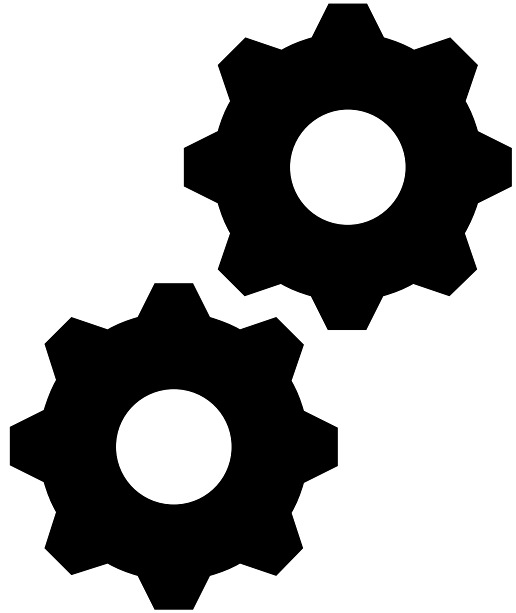


# Treniranje modela

# Treniranje modela

- Kako među svim hiperravnima-kandidatima $f(x) = \theta^T x = 0$ pronaći onu sa najvećom marginom?

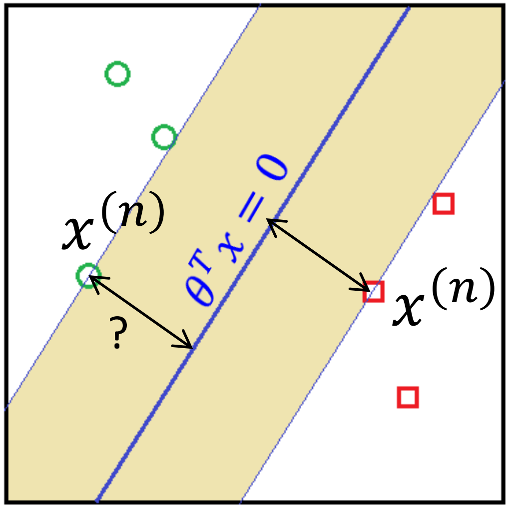
- Šta znamo:
  - Hiperravni su predstavljene jednačinom $f(x) = \theta^T x = 0$
  - Uslov: linearna separabilnost – za svako $\left(x^{(i)}, y^{(i)}\right)$ mora da važi $y^{(i)} \cdot f\left(x^{(i)}\right) > 0$
  - Ima beskonačno mnogo $\theta$ koji bi ovo zadovoljili
  - Od svih takvih jednačina, mi želimo da nađemo $\theta$ tako da $\theta^T x = 0$ bude ravan sa maskimalnom marginom?
- Naš cilj:
  1. Izraziti marginu kao funkciju parametara $\theta$
  2. Naći $\theta$ za koje ta funkcija dostiže maksimum

# Cilj: izraziti širinu margine kao funkciju $\theta$

- Uvešćemo par „tehnikalija“ koje će nam olakšati analizu

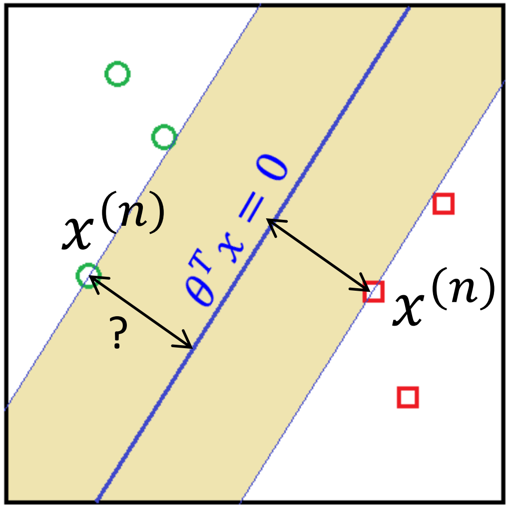
1. Klase smo obeležili sa $y = +1$ i $y = -1$
    - Kako ćemo obeležiti klase je svejedno, ovo biramo zato što nam je zgodno za analizu

# Cilj: izraziti širinu margine kao funkciju $\theta$




2. Uvešćemo ograničenje $\left|\theta^T x^{(n)}\right| = 1$
   - Sa $x^{(n)}$ smo obeležili vektore potpore
   - Za sve tačke skupa podataka važi $\left|\theta^T x^{(i)}\right| > 0$
   - A mi ćemo, pored toga, specijalno za vektore potpore $x^{(n)}$, insistirati i da je $\left|\theta^T x^{(n)}\right| = 1$

# Cilj: izraziti širinu margine kao funkciju $\theta$



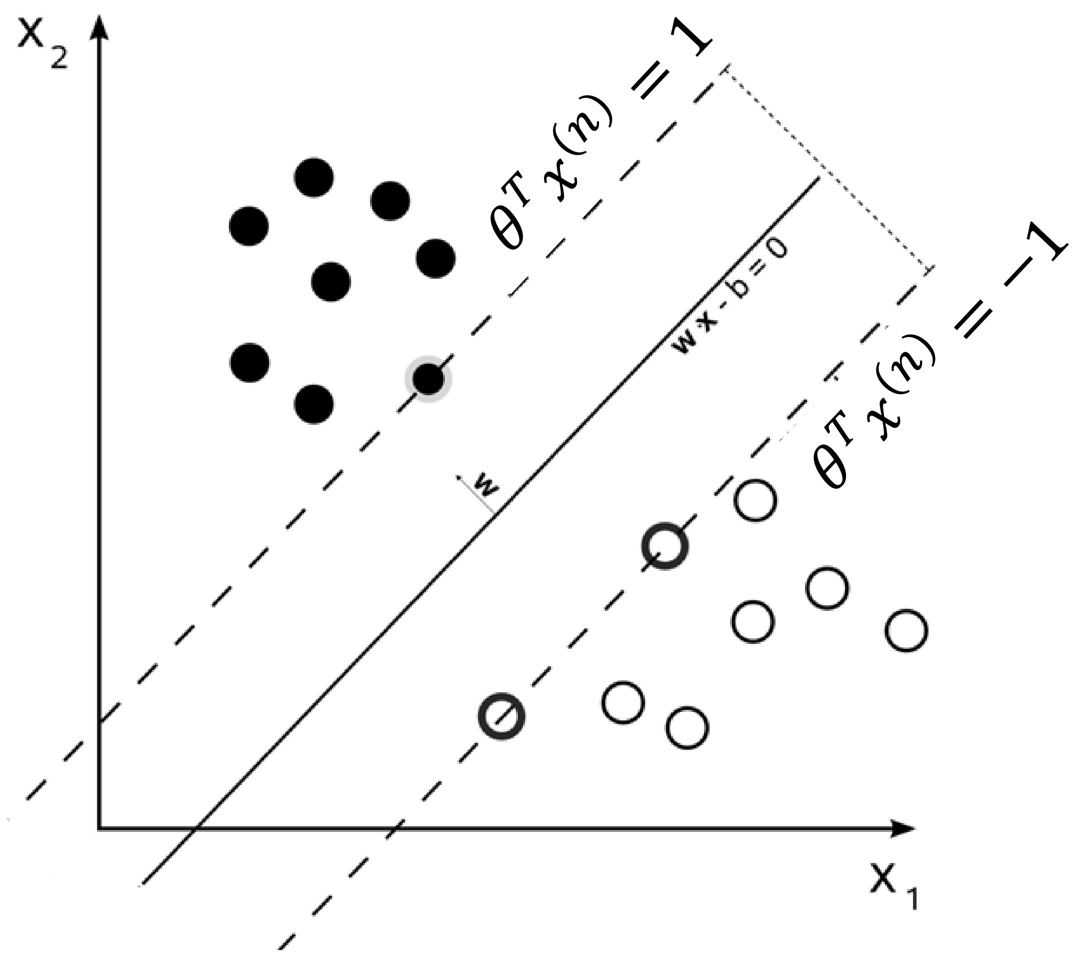

2. Uvešćemo ograničenje $\left|\theta^T x^{(n)}\right| = 1$

(Kako za istu hiperravan ne bismo imali beskonačno mnogo reprezentacija)

- Jednačina hiperravni je $\theta^T x = 0$. Ako pomnožimo $\theta$ bilo kojim skalarom, ovo je identična hiperravan!
- Npr. prave $2 \cdot x_1 - x_2 - 1 = 0$ i $6 \cdot x_1 - 3 \cdot x_2 - 3 = 0$ su identične, a opisane različitim parametrima: $\theta^{(1)} = [2, -1, -1]$ i $\theta^{(2)} = [6, -3, -3]$
- Biranjem $\left|\theta^T x^{(n)}\right| = 1$ nismo izgubili na opštosti: i dalje možemo predstaviti svaku hiperravan

# Cilj: izraziti širinu margine kao funkciju $\theta$

2. Uvešćemo ograničenje $\left|\theta^T x^{(n)}\right| = 1$

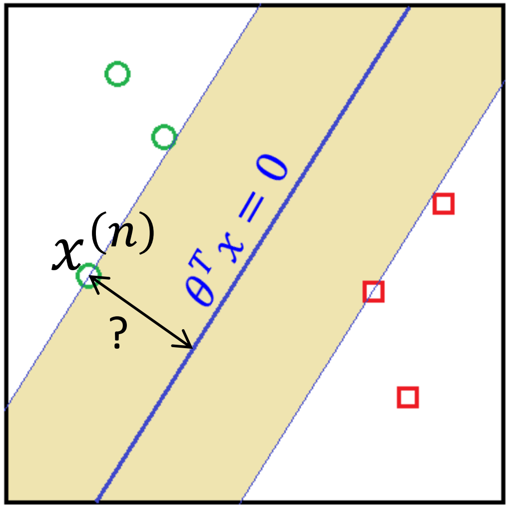


# Cilj: izraziti širinu margine kao funkciju $\theta$

3. Izvućićemo $\theta_0$ iz vektora i tretirati ga posebno (u odnosu na ostale parametre)

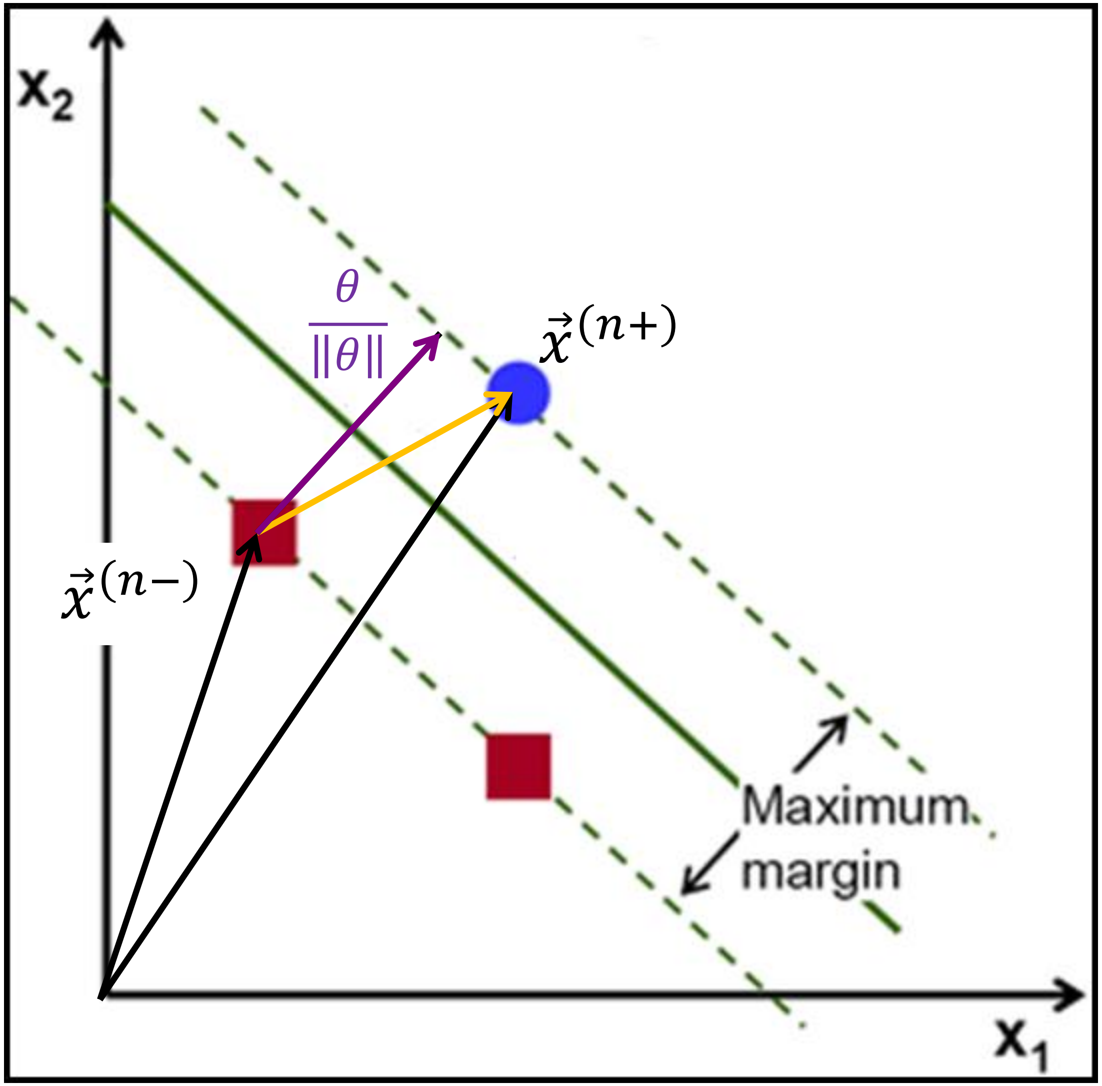
- Umesto dosadašnjeg obeležavanja $\theta = [\theta_0, \theta_1, \dots, \theta_D]$, usvojićemo sledeće obeležavanje: $\theta = [\theta_1, \dots, \theta_D]$, a $\theta_0$ ćemo označiti sa $b$
- Iz $x$ ćemo izbaciti $x_0 = 1$
- Hiperravan je: $f(x) = \theta^T x + b = 0$
- Ponovo nismo izgubili na opštosti, samo biramo malo drugačiju reprezentaciju parametara

# Cilj: izraziti širinu margine kao funkciju $\theta$




- Širina margine je rastojanje vektora potpore $x^{(n)}$ i hiperravni
- Dogovorili smo se da ćemo odabrati $\theta$ tako da važi $\left|\theta^T x^{(n)} + b\right| = 1$
- Znamo da su podaci linearno separabilni: $y^{(i)} \cdot f\left(x^{(i)}\right) > 0$
- Dakle, za vektore potpore $\left(x^{(n)}, y^{(n)}\right)$ važi $y^{(n)}\left(\theta^T x^{(n)} + b\right) = 1$

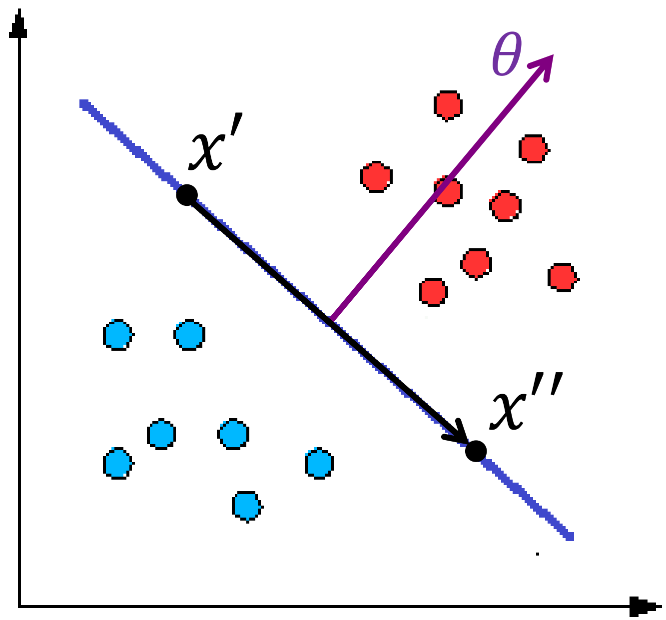
# Cilj: izraziti širinu margine kao funkciju $\theta$

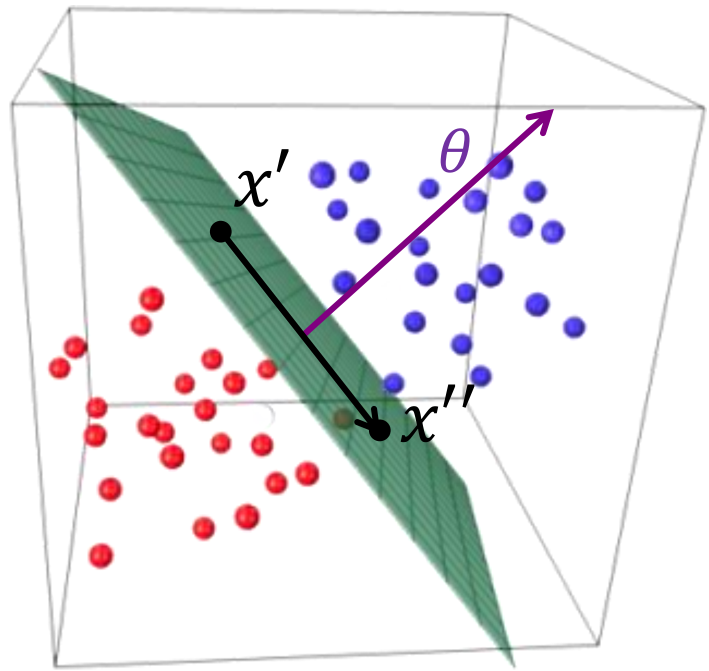



- Ako bismo vektor $\vec{x}^{(n+)} - \vec{x}^{(n-)}$ projektovali na jedinični vektor normalan na razdvajajuću hiperavan, dužina projekcije bi bila jednaka margini

1. Pokazaćemo da je vektor $\theta$ normalan na hiperravan
2. Projektovaćemo $\vec{x}^{(n+)} - \vec{x}^{(n-)}$ na $\frac{\vec{\theta}}{\|\theta\|}$ (normalizujemo $\theta$ tako da bude jedinični vektor)

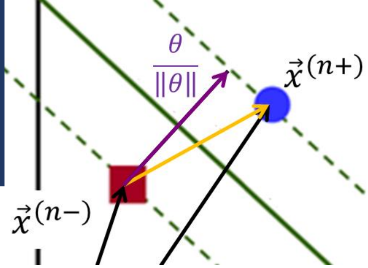
# 1. Vektor $\vec{\theta}$ je normalan na hiperravan




- Uzmimo bilo koje dve tačke $x'$ i $x''$ na hiperravni.
- One moraju da zadovoljavaju jednačinu hiperravni $\theta^T x + b = 0$, dakle: sledi $\theta^T x' + b = 0$ i $\theta^T x'' + b = 0$. Ako oduzmemo ovde dve jednačine: $\theta^T(x'' - x') = 0$
- Pošto su $\vec{\theta}$ i $\vec{x}'' - \vec{x}'$ ne-nula vektori, to znači da mora da su ortogonalni
- Dakle, pošto je $\vec{\theta}$ ortogonalan na bilo koji vektor na hiperravni, mora da je ortogonalan na hiperravan

# 2. Projektujemo $\vec{x}^{(n+)} - \vec{x}^{(n-)}$ na $\dfrac{\vec{\theta}}{\|\theta\|}$



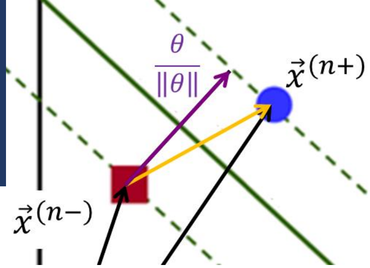

- Treba nam projekcija na jedinični vektor normalan na razdvajajuću hiperravan pa normalizujemo $\theta$:

$$\hat{\theta} = \theta / \|\theta\|, \|\theta\| = \sqrt{\theta_1^2 + \theta_2^2 + \cdots + \theta_D^2}$$

- Projekcija se dobija skalarnim proizvodom tih vektora
  - Projekcija je $\hat{\theta}^T\left(x^{(n+)} - x^{(n-)}\right)$, ali ovaj broj može biti pozitivan ili negativan
  - Zbog toga nam za razdaljinu treba apsolutna vrednost

$$distance = \left|\hat{\theta}^T\left(x^{(n+)} - x^{(n-)}\right)\right| = \frac{1}{\|\theta\|}\left|\theta^T x^{(n+)} - \theta^T x^{(n-)}\right|$$

# Margina izražena preko $\theta$




$$distance = \frac{1}{\|\theta\|}\left|\theta^T x^{(n+)} - \theta^T x^{(n-)}\right|$$

- Tačka $x^{(n+)}$:
  - Njena labela je $y^{(n+)} = +1$
  - Postavili smo uslov da za vektore potpore važi: $y^{(n+)}\left(\theta^T x^{(n+)} + b\right) = 1 \Rightarrow \theta^T x^{(n+)} = 1 - b$

- Tačka $x^{(n-)}$:
  - Njena labela je $y^{(n-)} = -1$
  - Postavili smo uslov da za vektore potpore važi: $y^{(n-)}\left(\theta^T x^{(n-)} + b\right) = 1 \Rightarrow \theta^T x^{(n-)} = -1 - b$

- Dakle: $distance = \frac{1}{\|\theta\|}|1 - b - (-1 - b)| = \frac{2}{\|\theta\|}$

# Margina izražena preko $\theta$

- Uspeli smo da izrazimo marginu kao funkciju $\theta$
- Želimo da odaberemo $\theta$ tako da margina bude maksimalna, dakle, naš optimizacioni problem je:

$$\max_{\theta} \frac{2}{\|\theta\|}$$

  - Možemo izbaciti konstantu 2
  - Umesto maksimizacije $\frac{1}{\|\theta\|}$ možemo minimizovati $\|\theta\|$
  - Ovo je ekvivalentno minimizaciji $\frac{1}{2}\|\theta\|^2 = \frac{1}{2}\theta^T\theta$
    - (ovo nam je matematički pogodnije jer je $\|\theta\| = \sqrt{\theta_1^2 + \theta_2^2 + \cdots + \theta_D^2}$ )

- Naš optimizacioni problem postaje:

$$\min_{\theta} \frac{1}{2}\theta^T\theta$$

# Pod kojim uslovom?

- Ranije smo uveli uslov da za vektore potpore važi $\left|\theta^T x^{(n)} + b\right| = 1$
- Vektori potpore su tačke najbliže razdvajajućoj hiperravni pa je za njih vrednost $\left|\theta^T x^{(n)} + b\right|$ najmanja u celom skupu podataka
- Dakle uslov možemo zapisati kao $\min_{i=1,2,\ldots,N} \left|\theta^T x^{(i)} + b\right| = 1$

# Pod kojim uslovom?

- Nezgodan nam je *min* u uslovu, pa ćemo zameniti uslov ekvivalentnim:

$$\min_{i=1,2,\dots,N} \left|\theta^T x^{(i)} + b\right| = 1 \Leftrightarrow \left|\theta^T x^{(i)} + b\right| \geq 1, \text{za svako } i \in \{1, \dots, N\}$$

(1) (2)

- Da li su uslovi (1) i (2) ekvivalentni?
  - Ako je uslov (1) ispunjen, onda je svakako i uslov (2)
  - Problem: šta ako dobijemo $\theta$ takvo da važi $\left|\theta^T x^{(i)} + b\right| > 1$ za svako $i \in \{1, \dots, N\}$?
    - To se ne slaže time da mora postojati takvo $x^{(n)}$ da je $\left|\theta^T x^{(n)} + b\right| = 1$
  - Da li se ovakva situacija može desiti?
    - Recimo da postoji rešenje $\min_{\theta} \frac{1}{2}\theta^T\theta$ takvo da je $\left|\theta^T x^{(i)} + b\right| > 1$ za svako $i \in \{1, \dots, N\}$
    - Onda mi možemo naći bolje! Skaliraćemo $\theta$ i $b$, sve dok se ne desi da postoji $x^{(n)}$ da je $\left|\theta^T x^{(n)} + b\right| = 1$
    - Dobili smo manje $\theta$ i $b$ – dakle, bolje rešenje optimizacionog problema $\min_{\theta} \frac{1}{2}\theta^T\theta$

# Pod kojim uslovom?

$$\left|\theta^T x^{(i)} + b\right| \geq 1, \text{za svako } i \in \{1, \dots, N\}$$

- Da bismo se rešili apsolutne vrednosti:

$$\left|\theta^T x^{(i)} + b\right| = y^{(i)}\left(\theta^T x^{(i)} + b\right)$$

  - Za pozitivne primere: $y^{(i)} = +1$, a $\theta^T x^{(i)} + b > 0$
  - Za negativne primere: $y^{(i)} = -1$, a $\theta^T x^{(i)} + b < 0$

Dakle, imamo optimizacioni problem:

$$\min_{\theta} \frac{1}{2} \theta^T \theta$$

Sa ograničenjima:

$$y^{(i)}\left(\theta^T x^{(i)} + b\right) - 1 \geq 0 \text{ za svako } i \in \{1, \dots, N\}$$

# Kako se pronalazi minimum pod ograničenjima?

- Standardan način za rešavanje optimizacionih problema sa ograničenjima tipa jednakosti i nejednakosti je KKT (*Karush-Kuhn-Tucker*)
  - Generalizacija metode Lagranžovih množilaca
  - Problem pronalaženja ekstrema pod ograničenjima zamenjujemo ekvivalentnim problemom bez ograničenja

$$L(\theta, b, \alpha) = \frac{1}{2}\theta^T\theta - \sum_{i=1}^{N} \alpha_i \left[y^{(i)}\left(\theta^T x^{(i)} + b\right) - 1\right]$$

- Optimizacioni problem $\min_{\theta,b} \max_{\alpha,\alpha\geq 0} L(\theta, b, \alpha)$ je ekvivalentan našem problemu
  - Ako su uslovi $y^{(i)}\left(\theta^T x^{(i)} + b\right) - 1 \geq 0$ ispunjeni, $\max_{\alpha,\alpha\geq 0} L(\theta, b, \alpha) = \frac{1}{2}\theta^T\theta$
  - Ako je neki od uslova narušen, $\max_{\alpha,\alpha\geq 0} L(\theta, b, \alpha) = \infty$

# Kako se pronalazi minimum pod ograničenjima?

$$L(\theta, b, \alpha) = \frac{1}{2}\theta^T\theta - \sum_{i=1}^{N} \alpha_i \left[y^{(i)}\left(\theta^T x^{(i)} + b\right) - 1\right]$$

- Izjednačimo parcijalni izvod po $\theta$ i $b$ sa 0:
  - $\frac{\partial L(\theta,b,\alpha)}{\partial\theta} = \theta - \sum_{i=1}^{N} \alpha_i\, y^{(i)} x^{(i)} = 0 \Rightarrow \theta = \sum_{i=1}^{N} \alpha_i y^{(i)} x^{(i)}$
  - $\frac{\partial L(\theta,b,\alpha)}{\partial b} = \sum_{i=1}^{N} \alpha_i y^{(i)} = 0$
- Zamenićemo dobijene izraze u $L(\theta, b, \alpha)$:

$$L(\theta, b, \alpha) = \sum_{i=1}^{N} \alpha_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} y^{(i)} y^{(j)} \alpha_i \alpha_j x^{(i)T} x^{(j)}$$

  - Ispalo je da $L(\theta, b, \alpha)$ zavisi samo od $\alpha$ ($\theta$ i $b$ su nestali)!
  - Treba da maksimizujemo ovaj izraz po $\alpha$ pod uslovom $\alpha_i \geq 0$ za $i \in \{1, \dots, N\}$ i $\sum_{i=1}^{N} \alpha_i y^{(i)} = 0$

Za detaljno izvođenje ove zamene i sređivanja izraza pogledajte https://youtu.be/eHsErlPJWUU?t=2172

# Kako se pronalazi minimum pod ograničenjima?

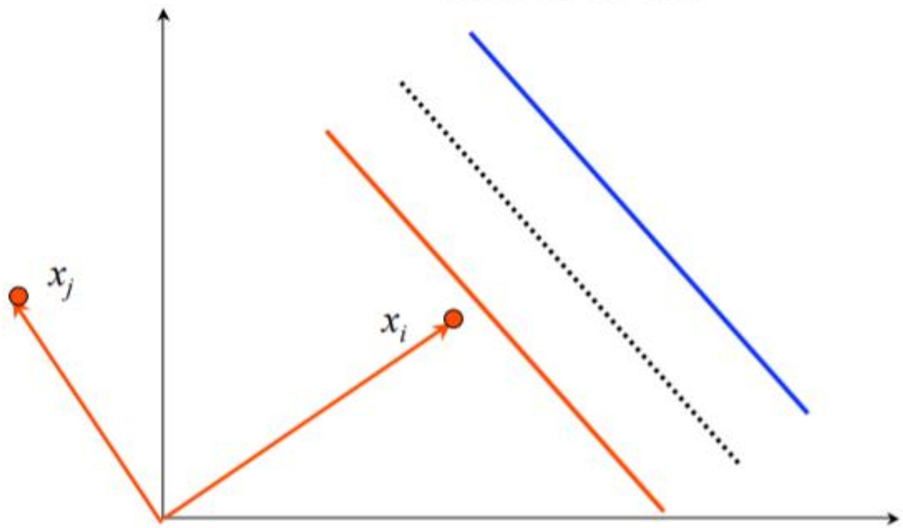
- Treba da maksimizujemo ovaj izraz:

$$L(\theta, b, \alpha) = \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} y^{(i)} y^{(j)} \alpha_i \alpha_j \boxed{x^{(i)T} x^{(j)}}$$

- Optimizacija zavisi od skalarnog proizvoda parova primera
- Intuicija za skalarni proizvod:
  - Skalarni proizvod daje neku meru *sličnosti*: skalarni proizvod dva jedinična vektora predstavlja kosinus ugla između njih (koliko su udaljeni)
  - Ako su vektori paralelni, njihov skalarni proizvod je 1 (potpuno slični)
  - Ako su normalni, njihov skalarni proizvod je 0 (potpuno neslični)

# Intuicija iza rešenja

2 dissimilar (orthogonal) vectors don't count at all

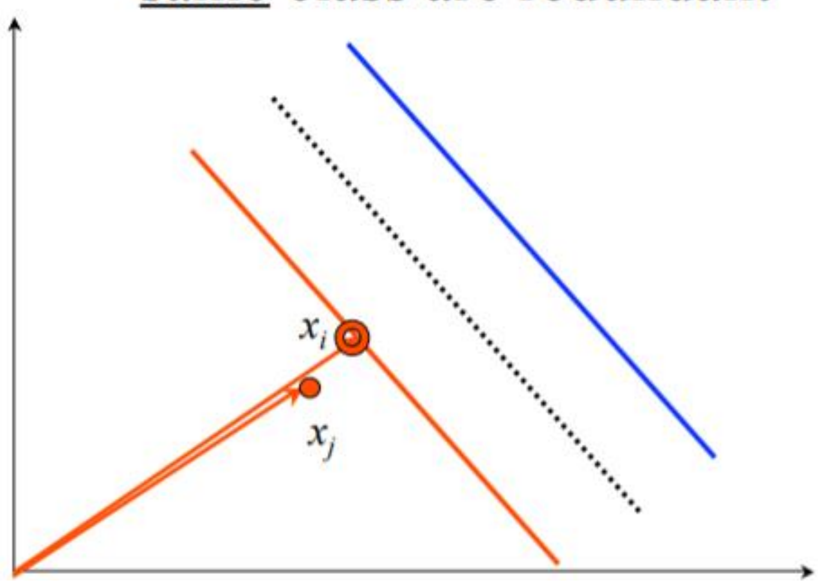



$$L(\theta, b, \alpha) = \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} y^{(i)} y^{(j)} \alpha_i \alpha_j x^{(i)T} x^{(j)}$$

- Za potpuno različite primere, $x^{(i)} \cdot x^{(j)} = 0 \rightarrow$ ne utiču na $L$

# Intuicija iza rešenja

2 vectors that are similar but predict the same class are redundant



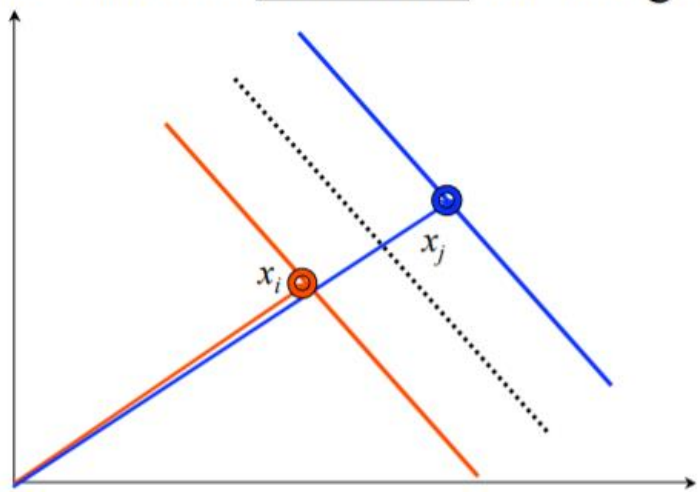

$L(\theta, b, \alpha)$

$$= \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} y^{(i)} y^{(j)} \alpha_i \alpha_j x^{(i)T} x^{(j)}$$

- Za potpuno slične primere, važi $x^{(i)} \cdot x^{(j)} = 1$
- Ako su im i klase iste, važi $y^{(i)} y^{(j)} = 1$
- Dakle, $\alpha_i \alpha_j x^{(i)} x^{(j)} y^{(i)} y^{(j)} > 0$ i $L$ se smanjuje
- Optimizacioni algoritam koji maksimizuje $L$ će težiti da smanji $\alpha_i$ i $\alpha_j$

# Intuicija iza rešenja

Insight into inner products, graphically: 2 very very similar $x_i$, $x_j$ vectors that predict difft classes tend to <u>maximize</u> the margin width




$$L(\theta, b, \alpha) = \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} y^{(i)} y^{(j)} \alpha_i \alpha_j x^{(i)T} x^{(j)}$$

- Za potpuno slične primere, važi $x^{(i)} \cdot x^{(j)} = 1$
- Ako su im i klase različite, važi $y^{(i)} y^{(j)} = -1$
- Dakle, $\alpha_i \alpha_j x^{(i)} x^{(j)} y^{(i)} y^{(j)} < 0$ i $L$ se povećava
- Optimizacioni algoritam koji maksimizuje $L$ će težiti da poveća $\alpha_i$ i $\alpha_j$
- Ovo su upravo „kritični“ primeri za razlikovanje klasa koje tražimo

# Rešenje SVM optimizacionog problema

- Rešenje se dobija kvadratnim programiranjem i oblika je:

$$\theta = \sum_{i=1}^{N} \alpha_i y^{(i)} x^{(i)}$$

  gde su $\alpha_i$ Lagranžovi množioci za koje važi $0 \leq \alpha_i$

- Primetićemo da je većina dobijenih $\alpha_i = 0$!
  - Ovo je posledica našeg uslova $\alpha_i\left(y^{(i)}\left(\theta^T x^{(i)} + b\right) - 1\right) = 0$
  - Ako važi $y^{(i)}\left(\theta^T x^{(i)} + b\right) > 1$, onda mora da važi $\alpha_i = 0$
  - Ako važi $y^{(i)}\left(\theta^T x^{(i)} + b\right) = 1$ (potporni vektor), onda $\alpha_i$ može biti veće od 0
- Podaci $\left(x^{(i)}, y^{(i)}\right)$ za koje je $\alpha_i > 0$ su potporni vektori
- Težine $\theta \in \mathbb{R}^D$ je izražen putem svega nekoliko primera $\left(x^{(i)}, y^{(i)}\right)$ našeg skupa podataka – umesto $D$-dimenzionog sistema, imamo svega nekoliko efektivnih parametara – jako smo smanjili dimenzionalnost sistema!

# Rešenje SVM optimizacionog problema

- Hipoteza je:

$$f_{\theta,b}(x) = \sum_{i=1}^{N} \alpha_i y^{(i)} x^{(i)} \cdot x + b$$

- $\alpha_i > 0$ samo za vektore potpore

- Klasa se određuje funkcijom $sign\left(f_{\theta,b}(x)\right)$