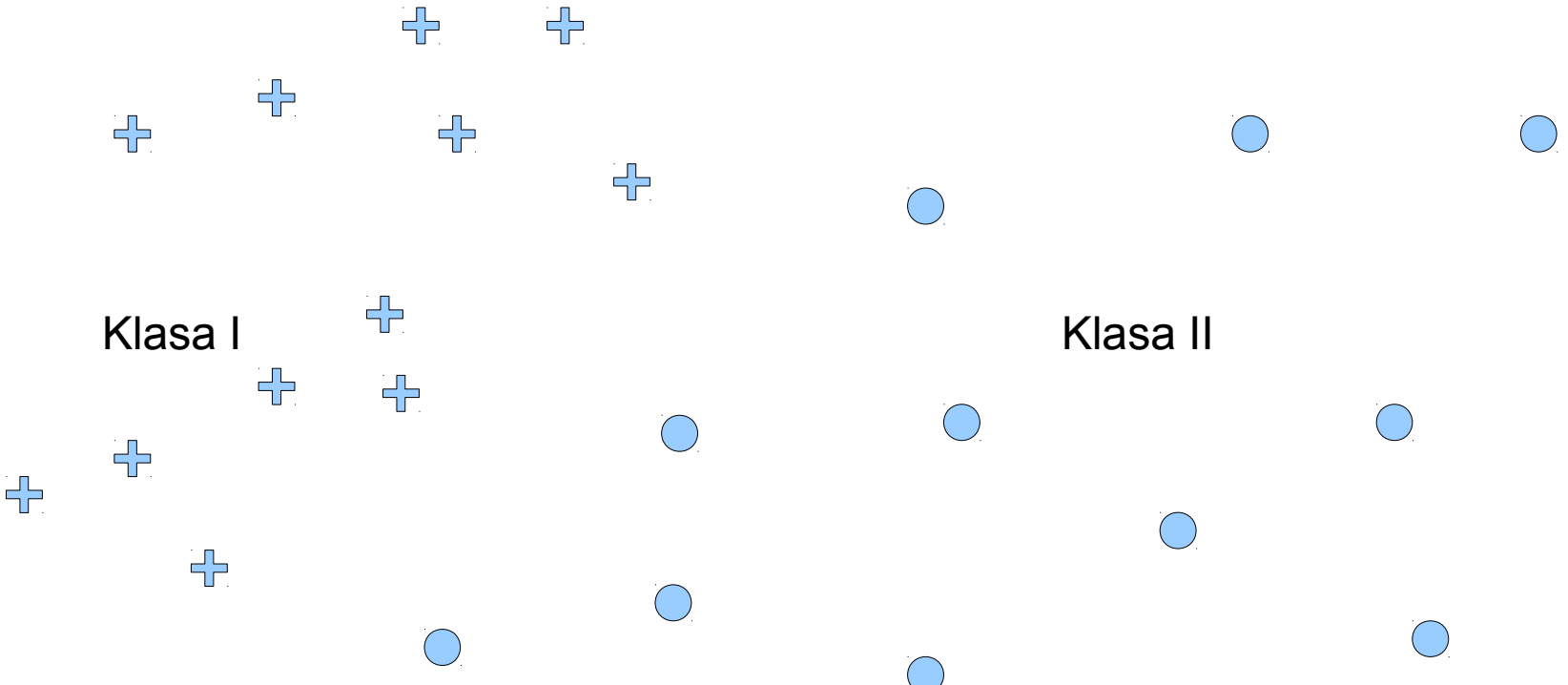
- Sličnost sa višeslojnim perceptronom i radijalnom mrežom
  - Feed forward mreža
  - Primjena u klasifikaciji i nelinearnoj regresiji
  - Inherentno dobra generalizacijska svojstva
- Razlike
  - Treniranje SVM-a se ne provodi iterativno s pojedinim uzorcima za treniranje
  - SVM minimizira broj uzoraka za treniranje unutar margine – MLP minimizaira prosječnu kvadratnu pogrešku
- Generalni algoritam za treniranje feedforward neuronskih mreža
- Feedforward mreža s jednim nelinearnim skrivenim slojem

Cilj

- Klasifikacija uzoraka dviju klasa
- Pronalazak ravnine razdvajanja dviju klasa koja maksimizira margine razdvajanja
- Ravninu razdvajanja definirati pomoću "ključnih" uzoraka za treniranje – potpornih vektora

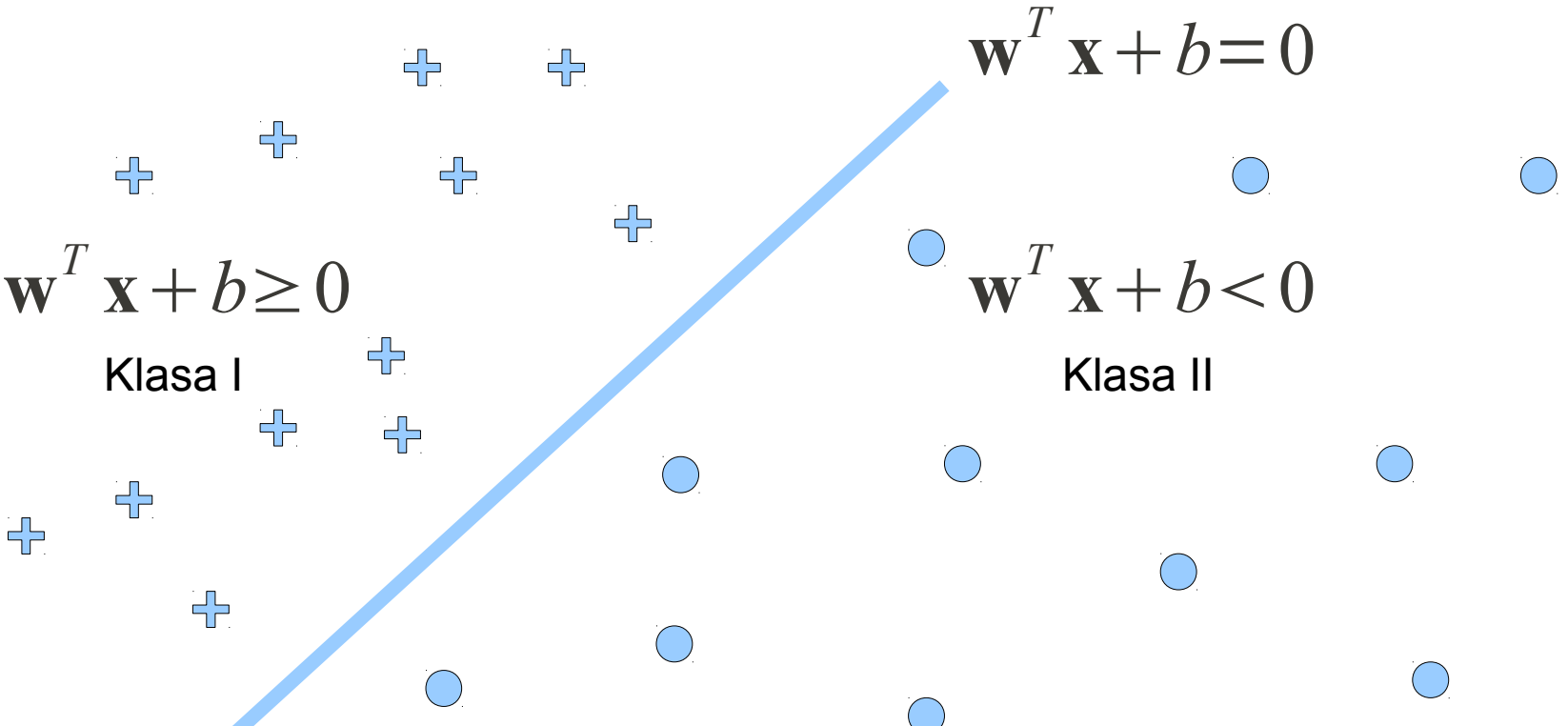
Razdvajanje linearno separabilnih klasa

- Jednostavan problem
- Kompleksnije probleme možda možemo svesti na ovaj jednostavniji...



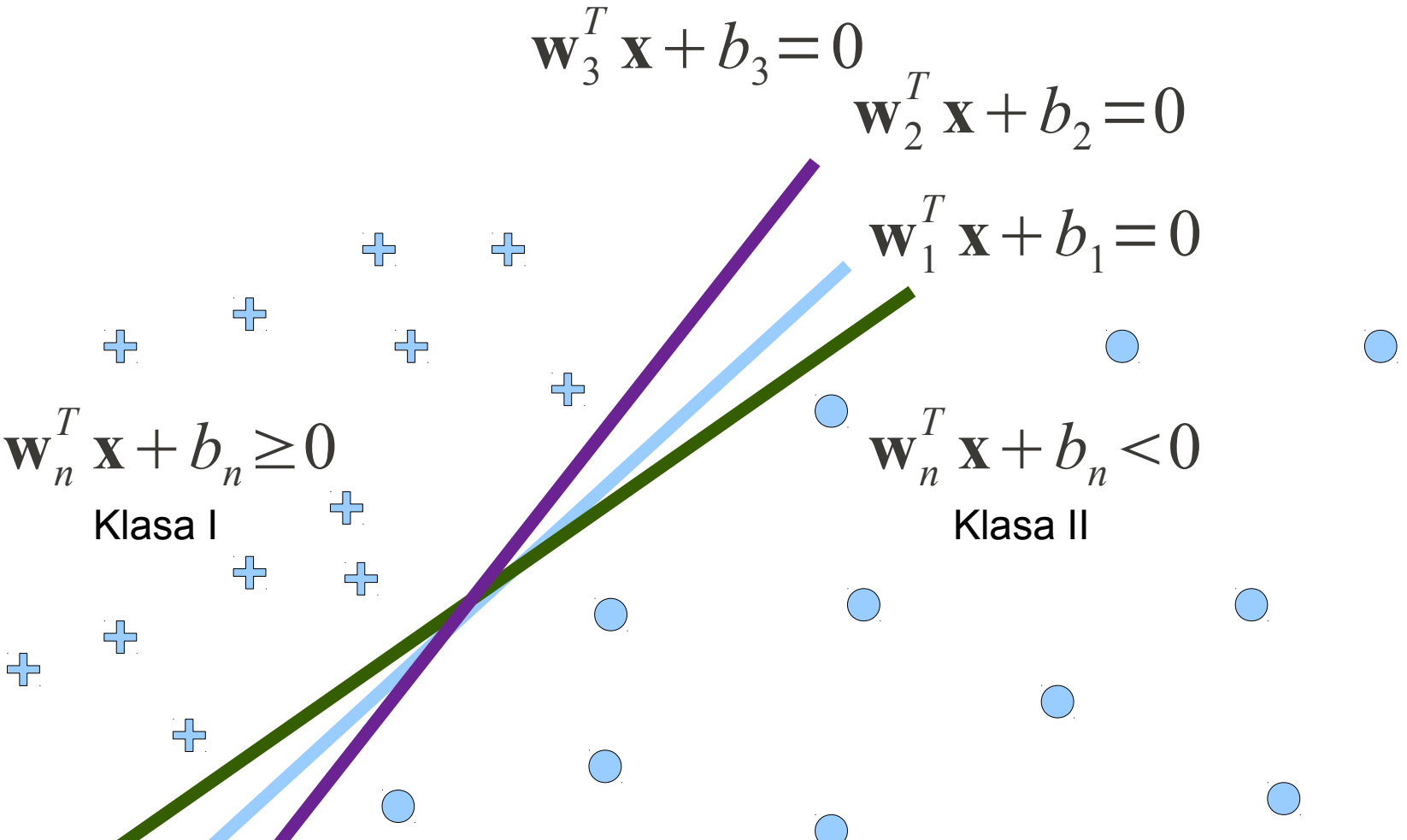
Razdvajanje linearno separabilnih klasa

- Jednadžba hiperravnine
  - **w** - vektor težina, **x** - ulazni vektor, *b* - pomak



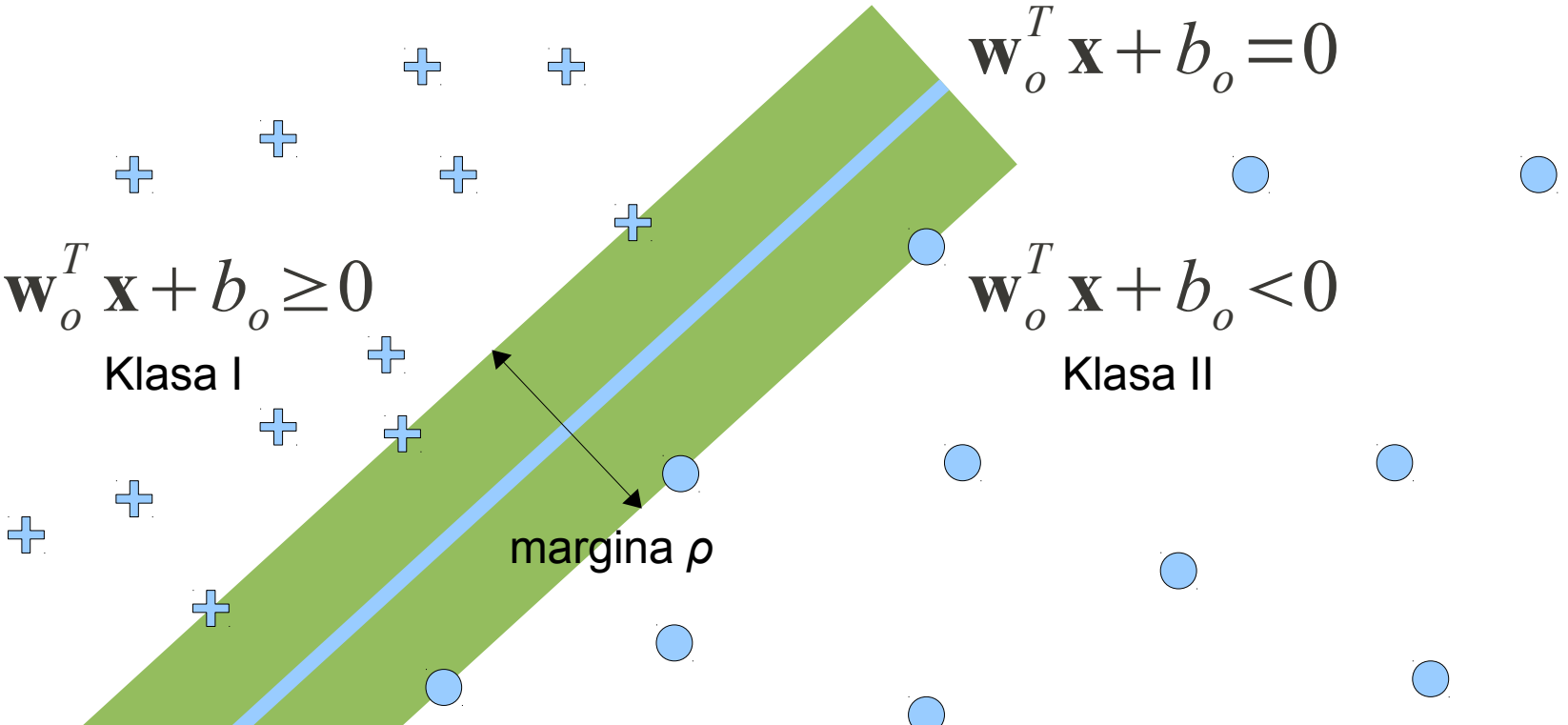
Razdvajanje linearno separabilnih klasa

- Postoji više mogućih ravnina razdvajanja
- Koja je najbolja?



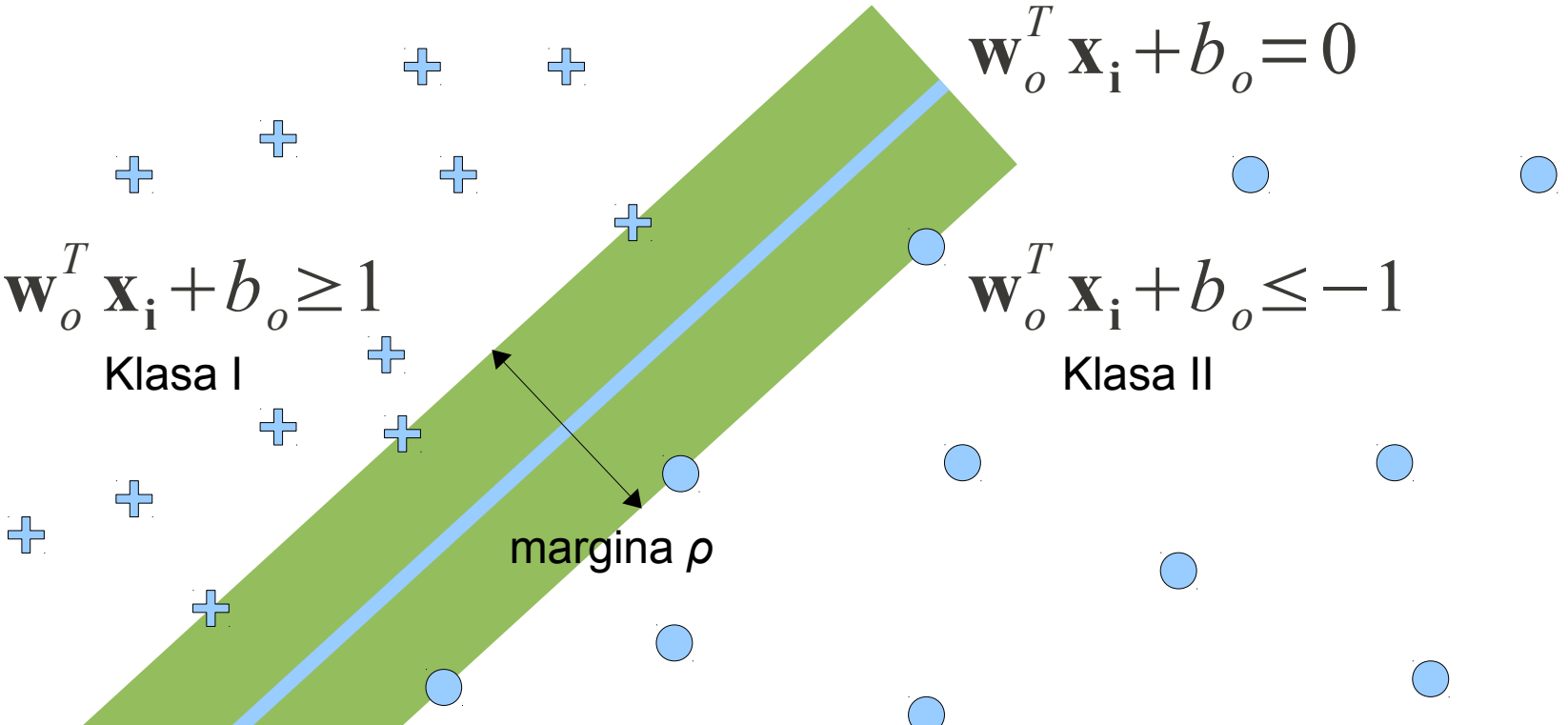
Margina razdvajanja

- Udaljenost od hiperravnine do najbližeg uzorka $\mathbf{x}_i$ bilo koje klase
- SVM-a traži optimalnu ravninu razdvajanja ($\mathbf{w}_o$ i b_o) koja maksimizira marginu razdvajanja ρ



Vektori potpore

- Vektori potpore $\mathbf{x}_i$ takvi da je $\mathbf{w}_o^T \mathbf{x}_i + b_o = \pm 1$
  - Najbliži ravnini razdvajanja
  - Najteže ih je klasificirati
  - Najbitniji za određivanje $\mathbf{w}_o$ i b_o



Ravnina razdvajanja

- Ravnina razdvajanja

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

- Diskriminacijska funkcija

$$g(\mathbf{x}) = \mathbf{w}_o^T \mathbf{x} + b_o$$

- Ovisno o predznaku određuje pripadnost pojedinoj klasi ovisno o predznaku

Udaljenost od ravnine razdvajanja

- Pozicija uzorka **x** izražena preko projekcije **x**-a na ravninu razdvajanja $\mathbf{x}_p$ i udaljenosti r
- r određuje amplitudu vektora kojem smjer određuje $\mathbf{w}_o$ – okomica na ravninu projekcije (normala)

$$\mathbf{x} = \mathbf{x}_p + r\frac{\mathbf{w}_o}{\|\mathbf{w}_o\|}$$

- Predznak od r ovisi o tome s koje strane ravnine razdvajanja se uzorak nalazi

Udaljenost od ravnine razdvajanja

- Izrazimo udaljenost pomoću diskriminacijske funkcije

$$g(\mathbf{x}) = \mathbf{w}_o^T \mathbf{x} + b_o = r\|\mathbf{w}_o\|$$

$$g(\mathbf{x}_p) = 0$$

$$r = \frac{g(\mathbf{x})}{\|\mathbf{w}_o\|}$$

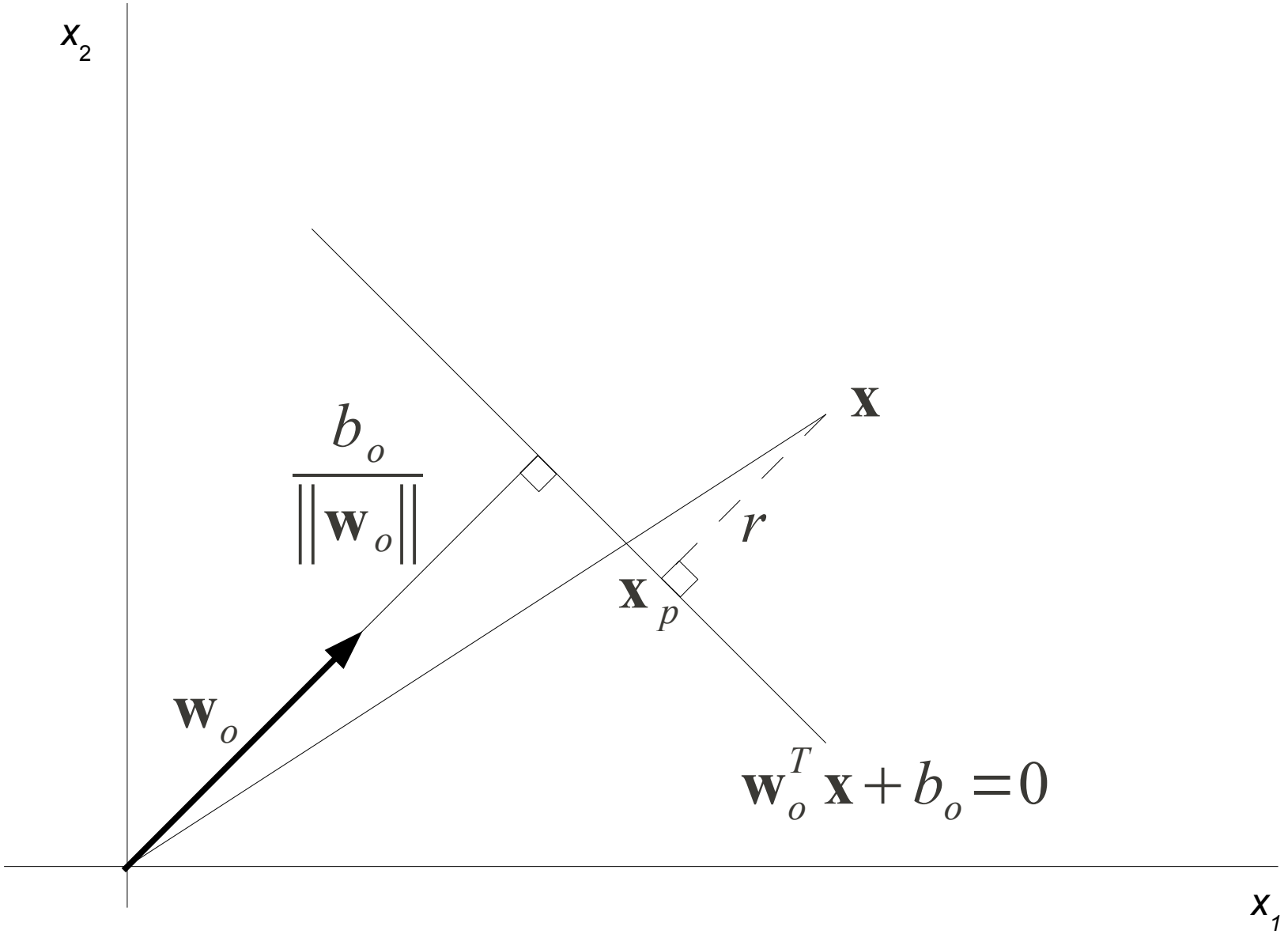
Udaljenost ravnine razdvajanja od ishodišta

- Udaljenost ravnine razdvajanja od ishodišta

$$\frac{b_o}{\|\mathbf{w}_o\|}$$

- Skaliranjem $\mathbf{w}_o$ i b_o ne mijenjamo ravninu razdvajanja
- Smjer vektora $\mathbf{w}_o$ ostaje nepromijenjen

Položaj ravnine razdvajanja



"Odabir" potpornih vektora

- Odaberimo takve potporne vektore $\mathbf{x}^{(s)}$ da vrijedi

$$g(\mathbf{x}^{(s)}) = \mathbf{w}_o^T \mathbf{x}^{(s)} + b_o = \pm 1 \qquad \text{za} \qquad d = \pm 1$$

- Tada je njihova udaljenost od ravnine razdvajanja

$$r = \frac{\mathbf{g}(\mathbf{x}^{(s)})}{\|\mathbf{w}_o\|} = \frac{\pm 1}{\|\mathbf{w}_o\|}$$

- Širina margine ρ tada postaje

$$\rho = 2r = \frac{2}{\|\mathbf{w}_o\|}$$

Maksimiziranje margine razdvajanja

- Maksimiziranje širine margine je ekvivalentno minimiziranju euklidske norme $\mathbf{w}_o$

$$\rho = 2r = \frac{2}{\|\mathbf{w}_o\|}$$

- Rezultat je optimalna ravnina razdvajanja koja ima maksimalnu marginu razdvajanja

Postupak optimizacije

- Uvjet za sve uzorke za treniranje

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

- Cilj je pronalazak minimuma funkcije (norme vektora težina)

$$\Phi(\mathbf{w}) = \frac{1}{2}\mathbf{w}^T\mathbf{w}$$

Postupak optimizacije

- Rješenje pomoću metode Lagrangeovih multiplikatora (α_i)

$$J(\mathbf{w}, b, \alpha) = \frac{1}{2}\mathbf{w}^T\mathbf{w} - \sum_{i=1}^{N} \alpha_i [d_i(\mathbf{w}^T\mathbf{x}_i + b) - 1]$$

$$\alpha_i \geq 0$$

$$\min_{\mathbf{w}, b} \max_{\alpha_i} J(\mathbf{w}, b, \alpha)$$

- Rješenje je u sedlu

Postupak optimizacije

- Parcijalne derivacije po **w** i *b* izjednačiti s nulom

$$J(\mathbf{w}, b, \alpha) = \frac{1}{2}\mathbf{w}^T\mathbf{w} - \sum_{i=1}^{N} \alpha_i [d_i(\mathbf{w}^T\mathbf{x}_i + b) - 1]$$

$$\frac{\partial J(\mathbf{w}, b, \alpha)}{\partial \mathbf{w}} = \mathbf{0} \qquad \frac{\partial J(\mathbf{w}, b, \alpha)}{b} = 0$$

$$\mathbf{w} = \sum_{i=1}^{N} \alpha_i d_i \mathbf{x}_i \qquad \sum_{i=1}^{N} \alpha_i d_i = 0$$

Postupak optimizacije

- Određivanje Lagrangeovih multiplikatora (α_i)

$$\min_{\mathbf{w},b} \max_{\alpha_i} J(\mathbf{w},b,\alpha) = \frac{1}{2}\mathbf{w}^T\mathbf{w} - \sum_{i=1}^{N} \alpha_i [d_i(\mathbf{w}^T\mathbf{x}_i + b) - 1]$$

$$\max_{\alpha_i} J(\mathbf{w},b,\alpha) = -\sum_{i=1}^{N} \alpha_i [d_i(\mathbf{w}^T\mathbf{x}_i + b) - 1]$$

$$\alpha_i \geq 0 \qquad\qquad d_i(\mathbf{w}^T\mathbf{x}_i + b) - 1 \geq 0$$

- Maksimum je kada su svi pribrojnici jednaki nuli
  - *α_i će biti različiti od nule samo kada vrijedi*

$$d_i(\mathbf{w}^T\mathbf{x}_i + b) - 1 = 0$$

Postupak optimizacije

- Jednadžba opisuje potporne vektore

$$d_i(\mathbf{w}^T\mathbf{x}_i + b) - 1 = 0$$

- Lagrangeovi multiplikatori α_i koji su različiti od nule automatski odabiru potporne vektore

Postupak optimizacije

- Za određivanje Lagrangeovih multiplikatora koristi se dualni problem

$$Q(\alpha)=\sum_{i=1}^{N}\alpha_i-\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j d_i d_j \mathbf{x}_i^T\mathbf{x}_j$$

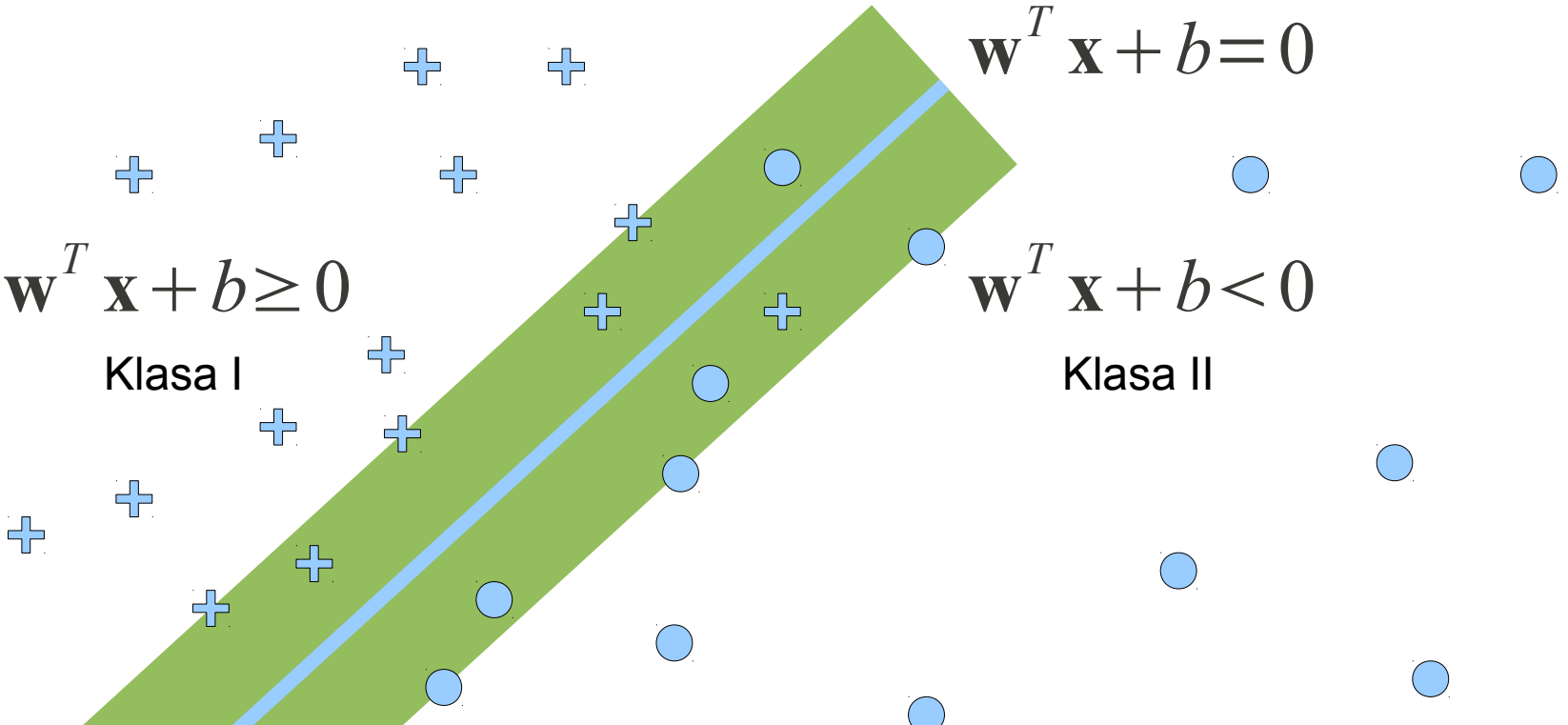
$$\sum_{i=1}^{N}\alpha_i d_i=0$$

$$\alpha_i\geq 0$$

$$\mathbf{w}_o=\sum_{i=1}^{N}\alpha_{o,i}d_i\mathbf{x}_i \qquad b_o=1-\mathbf{w}_i^T\mathbf{x}^{(s)},\quad d^{(s)}=1$$

Razdvajanje linearno neseparabilnih klasa

- Konačni postupak je praktički identičan kao u slučaju linearno separabilnih klasa



Razdvajanje linearno neseparabilnih klasa

$$d_i(\mathbf{w}^T \mathbf{x_i} + b) \geq 1 - \xi$$

- Cilj je smanjiti prosječnu grešku klasifikacije

$$\Phi(\xi) = \sum_{i=1}^{N} I(\xi_i - 1)$$

$$I(\xi) = \begin{cases} 0 & ako\ \xi \leq 0 \\ 1 & ako\ \xi > 0 \end{cases}$$

Optimizacija

- Pojednostavimo problem aproksimacijom

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

- I proširimo ga sa minimizacijom euklidske norme od **w**

$$\Phi(\xi, \mathbf{w}) = \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{i=1}^{N} \xi_i$$

Optimizacija

- Rješenje se opet traži kroz dualni problem

$$Q(\alpha)=\sum_{i=1}^{N}\alpha_i-\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j d_i d_j \mathbf{x}_i^T\mathbf{x}_j$$

$$\sum_{i=1}^{N}\alpha_i d_i=0$$

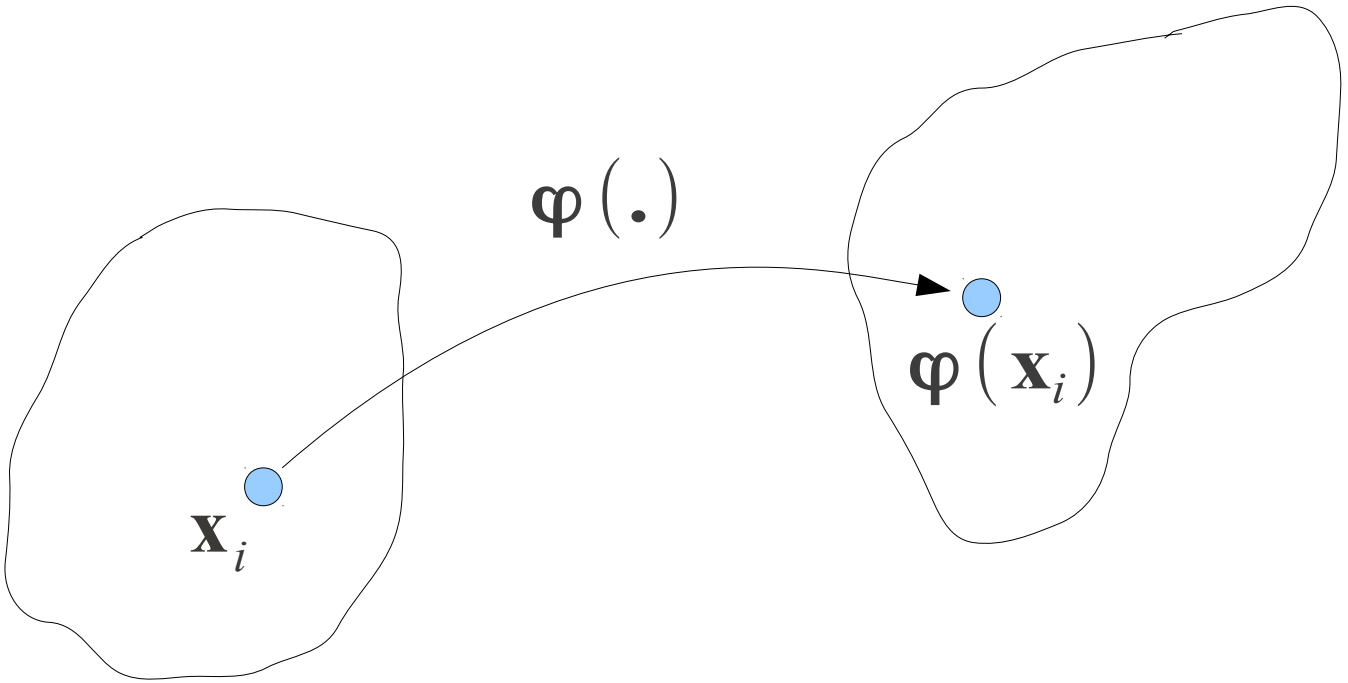
$$0\le\alpha_i\le C$$

$$\mathbf{w_o}=\sum_{i=1}^{N}\alpha_{o,i}d_i\mathbf{x}_i \qquad b_o=\frac{1}{N_s}\sum_{i=1}^{N_s}d_i\left(1-\mathbf{w}_i^T\mathbf{x}_i^{(s)}\right)$$

Stroj s potpornim vektorima za klasifikaciju

- Ukoliko uzorci nisu linearno separabilni, bilo bi lijepo kada bi ih mogli takvima učiniti
- Tada bi lagano mogli primijeniti proučenu separaciju
- Prelaskom u višedimenzionalni prostor, raste vjerojatnost linearne separabilnosti (Coverov teorem)
- Osnovna ideja je
  - Nelinearno mapiranje ulaznog prostora u novi prostor značajki više dimenzionalnosti
  - Konstrukcija optimalne ravnine razdvajanja u novom prostoru značajki

Nelinearno preslikavanje



$$\boldsymbol{\varphi}(\mathbf{x}) = [\varphi_0(\mathbf{x}), \varphi_1(\mathbf{x}), \ldots \varphi_m(\mathbf{x})]$$

Linearna separacija u prostoru značajki

- Optimalna ravnina razdvajanja konstruira se u novom višedimenzionalnom prostoru značajki

$$\sum_{j=1}^{m} w_j \varphi_j(\mathbf{x}) + b = 0$$

- $\varphi_j(\mathbf{x})$ su m transformacijksih funkcija
- m je broj dimenzija u novom prostoru značajki

Linearna separacija u prostoru značajki

- Uključivanje pomaka *b* u vektor težina **w**, kao prvog člana

$$\sum_{j=1}^{m} w_j \varphi_j(\mathbf{x}) = 0$$

$$\varphi_0(\mathbf{x}) = 1$$

$$w_0 = b_o$$

Ravnina razdvajanja

- Ravnina razdvajanja

$$\mathbf{w}^T \boldsymbol{\varphi}(\mathbf{x}) = 0$$

- Traženi **w** sada možemo izraziti kao

$$\mathbf{w} = \sum_{i=1}^{N} \alpha_i d_i \boldsymbol{\varphi}(\mathbf{x}_i)$$

- Kombiniranjem gornje dvije jednadžbe dobijemo

$$\sum_{i=1}^{N} \alpha_i d_i \boldsymbol{\varphi}^T(\mathbf{x}_i) \boldsymbol{\varphi}(\mathbf{x}) = 0$$

Funkcija jezgre unutarnjeg produkta

- Unutarnji produkt dvaju vektora u novom prostoru značajki

$$\boldsymbol{\varphi}^T(\mathbf{x}_i)\boldsymbol{\varphi}(\mathbf{x})$$

- Uvodimu novu funkciju jezgre *K*

$$K(\mathbf{x}_i,\mathbf{x})=\boldsymbol{\varphi}^T(\mathbf{x}_i)\boldsymbol{\varphi}(\mathbf{x})$$

- Čime dobijemo novu jednadžbu ravnine razdvajanja

$$\sum_{i=1}^{N}\alpha_i d_i K(\mathbf{x}_i,\mathbf{x})=0$$

Mercerov teorem

- Neka je K(**x**,**x**') simetrična funkcija jezgre definirana na zatvorenim intervalima od **x** i **x**'
- Takva se jezgra može rastaviti na slijedeći niz:

$$K(\mathbf{x}_i, \mathbf{x}) = \sum_{i=1}^{\infty} \lambda_i \varphi_i(\mathbf{x}) \varphi_i(\mathbf{x}')$$

- Uvjeti za to predstavljaju uvjete uz koje je jezgra K ujedno i jezgra unutarnjeg produkta
  - Broj dimenzija prostora značajki teoretski je beskonačan

Optimizacija

$$Q(\alpha)=\sum_{i=1}^{N}\alpha_i-\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j d_i d_j K(\mathbf{x}_i,\mathbf{x}_j)$$

$$\sum_{i=1}^{N}\alpha_i d_i=0$$

$$0\le\alpha_i\le C$$

$$\mathbf{w_o}=\sum_{i=1}^{N}\alpha_{o,i}d_i\boldsymbol{\varphi}(\mathbf{x}_i)$$

Primjeri funkcija jezgre unutarnjeg produkta

- Sloboda u izboru funkcija jezgre postoji no sve moraju zadovoljavati Mercerov teorem
- Tipični primjeri su:
  - Polinomna jezgra
  - Radijalna jezgra
  - Dvoslojni perceptron
- Dimenzionalnost prostora značajki ovisi o broju potpornih vektora

Polinomska jezgra

$$K(\mathbf{x}, \mathbf{x}_i) = (\mathbf{x}^T \mathbf{x}_i + 1)^p$$

- Parametar *p* se određuje apriori
- $\mathbf{x}_i$ su odabrani vektori potpore

Radijalna jezgra

$$K(\mathbf{x},\mathbf{x}_i)=e^{\left(-\frac{1}{2\sigma^2}\|\mathbf{x}-\mathbf{x}_i\|^2\right)}$$

- Parametar širine radijalne funkcije σ^2 određuje se a priori
- Broj radijalnih funkcija i njihovi centri određeni su izborom vektora potpore

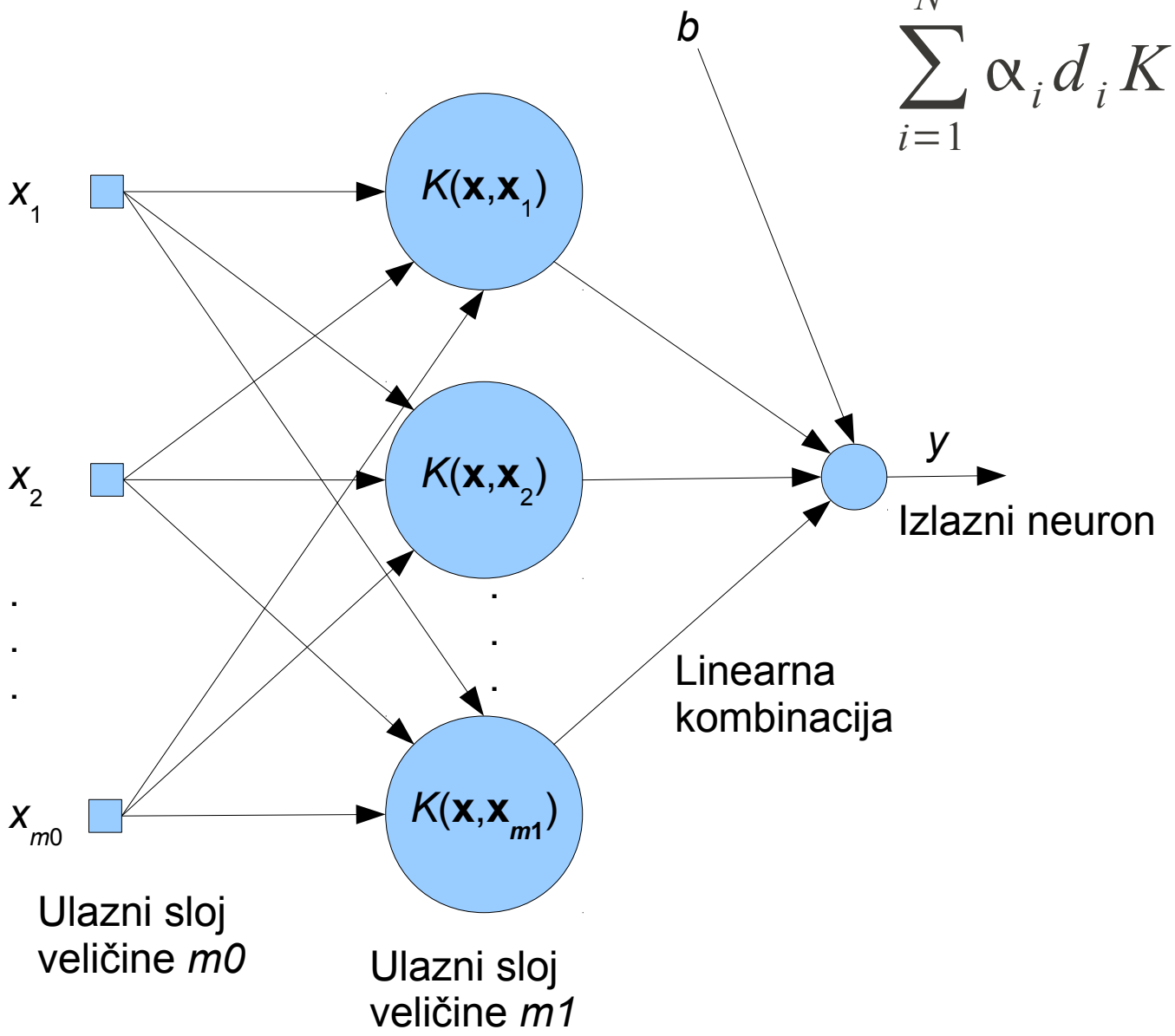
Dvoslojni perceptron

$$K(\mathbf{x}, \mathbf{x}_i) = \text{tnah}(\beta_0 \mathbf{x}^T \mathbf{x}_i + \beta_1)$$

- Mercerov teorem zadovoljen je samo za neke kombinacije parametara β_0 i β_1

Arhitektura SVM

$$\sum_{i=1}^{N} \alpha_i d_i K(\mathbf{x}_i, \mathbf{x}) = 0$$



Primjer: XOR problem

XOR problem	
Ulaz **x**	Željeni izlaz d_i
(-1,-1)	- 1
(-1,+1)	+1
(+1,-1)	+1
(+1,+1)	-1

$$K(\mathbf{x},\mathbf{x}_i)=(\mathbf{x}^T\mathbf{x}_i+1)^2$$

$$\mathbf{x}=[x_1, x_2]^T$$

$$K(\mathbf{x},\mathbf{x}_i)=1+x_1^2x_{i1}^2+2\,x_1x_2x_{i1}x_{i2}+x_2^2x_{i2}^2+2\,x_1x_{i1}+2\,x_2x_{i2}$$

$$\boldsymbol{\phi}(\mathbf{x})=\left[1, x_1^2, \sqrt{2}\,x_1x_2, x_2^2, \sqrt{2}\,x_1, \sqrt{2}\,x_2\right]$$

Primjer: XOR problem

$$Q(\alpha)=\sum_{i=1}^{4}\alpha_i-\frac{1}{2}\sum_{i=1}^{4}\sum_{j=1}^{4}\alpha_i\alpha_j d_i d_j K(\mathbf{x}_i,\mathbf{x}_j)$$

- Optimizacija daje slijedeće vrijednosti Lagrangeovih koeficijenata

$$\alpha_{o1}=\alpha_{o2}=\alpha_{o3}=\alpha_{o4}=\frac{1}{8}$$

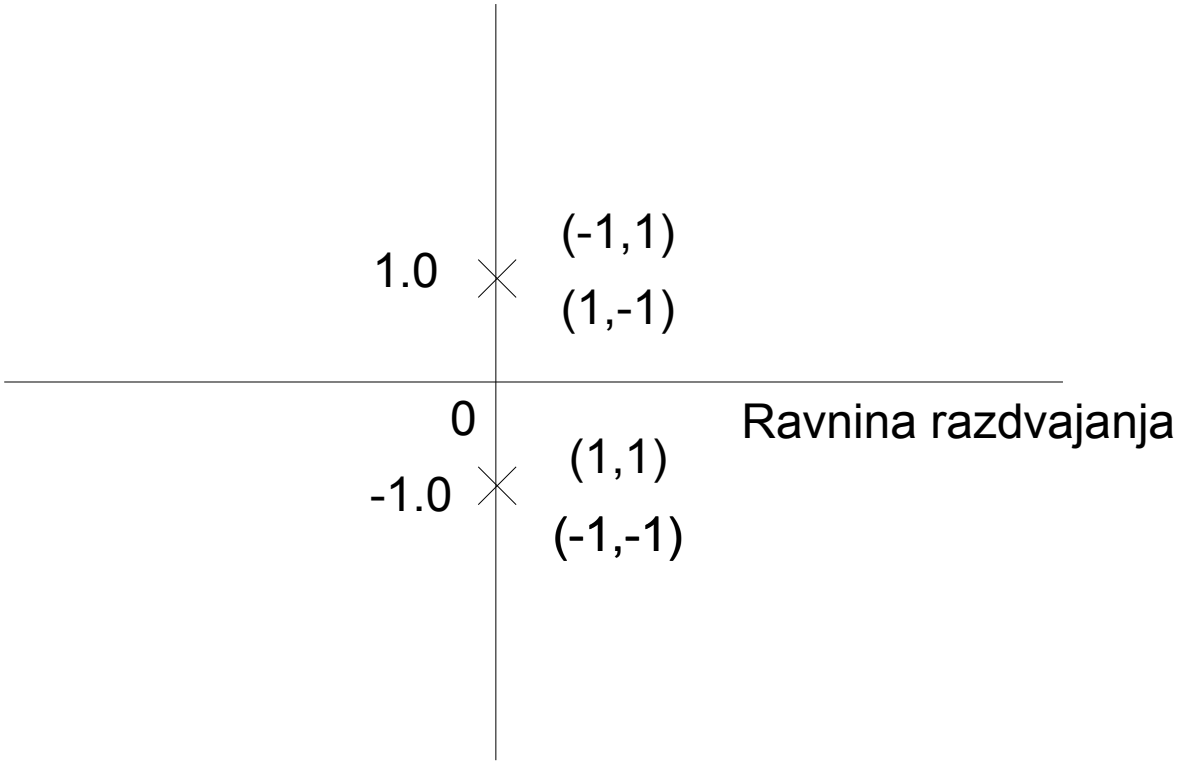
$$\mathbf{w_o}=\sum_{i=1}^{N}\alpha_{o,i}d_i\boldsymbol{\varphi}(\mathbf{x}_i)=\begin{bmatrix}0\\0\\-1/\sqrt{2}\\0\\0\\0\end{bmatrix}$$

Primjer: XOR problem

$$0 = \mathbf{w}_o^T \boldsymbol{\varphi}(\mathbf{x}) = [0, 0, -1/\sqrt{,} 2, 0, 0, 0] \begin{bmatrix} 1 \\ x_1^2 \\ \sqrt{2}\, x_1 x_2 \\ x_2^2 \\ \sqrt{2}\, x_1 \\ \sqrt{2}\, x_2 \end{bmatrix} = -x_1 x_2$$

Primjer: XOR problem

$$-x_1 x_2 = 0$$



Kako sve funkcionira

1. Pripremiti skup za treniranje
2. Odabrati funkciju jezgre unutarnjeg produkta *K* koja zadovoljava Mercerov teorem
3. Određivanje optimalnih α_i

$$Q(\alpha)=\sum_{i=1}^{N}\alpha_i-\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j d_i d_j K(\mathbf{x}_i,\mathbf{x}_j),\ \sum_{i=1}^{N}\alpha_i d_i=0,\ 0\le\alpha_i\le C$$

   (time su određeni i potporni vektori)

4. Klasifikacija prema jednadžbi diskriminacije

$$g(\mathbf{x})=\sum_{i=1}^{N}\alpha_i d_i K(\mathbf{x}_i,\mathbf{x})$$

SVM: Prednosti i nedostaci

- Prednosti
  - Pronalazak ekstrema funkcije cilja je zagarantirana
  - Mogućnost efikasne implementacije optimizacije
  - Separacija u višedimenzionalnom prostoru značajki bez da ga ikad posjetimo
- Nedostaci
  - Brzina izvođenja – nema direktne kontrole broja potpornih vektora
  - Nije moguće prilagođavati arhitekturu mreže prema apriori znanju o problemu
    - Rješenje: konstrukcija "umjetnih" uzoraka za treniranje prema apriori znanju
    - Rješenje: uvođenje dodatnih uvjeta u funkciju cilja

Teme predavanja

- Problem kalsifikacije linearno separabilnih klasa
- Margina razdvajanja
- Vektori potpore
- Klasifikacija linearno neseparabilnih klasa
- Nelinearno preslikavanje u prostor značajki

Zadaci

1. Pokažite da je margina razdvajanja jednaka $2/\|\mathbf{w}_o\|$ ako je za ravninu razdvajanja $\mathbf{w}_o^T \mathbf{x} + b_o = 0$ zadan dodatni uvjet $\min_{i=1,2,...,N} |\mathbf{w}_o^T \mathbf{x} + b_o| = 1$

2. Kod polinomne jezgre u XOR primjeru, odredite minimalni pozitivni eksponent *p* za koji je moguće naći rješenje problema