Metode na bazi vektora nosača

- Metode na bazi vektora nosača
  - Klasifikator maksimalne margine (linearno razdvojiv slučaj)
    - Matematički model i interpretacija pojma vektora nosača
    - Ograničenja klasifikatora maksimalne margine
  - Klasifikator blage margine (linearno nerazdvojiv slučaj)
    - Modifikacija matematičkog modela
    - Potreba za daljim proširenjem
  - Mašine na bazi vektora nosača (eng. *Support Vector Machines* – SVM)
    - Preslikavanje problema u višedimenzionalni prostor

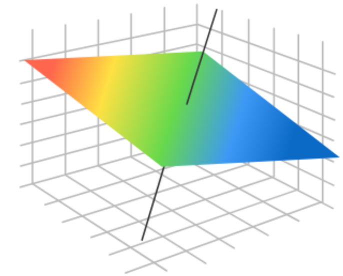
Metode na bazi vektora nosača (uvod)

- Pristup rešavanju problema binarne klasifikacije pronalaženjem hiperravni koja razdvaja uzorke različitih klasa u prostoru obeležja
- Osnovna ideja ovog pristupa bazira se na **klasifikatoru maksimalne margine**
  - Ovaj klasifikator, iako vrlo elegantan i jednostavan, primenljiv je samo u slučaju kada su uzorci dveju klasa linearno razdvojivi
  - Ako uzorci nisu linearno razdvojivi, može se ići u dva pravca:
    - Ne zahtevati striktno razdvajanje svih uzoraka već samo razdvajanje u što većoj meri
    - Preslikati problem u prostor više dimenzionalnosti u kom će uzorci biti barem približno linearno razdvojivi
- **Klasifikator blage margine** predstavlja proširenje klasifikatora maksimalne margine, primenljivo u slučaju kada su uzorci barem približno linearno razdvojivi
- **Mašine na bazi vektora nosača** (eng. *support vector machines – SVM*) predstavljaju dalje proširenje, kojim se mogu modelovati i nelinearne granice odlučivanja

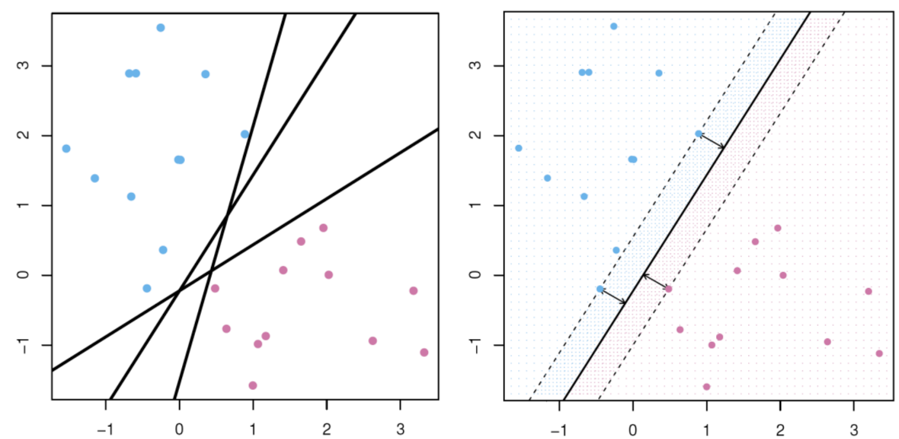
Klasifikator maksimalne margine

- Hiperravan u *d* dimenzija predstavlja ravan afini potprostor dimenzije *d*−1:

$$\vartheta_0 + \vartheta_1 x_1 + \vartheta_2 x_2 + \ldots + \vartheta_d x_d = 0$$

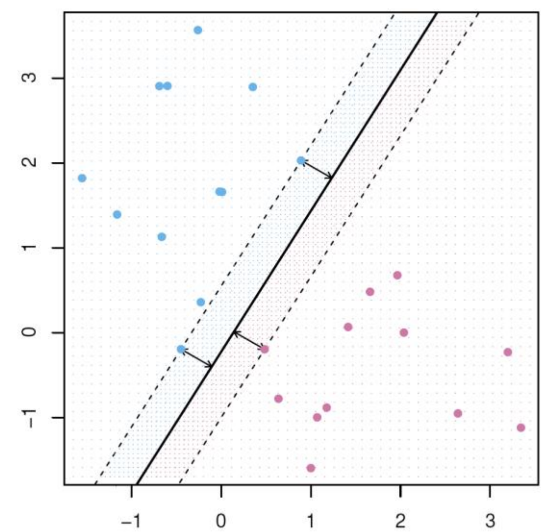


  - Tačke za koje je dati zbir pozitivan nalaze se sa jedne strane hiperravni a tačke gde je on negativan sa druge strane
  - Ako je $\vartheta_0 = 0$, hiperravan prolazi kroz koordinatni početak, inače ne
  - Vektor $\boldsymbol{\theta} = \begin{bmatrix}\vartheta_1 & \vartheta_2 & \ldots & \vartheta_d\end{bmatrix}^{\mathrm{T}}$ ortogonalan je na hiperravan
  - Hiperravan u 2D prostoru (u ravni) je prava
- Za problem binarne klasifikacije, hiperravan odlučivanja je hiperravan za koju važi:
  - za sve uzorke jedne klase $\vartheta_0 + \boldsymbol{\theta}^{\mathrm{T}}\mathbf{x} < 0$
  - za sve uzorke druge klase $\vartheta_0 + \boldsymbol{\theta}^{\mathrm{T}}\mathbf{x} > 0$
- Hiperravan koja razdvaja uzorke pojedinih klasa ne mora postojati, a ako postoji, obično nije jedinstvena
- Može se, međutim, uočiti hiperravan koja u najvećoj meri (sa najvećom marginom) razdvaja uzorke
  - Udaljenost novog uzorka od te hiperravni u vezi je s pouzdanošću odluke



Klasifikator maksimalne margine

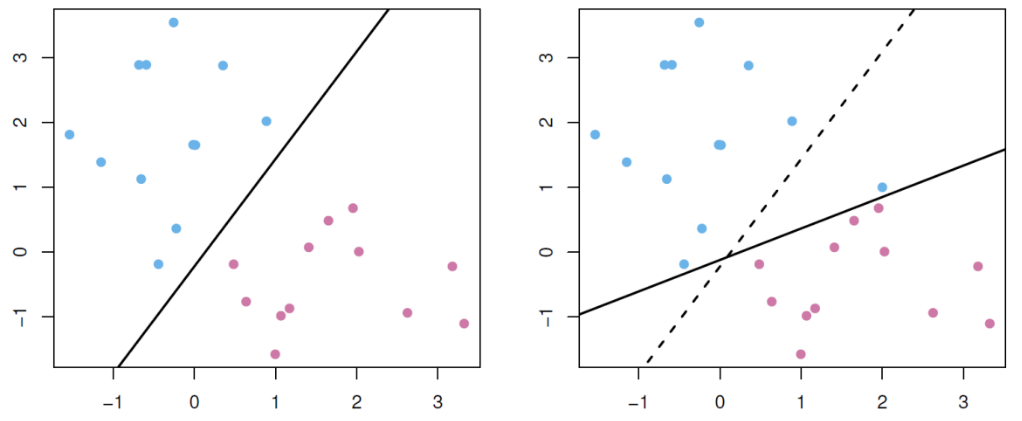
- *Margina M* je najmanje rastojanje od hiperravni odlučivanja do bilo kog uzorka iz skupa za obuku
- Vektori koji odgovaraju uzorcima na rastojanju *M* od hiperravni odlučivanja predstavljaju „vektore nosače“
  - pomeranje vektora nosača u opštem slučaju izazvalo bi i pomeranje hiperravni odlučivanja
  - relativno mali pomeraji ostalih uzoraka iz skupa za obuku nemaju nikakav uticaj
- Klasifikator maksimalne margine dobija se kao rešenje optimizacionog problema:
  - Maksimizovati *M* u odnosu na **θ** pod ograničenjima:
    - $\sum_{j=0}^{d} \vartheta_j^2 = 1$ (samo radi jedinstvenosti jednačine hiperravni)
    - $y_i \cdot (\vartheta_0 + \vartheta_1 x_1^{(i)} + \ldots + \vartheta_d x_d^{(i)}) \geq M, \quad \forall i = 1,2,\ldots,N$ (y_i = ±1 zavisno od klase)



- Dati optimizacioni problem može se rešiti korišćenjem Lagrangeovih multiplikatora, i hiperravan odlučivanja dobijena na ovaj način je jedinstvena
  - Vektor **θ** dobija se kao linearna kombinacija vektora nosača

Ograničenja klasifikatora maksimalne margine

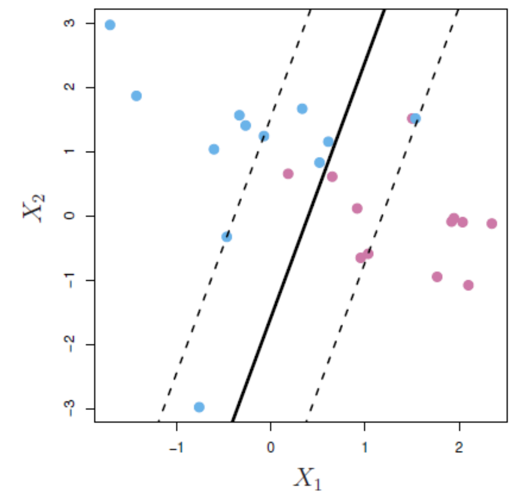
- Metod je neprimenljiv ako uzorci nisu linearno razdvojivi
- Položaj hiperravni odlučivanja je veoma osetljiv na položaj pojedinačnih uzoraka
  - Primera radi, uvođenje samo jednog novog uzorka može drastično promeniti procenjeni položaj hiperravni odlučivanja



  - Ovo ukazuje na to da je, u nekom smislu, nastupilo natprilagođenje (*overfitting*)
  - Generalno, dobijena procena, iako je možda dobro centrirana, ima izuzetno visoku varijansu
- Treba dozvoliti mogućnost da poneki uzorak za obuku ne bude dobro klasifikovan
  - Na račun toga bio bi dobijen klasifikator koji je mnogo robustniji na pojedinačne uzorke

Klasifikator blage margine

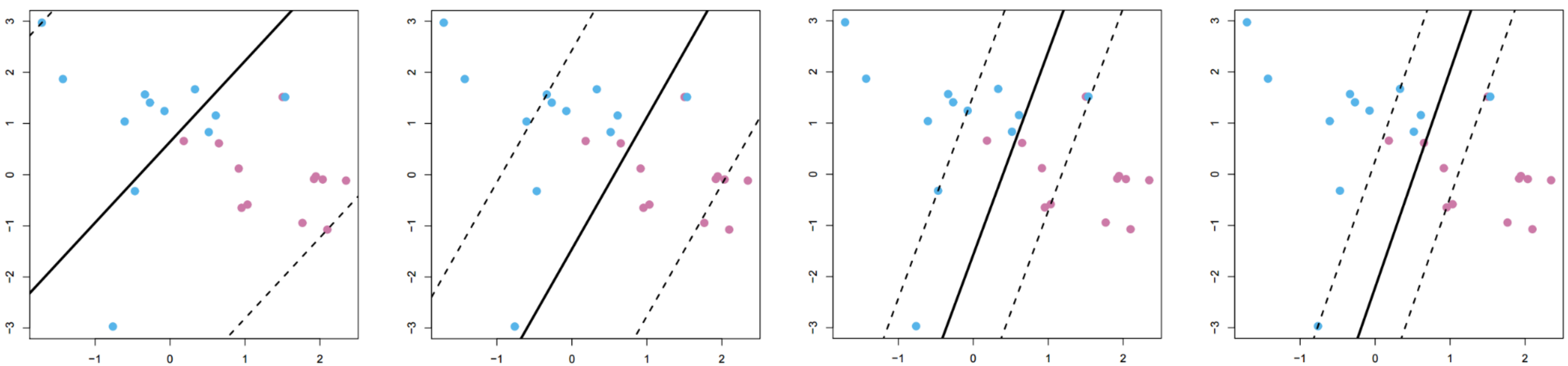
- Dozvoljava da poneki uzorak iz skupa za obuku bude sa pogrešne strane marginalne hiperravni, pa i sa pogrešne strane hiperravni odlučivanja
  - Ako uzorci nisu linearno razdvojivi, ovo poslednje je i neizbežno
- Ovako definisan klasifikator dobija se kao rešenje optimizacionog problema:
  - Maksimizovati M u odnosu na $\boldsymbol{\theta}$ i $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_N$ pod ograničenjima:
    - $\sum_{j=0}^{d} \vartheta_j^2 = 1$
    - $y_i \cdot (\vartheta_0 + \vartheta_1 x_1^{(i)} + \ldots + \vartheta_d x_d^{(i)}) \geq M(1 - \varepsilon_i)$ ($y_i = \pm 1$ zavisno od klase)
    - $\varepsilon_i > 0, \quad \sum_{i=1}^{N} \varepsilon_i \leq C$
  - Ovde su ε_i pomoćne promenljive, koje se odnose na to koliko koji uzorak zalazi na pogrešnu stranu marginalne hiperravni (ili, za $\varepsilon_i > 1$, čak i na pogrešnu stranu hiperravni odlučivanja), dok je C regularizacioni hiperparametar, koji se može posmatrati kao „budžet“ za ukupno odstupanje ε_i
    - Ovaj problem matematički nije ništa složeniji od prethodnog
    - C se u praksi određuje unakrsnom validacijom



Klasifikator blage margine

- Na položaj hiperravni odlučivanja sada utiču samo uzorci koji se nalaze na marginalnoj hiperravni ili s njene pogrešne strane (to su sada vektori nosači)
  - Metoda vektora nosača, za razliku od mnogih drugih, uopšte ne uzima u obzir položaje uzoraka koji su daleko od granice odlučivanja (a sa njene prave strane)
- Kao i u slučaju klasifikatora maksimalne margine, vektor **θ** dobija se kao linearna kombinacija vektora nosača, čiji broj s porastom *C* raste*

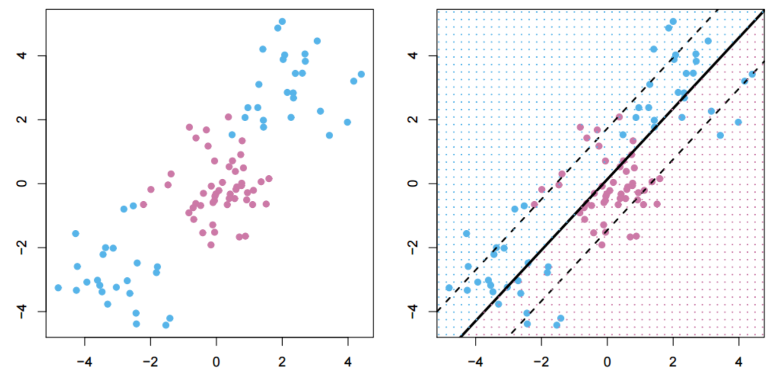
veliko *C* malo *C*



*u literaturi ima i drugačijih formulacija optimizacionog problema, kod kojih se regularizacioni parametar *C* uvodi na drugačiji način, takav da sa porastom *C* broj vektora nosača *opada*

Ograničenja klasifikatora blage margine

- Postoje situacije kada linearne granice odlučivanja, bez obzira na vrednost *C*, nisu odgovarajuće
- Moguće je proširiti prostor obeležja nelinearnim članovima, odnosno, umesto *d* obeležja $x_1, x_2, \ldots, x_d$ koristiti npr. 2*d* obeležja $x_1, x_2, \ldots, x_d$, $x_1^2, x_2^2, \ldots, x_d^2$



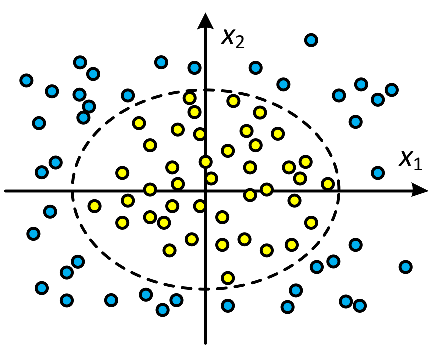
  - Ovako redefinisanom problemu odgovarale bi linearne granice odlučivanja u proširenom prostoru, ali nelinearne u originalnom, jer bi ograničenja bila:
    - $y_i \cdot (\vartheta_0 + \sum_{j=1}^{d} \vartheta_{j1} x_j^{(i)} + \sum_{j=1}^{d} \vartheta_{j2} (x_j^{(i)})^2) \geq M(1-\varepsilon_i)$
    - $\varepsilon_i > 0, \quad \sum_{i=1}^{N} \varepsilon_i \leq C, \quad \vartheta_0 + \sum_{j=1}^{d} (\vartheta_{j1}^2 + \vartheta_{j2}^2) = 1$
  - Ovo je i bilo potrebno postići, ali prelazak u višedimenzionalni prostor pogoršava problem manjka podataka, a povećava se i računska složenost
    - Zbog toga je potrebno dobro osmisliti način uvođenja ove nelinearnosti

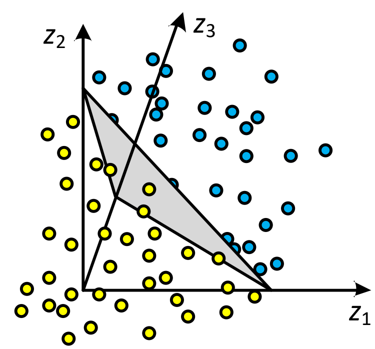
Mašine na bazi vektora nosača

- Problem konstrukcije nelinearnih granica odlučivanja rešava se u dve etape
  - Nelinearno preslikavanje uzoraka u višedimenzionalni prostor ($\mathbf{z} = \varphi(\mathbf{x})$)
  - Konstrukcija optimalne hiperravni odlučivanja u tom višedimenzionalnom prostoru
- Primer:
  - Slika levo ilustruje uzorke u 2D prostoru kod kojih nije moguće konstruisati hiperravan odlučivanja koja bi dala zadovoljavajuće rezultate
  - Uvođenjem preslikavanja $\mathbf{z} = \varphi(\mathbf{x})$ u 3D prostor prema sledećem pravilu:

$$\mathbf{x} = (x_1, x_2) \rightarrow \mathbf{z} = (z_1, z_2, z_3) = (x_1^2, \sqrt{2}x_1 x_2, x_2^2)$$

obezbeđuje se linearna razdvojivost u 3D prostoru (slika desno)





- Optimalnoj hiperravni odlučivanja u visokodimenzionalnom prostoru odgovara optimalna nelinearna granica odlučivanja u polaznom

Kernel funkcije

- Može se pokazati sledeće:
  - U postupku konstrukcije optimalne hiperravni odlučivanja pojedini uzorci figurišu samo u okviru skalarnih proizvoda njihovih parova $\langle \mathbf{x}_i, \mathbf{x}_j \rangle$
  - Pri klasifikaciji nepoznatog uzorka $\mathbf{x}'$ ni njegova obeležja u izračunavanju ne učestvuju kao pojedinačna, već samo u okviru skalarnog proizvoda $\langle \mathbf{x}', \mathbf{x}_i \rangle$ tog uzorka s uzorcima iz skupa za obuku (i to samo sa vektorima nosačima jer ostali ne utiču na hiperravan odlučivanja)
- Ako bi se svaki skalarni proizvod vektora $\mathbf{x}_i$ i $\mathbf{x}_j$ u ovim postupcima zamenio funkcijom $K(\mathbf{x}_i, \mathbf{x}_j)$, koja izračunava vrednost skalarnog proizvoda transformisanih uzoraka a da pri tome uopšte ne primenjuje eksplicitno preslikavanje, dobili bi se postupci koji obavljaju isti zadatak razdvajanja uzoraka, ali u visokodimenzionalnom prostoru u kom je granica odlučivanja linearna
  - Ovakvi klasifikatori nazivaju se mašinama na bazi vektora nosača (eng. *support vector machines* – SVM) i u zavisnosti od problema koriste neku od standardnih kernel funkcija

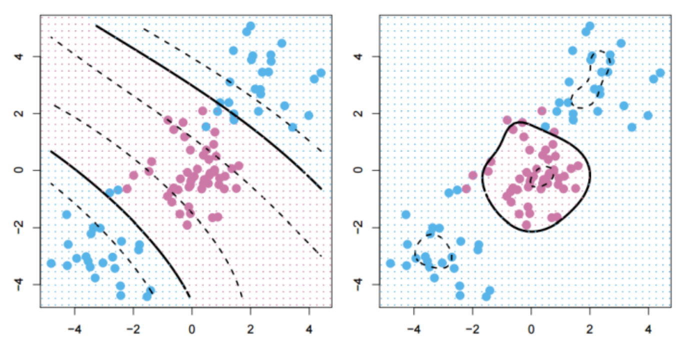
Polinomijalni kernel stepena $m > 0$

$$K(\mathbf{x}_i, \mathbf{x}_j) = \left(1 + \sum_{k=1}^{d} x_{ik} x_{jk}\right)^m$$

Radijalni kernel

$$K(\mathbf{x}_i, \mathbf{x}_j) = e^{-\gamma \sum_{k=1}^{d} (x_{ik} - x_{jk})^2}, \gamma > 0$$

Polinomijalni kernel Radijalni kernel



Klasifikacija u više od dve klase

- Kao i kod linearne klasifikacije, klasifikacija u više od dve klase (ω_1, ω_2,..., ω_K) uvek se može realizovati višestrukom primenom klasifikacije u dve klase, ali se to radije izbegava zbog problema regiona nedefinisane pripadnosti
  - **Jedan protiv svih** (eng. *one versus rest* – OVR)
    - Konstruiše se K SVM klasifikatora i svaki od njih se obučava za klasifikaciju u klase ω_i i „ne ω_i", i na kraju se uzorak $\mathbf{x}'$ dodeli onoj klasi ω_i koja maksimizuje vrednost izraza koji opisuje nivo pouzdanosti da joj uzorak $\mathbf{x}'$ pripada (kod linearnih SVM to je $\boldsymbol{\theta}_i^{\mathrm{T}}\mathbf{x}'$, a kod nelinearnih se računa na osnovu $K(\mathbf{x}',\mathbf{x}_i)$)
    - U ovom slučaju izbegnuta je višestruka primena binarne klasifikacije, pa i problem regiona nedefinisane pripadnosti
  - **Svaki protiv svakog** (eng. *one versus one* – OVO)
    - Konstruiše se $\binom{K}{2}$ SVM klasifikatora (što je razumno dogod K nije suviše velik broj), svaki od njih se obučava za klasifikaciju u određen par klasa ω_i i ω_j, i na kraju se uzorak $\mathbf{x}$ dodeli onoj klasi koja je najveći broj puta bila uspešnija u ovom nadmetanju po parovima
    - U ovom slučaju problem regiona nedefinisane pripadnosti i dalje postoji

Rezime

- Metoda vektora nosača prevazilazi određene probleme i ograničenja koji su postojali kod nekih drugih metoda
  - Daju dobre rezultate čak i kad je količina podataka relativno mala u odnosu na broj dimenzija, odnosno, obeležja
  - Rezultati su stabilni i ponovljivi (za razliku npr. od neuronskih mreža, gde ima slučajnosti u izboru težinskih koeficijenata)
  - Postoji jasna geometrijska interpretacija
  - Optimalno rešenje uključuje samo vektore nosače, tako da je *retko*
  - Optimalno rešenje traži se *analitički* i može se naći u polinomijalnom vremenu, pri čemu je prevaziđen problem lokalnih minimuma