Metoda podpornih vektorjev in jedrne funkcije

Ljupčo Todorovski

Univerza v Ljubljani, Fakulteta za upravo
Institut Jožef Stefan, Odsek za tehnologije znanja (E-8)

Marec 2020

Pregled predavanja

Metoda podpornih vektorjev

- Rob (*margin*) in širina roba
- Optimizacija roba in dualni problem
- Kontrola predsodka in variance

Jedrne (*kernel*) funkcije

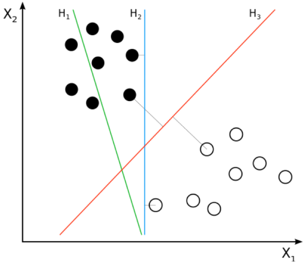
- Jedrni trik
- Izbira jedra

Klasifikacija kot optimizacijski problem

Linearni modeli za binarno klasifikacijo, $D_Y = \{-1, 1\}$

$$\hat{Y} = g(\beta^T \boldsymbol{x} + \beta_0),\ g(z) = \begin{cases} 1 & ; z \geq 0 \\ -1 & ; z < 0 \end{cases}$$

Vprašanje: kateri model je najboljši?



Odgovor metode podpornih vektorjev: model z **največjim robom**!?

Rob: grafična ponazoritev in odločitveno pravilo

Odločitveno pravilo

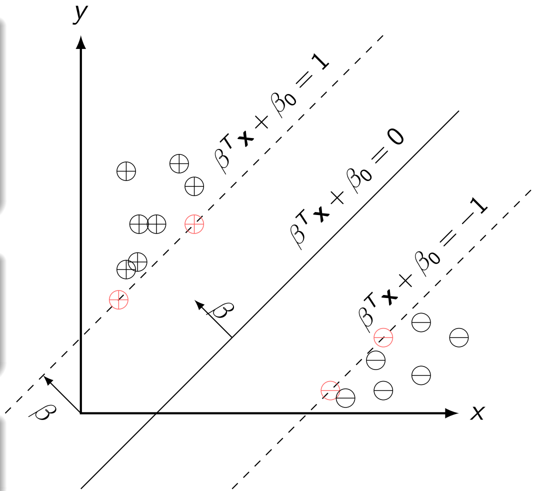
- Za pozitivni primer $\mathbf{x}_\oplus$: $\beta^T \mathbf{x}_\oplus + \beta_0 \geq 1$
- Za negativni primer $\mathbf{x}_\ominus$: $\beta^T \mathbf{x}_\ominus + \beta_0 \leq -1$

Vpeljemo y

- $y = 1$ za primere $\oplus$
- $y = -1$ za primere $\ominus$

$(\beta^T \mathbf{x} + \beta_0)\, y - 1 \geq 0$

Enačaj velja točno na robu!



Rob: grafična ponazoritev in širina roba

Širina roba

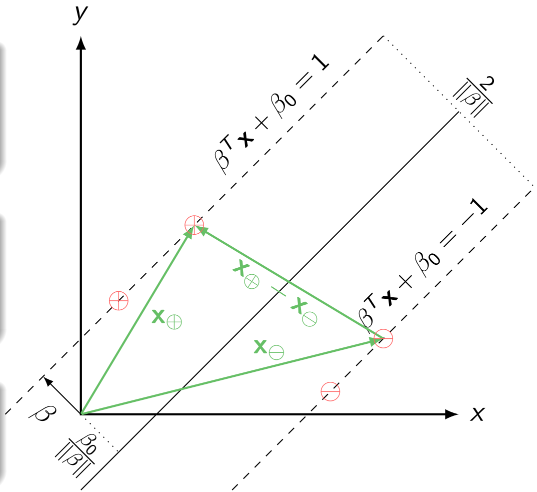
$$\frac{\beta^T(\mathbf{x}_\oplus - \mathbf{x}_\ominus)}{\|\beta\|}$$

Prejšnja prosojnica

- $\beta^T\mathbf{x}_\oplus + \beta_0 = 1$
- $\beta^T\mathbf{x}_\ominus + \beta_0 = -1$

Če odštejemo ti dve enačbi

$$\beta^T(\mathbf{x}_\oplus - \mathbf{x}_\ominus) = 2$$



Osnovna formulacija in konveksna ciljna funkcija

Osnovni optimizacijski problem

$$\max_{\beta,\beta_0} \frac{2}{\|\beta\|}$$

$|S|$ omejitev

- $(\beta^T \boldsymbol{x} + \beta_0)\, y - 1 \geq 0,\ (\boldsymbol{x}, y) \in S$

Kako do konveksne ciljne funkcije?

- Upoštevamo $\max 1/\|\beta\| = 1/\min \|\beta\|$
- Upoštevamo da sta $\min \|\beta\|$ in $\min \frac{1}{2}\|\beta\|^2$ ekvivalentni

Končna formulacija za binarno klasifikacijo

$$\min_{\beta,\beta_0} \frac{1}{2}\|\beta\|^2$$

$|S|$ omejitev

- $1 - (\beta^T \boldsymbol{x} + \beta_0)\, y \leq 0,\ (\boldsymbol{x}, y) \in S$

Naloga kvadratnega programiranja

- Kvadratna in konveksna ciljna funkcija
- Linearne omejitve
- Možno najti globalni minimum

Dualni problem

Lagrange-ova funkcija s koeficienti $\alpha_{\boldsymbol{x}}$, $(\boldsymbol{x}, y) \in S$

$$\mathcal{L}(\beta, \beta_0, \alpha) = \frac{1}{2}\|\beta\|^2 + \sum_{(\boldsymbol{x},y)\in S} \alpha_{\boldsymbol{x}} \left(1 - (\beta^T \boldsymbol{x} + \beta_0)\, y\right)$$

Karush-Kuhn-Tucker-jevi (KKT) pogoji za optimalno rešitev $\beta^*, \beta_0^*, \alpha^*$

$$\begin{aligned}
\frac{\partial \mathcal{L}(\beta^*, \beta_0^*, \alpha^*)}{\partial \beta} = \frac{\partial \mathcal{L}(\beta^*, \beta_0^*, \alpha^*)}{\partial \beta_0} &= 0 \\
\alpha_{\boldsymbol{x}}^* \left(1 - ({\beta^*}^T \boldsymbol{x} + \beta_0^*)\, y\right) &= 0 \\
1 - ({\beta^*}^T \boldsymbol{x} + \beta_0^*)\, y &\leq 0 \\
\alpha_{\boldsymbol{x}}^* &\geq 0
\end{aligned}$$

Reševanje dualnega problema

Prvi KKT pogoj, upoštevaje $\|\beta\|^2 = \beta^T\beta$ oziroma $\frac{\partial\|\beta\|^2}{\partial\beta} = 2\,\beta$

$$
\begin{aligned}
\frac{\partial\mathcal{L}}{\partial\beta} &= \beta - \sum_{(\boldsymbol{x},y)\in S} \alpha_{\boldsymbol{x}} y \boldsymbol{x} = 0 \\
\frac{\partial\mathcal{L}}{\partial\beta_0} &= -\sum_{(\boldsymbol{x},y)\in S} \alpha_{\boldsymbol{x}} y = 0
\end{aligned}
$$

Je izpolnjen pri

- $\beta = \sum_{(\boldsymbol{x},y)\in S} \alpha_{\boldsymbol{x}} y \boldsymbol{x}$
- $\sum_{(\boldsymbol{x},y)\in S} \alpha_{\boldsymbol{x}} y = 0$

Če upoštevamo $\beta = \sum_{(\boldsymbol{x},y)\in S} \alpha_{\boldsymbol{x}} y \boldsymbol{x}$

$\langle u, v\rangle = u^T v = v^T u$ je oznaka za skalarni produkt vektorjev u in v

$$
\begin{aligned}
\frac{1}{2}\|\beta\|^2 &= \frac{1}{2}\langle \beta, \beta\rangle \\
&= \frac{1}{2} \sum_{(\boldsymbol{x_1},y_1)\in S} \sum_{(\boldsymbol{x_2},y_2)\in S} \alpha_{\boldsymbol{x_1}} \alpha_{\boldsymbol{x_2}} y_1 y_2 \langle \boldsymbol{x_1}, \boldsymbol{x_2}\rangle \\
\beta^T \boldsymbol{x} &= \langle \beta, \boldsymbol{x}\rangle \\
&= \sum_{(\boldsymbol{x_2},y_2)\in S} \alpha_{\boldsymbol{x_2}} y_2 \langle \boldsymbol{x}, \boldsymbol{x_2}\rangle
\end{aligned}
$$

Če upoštevamo $\beta = \sum_{(\boldsymbol{x},y)\in S} \alpha_{\boldsymbol{x}} y \boldsymbol{x}$

$$
\begin{aligned}
\sum_{(\boldsymbol{x},y)\in S} \alpha_{\boldsymbol{x}} \left(1 - y\left(\beta^T \boldsymbol{x} + \beta_0\right)\right) &= \sum_{(\boldsymbol{x},y)\in S} \alpha_{\boldsymbol{x}} \\
&- \sum_{(\boldsymbol{x},y)\in S} \sum_{(\boldsymbol{x_2},y_2)\in S} \alpha_{\boldsymbol{x}} \alpha_{\boldsymbol{x_2}}\, y\, y_2 \left\langle \boldsymbol{x}, \boldsymbol{x_2} \right\rangle \\
&- \beta_0 \sum_{(\boldsymbol{x},y)\in S} \alpha_{\boldsymbol{x}} y
\end{aligned}
$$

Če upoštevamo zadnje dve prosojnici ter $\sum_{(\boldsymbol{x},y)\in S} \alpha_{\boldsymbol{x}} y = 0$

$$\mathcal{L}(\beta, \beta_0, \alpha) = \sum_{(\boldsymbol{x},y)\in S} \alpha_{\boldsymbol{x}} - \frac{1}{2} \sum_{(\boldsymbol{x_1},y_1)\in S} \sum_{(\boldsymbol{x_2},y_2)\in S} \alpha_{\boldsymbol{x_1}} \alpha_{\boldsymbol{x_2}} \, y_1 \, y_2 \, \langle \boldsymbol{x_1}, \boldsymbol{x_2} \rangle$$

Končna formulacija dualnega problema

$$\max_{\alpha} \sum_{(\boldsymbol{x},y)\in S} \alpha_{\boldsymbol{x}} - \frac{1}{2} \sum_{(\boldsymbol{x_1},y_1)\in S} \sum_{(\boldsymbol{x_2},y_2)\in S} \alpha_{\boldsymbol{x_1}} \alpha_{\boldsymbol{x_2}} \, y_1 \, y_2 \, \langle \boldsymbol{x_1}, \boldsymbol{x_2} \rangle$$

$|S| + 1$ linearnih omejitev iz tretjega in četrtega KKT pogoja

- $\alpha_{\boldsymbol{x}} \geq 0$, $(\boldsymbol{x}, y) \in S$
- $\sum_{(\boldsymbol{x},y)\in S} \alpha_{\boldsymbol{x}} y = 0$

Pozor: vse operacije v prostoru X so skalarni produkti.

Napoved in podporni vektorji

Napoved modela y_0 za primer $\boldsymbol{x_0}$

$$
\begin{aligned}
\hat{y}_0 &= g\left(\left\langle \sum_{(\boldsymbol{x},y)\in S} \alpha_{\boldsymbol{x}}\, y\, \boldsymbol{x}, \boldsymbol{x_0} \right\rangle + \beta_0\right) \\
&= g\left(\sum_{(\boldsymbol{x},y)\in S} \alpha_{\boldsymbol{x}}\, y\, \langle \boldsymbol{x}, \boldsymbol{x_0} \rangle + \beta_0\right)
\end{aligned}
$$

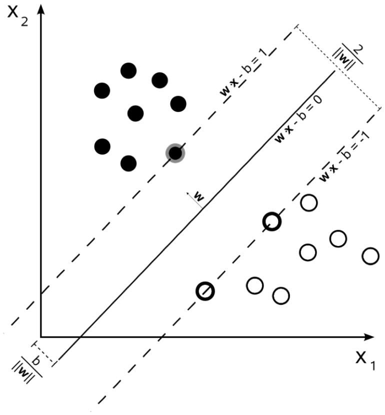
Vsoto v zgornji formuli računamo le za primere $\boldsymbol{x}$, kjer $\alpha_{\boldsymbol{x}} > 0$.

Podporni vektorji $\boldsymbol{x_s}$

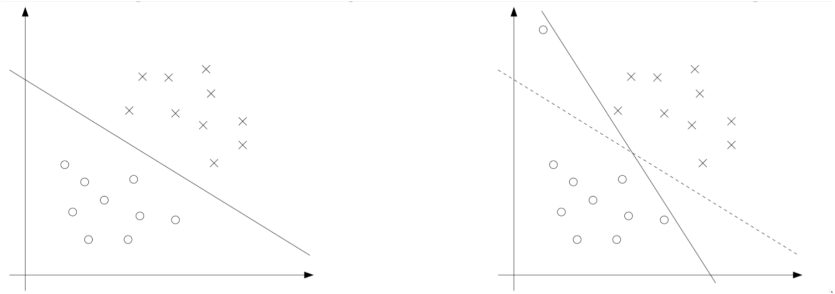
- Pri njih velja $\alpha_{\boldsymbol{x}} > 0$
- Zaradi drugega KKT pogoja velja $1 - y\,(\beta^T \boldsymbol{x_s} + \beta_0) = 0$
- Torej $\boldsymbol{x_s}$ ležijo na robu

Podporni vektorji in rob

Na sliki je w oznaka za β in b oznaka za β_0



Visoka varianca in občutljivost na šumne podatke



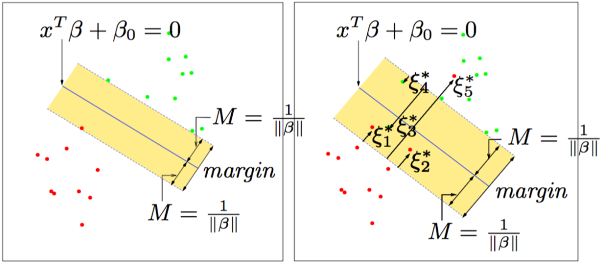
In kaj lahko naredimo, če razreda nista linearno ločljiva?

Alternativna formulacija za podatke, ki niso linearno ločljivi

$$\min_{\beta,\beta_0} \frac{1}{2}\|\beta\|^2 + C \sum_{(\boldsymbol{x},y)\in S} \xi_{\boldsymbol{x}}$$

$2\,|S|$ linearnih omejitev:

- $y\left(\beta^T \boldsymbol{x_s} + \beta_0\right) \geq 1 - \xi_{\boldsymbol{x}}$, $(\boldsymbol{x}, y) \in S$
- $\xi_{\boldsymbol{x}} \geq 0$, $(\boldsymbol{x}, y) \in S$



Dualni problem in podporni vektorji

$$\max_{\alpha} \sum_{(\mathbf{x},y)\in S} \alpha_{\mathbf{x}} - \frac{1}{2} \sum_{(\mathbf{x_1},y_1)\in S} \sum_{(\mathbf{x_2},y_2)\in S} \alpha_{\mathbf{x_1}} \alpha_{\mathbf{x_2}} y_1 y_2 \langle \mathbf{x_1}, \mathbf{x_2} \rangle$$

$2|S| + 1$ linearnih omejitev

- $0 \leq \alpha_{\mathbf{x}} \leq C$, $(\mathbf{x}, y) \in S$
- $\sum_{(\mathbf{x},y)\in S} \alpha_{\mathbf{x}} y = 0$

Podporni vektorji so lahko tudi znotraj roba

- Na robu $0 < \alpha_{\mathbf{x_s}} < C$: $y\,(\beta^T \mathbf{x_s} + \beta_0) = 1$
- Znotraj roba $\alpha_{\mathbf{x_s}} = C$: $y\,(\beta^T \mathbf{x_s} + \beta_0) < 1$

Vloga parametra C (cena): predsodek in varianca

$\min_{\beta,\beta_0} \frac{1}{2}\|\beta\|^2 + C\sum_{(\boldsymbol{x},y)\in S}\xi_{\boldsymbol{x}}$

Večanje vrednosti C

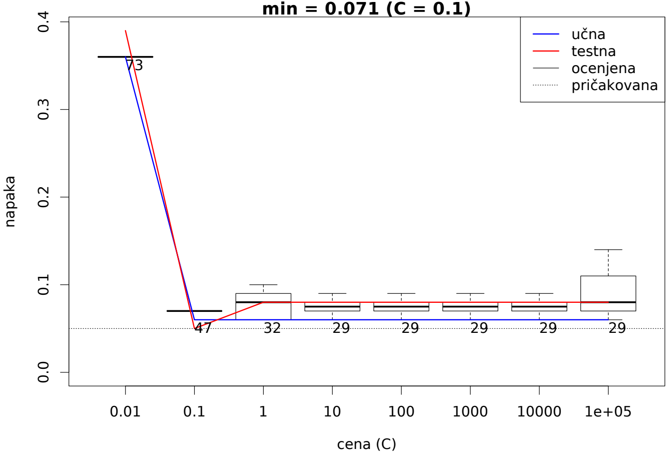
- Fokus optimizacije se prenese na drugi člen formule zgoraj, t.j., na zmanjševanju odmikov $\xi_{\boldsymbol{x}}$
- Fokus torej na čim boljšem ločevanju razredov
- Potegne model v smer višje variance in nižjega predsodka
- Podobno kot malo število sosedov k pri metodi najbližjih sosedov

Manjšanje vrednosti C

- $\xi_{\boldsymbol{x}}$ so lahko veliki, zato ločevanje razredov ni tako pomembno
- Zniža se varianca in poveča predsodek
- Pozor: **poveča se tudi število podpornih vektorjev znotraj roba**

Klasifikacija s podpornimi vektorji: cena C

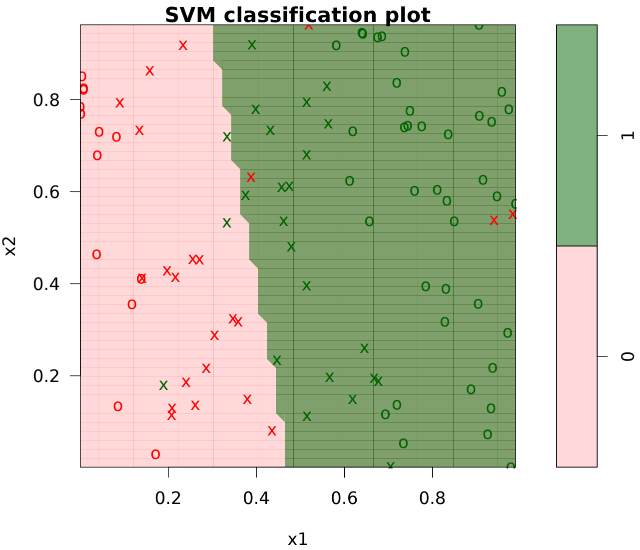
Številke na grafih: število podpornih vektorjev



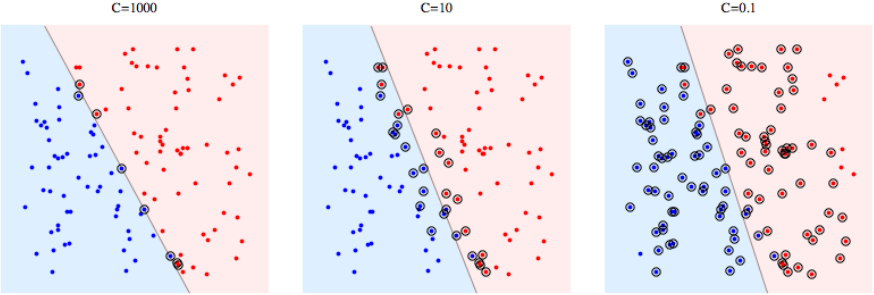
$Y = I((1 + X_1 + X_1 X_2)/3 \geq 0.5), D_i = [0, 1], i = 1..2, D_Y = \{0, 1\}$

zamenjamo vrednost Y 0.05 naključno izbranim primerom

Klasifikacije s podpornimi vektorji: $C = 0.1$



Klasifikacije s podpornimi vektorji: vpliv vrednosti C



Alternativni metodi za več kot dva razreda, $|D_Y| > 2$

Metoda ena-na-ena, $|D_Y|(|D_Y| - 1)/2$ modelov

- Po en za vsako podmnožico $\{v_1, v_2\} \subset D_Y$
  - Izberemo primere $S' = \{(\boldsymbol{x}, y) \in S : y \in \{v_1, v_2\}\}$
  - Zgradimo model iz S', kjer v_1 spremenimo v 1, v_2 v -1
- Iz napovedi veh modelov izberemo večinsko napoved

Metoda en-proti-vsem, $|D_Y|$ modelov

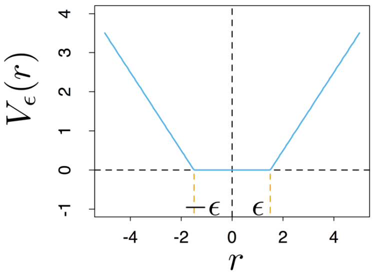
- Po en za vsako vrednost $v \in D_Y$
- Zgradimo model iz S, kjer v spremenimo v 1, druge vrednosti Y v -1
- Napovemo razred, katerega napoved je najdlje od roba

Regresija, $D_Y \subset \mathbb{R}$

$$\min_{\beta,\beta_0} \frac{C}{2}\|\beta\|^2 + \sum_{(\boldsymbol{x},y)\in S} V_\epsilon(y - (\beta^T \boldsymbol{x} + \beta_0))$$

Funkcija V_ϵ, $\epsilon \in \mathbb{R}^+$

$$V_\epsilon(r) = \begin{cases} 0 & ; |r| \leq \epsilon \\ |r| - \epsilon & ; |r| > \epsilon \end{cases}$$



Regresija: dualni problem

$$\min_{\alpha,\alpha^*} \quad \epsilon \sum_{(\boldsymbol{x},y)\in S} (\alpha^*_{\boldsymbol{x}} + \alpha_{\boldsymbol{x}}) - \sum_{(\boldsymbol{x},y)\in S} y(\alpha^*_{\boldsymbol{x}} - \alpha_{\boldsymbol{x}})$$
$$+ \quad \frac{1}{2} \sum_{(\boldsymbol{x}_1,y_1)\in S} \sum_{(\boldsymbol{x}_2,y_2)\in S} (\alpha^*_{\boldsymbol{x}_1} - \alpha_{\boldsymbol{x}_1})(\alpha^*_{\boldsymbol{x}_2} - \alpha_{\boldsymbol{x}_2}) \langle \boldsymbol{x}_1, \boldsymbol{x}_2 \rangle$$

$3|S| + 1$ linearnih omejitev

- $0 \leq \alpha_{\boldsymbol{x}} \leq C$, $(\boldsymbol{x}, y) \in S$
- $0 \leq \alpha^*_{\boldsymbol{x}} \leq C$, $(\boldsymbol{x}, y) \in S$
- $\alpha_{\boldsymbol{x}} \alpha^*_{\boldsymbol{x}} = 0$, $(\boldsymbol{x}, y) \in S$
- $\sum_{(\boldsymbol{x},y)\in S} (\alpha^*_{\boldsymbol{x}} - \alpha_{\boldsymbol{x}}) = 0$

Napoved regresijskega modela v točki $\mathbf{x_0}$

$$\hat{y}_0 = \sum_{(\mathbf{x},y)\in S} (\alpha^*_{\mathbf{x}} - \alpha_{\mathbf{x}}) \langle \mathbf{x}, \mathbf{x_0} \rangle + \beta_0$$

- Podporni vektorji so učni primeri $(\mathbf{x}, y) \in S$: $\alpha^*_{\mathbf{x}} \neq \alpha_{\mathbf{x}}$
- Vse operacije v prostoru X so še vedno skalarni produkti

Vloga parametrov cena (C) in ϵ

$$\min_{\beta,\beta_0} \frac{C}{2}\|\beta\|^2 + \sum_{(\mathbf{x},y)\in S} V_\epsilon(y - (\beta^T \mathbf{x} + \beta_0))$$

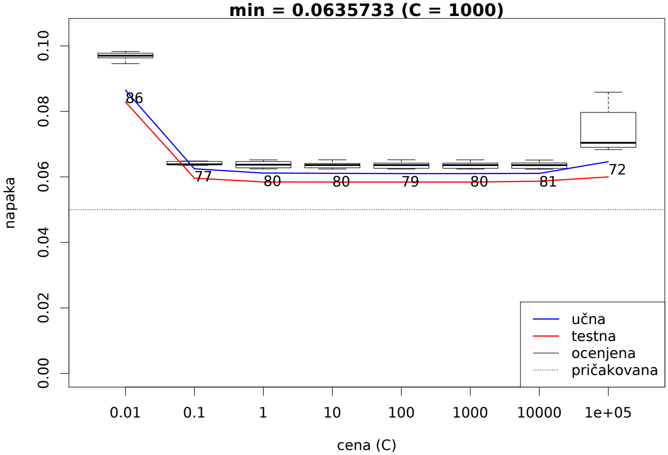
Parameter cena uravnava kompromis med predsodkom in varianco

- **Obratno** kot pri klasifikaciji
- Z večanjem C se optimizacija osredotoča na male vrednosti β (regularizacija), kar zmanjša varianco in poveča predsodek
- Večji predsodek, manj kompleksni modeli in **praviloma** manj podpornih vektorjev (obvezno opazovanje njihovega števila)

Parameter ϵ

- Vpliv nepredvidljiv, zato običajno privzeta vrednost $\epsilon = 0.1$
- Vrednosti spremenljivk X in Y običajno standardiziramo

Regresija s podpornimi vektorji: cena C

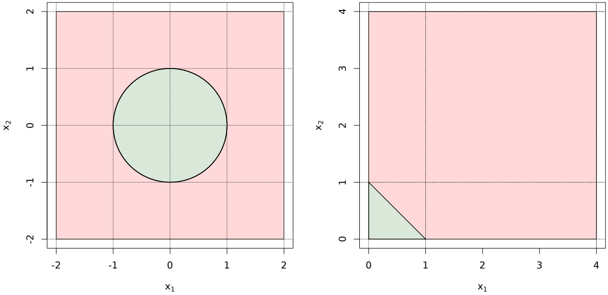


$$Y = (1 + X_1 + X_1 X_2)/3 + \mathcal{N}(0, 0.05),\ D_i = [0,1], i = 1..2, D_Y = [0,1]$$

Transformacija ϕ prostora napovednih spremenljivk X

$$X' = \phi(X)$$

Primer kvadratne transformacije $\phi(\{X_1, X_2\}) = \{X_1^2, X_2^2\}$



Nelinearna odločitvena meja postane linearna.

Definicija in pomen

Definicija jedrne funkcije $K : \bigtimes_{i=1}^{p} D_i \times \bigtimes_{i=1}^{p} D_i \to \mathbb{R}_0^+$

$$K(u, v) = \langle \phi(u), \phi(v) \rangle$$

Vrne vrednost skalarnega produkta transformiranih vektorjev.

Primer s prejšnje prosojnice

$$K(u, v) = \left\langle ({u_1}^2, {u_2}^2), ({v_1}^2, {v_2}^2) \right\rangle = u_1^2 v_1^2 + u_2^2 v_2^2$$

Jedrni trik

Z jedrno funkcijo skalarni produkt vektorjev v transformiranem prostoru izračunamo brez potrebe prehoda v transformirani prostor.

Primer jedrne funkcije $K(u, v) = (\langle u, v\rangle + 1)^2$

$$
\begin{aligned}
(\langle u, v\rangle + 1)^2 &= (u_1 v_1 + u_2 v_2 + 1)^2 \\
&= u_1^2 v_1^2 + u_2^2 v_2^2 + 1 + 2u_1 u_2 v_1 v_2 + 2u_1 v_1 + 2u_2 v_2 \\
&= \langle(\phi(u), \phi(v)\rangle
\end{aligned}
$$

$$
\phi(u) = (u_1^2, u_2^2, 1, \sqrt{2}u_1 u_2, \sqrt{2}u_1, \sqrt{2}u_2)
$$

S pomočjo jedrnega trika lahko izračunamo model podpornih vektorjev v 6-dimenzionalnem prostoru, čeprav je X dvodimenzionalen.

Pogosto uporabljene jedrne funkcije (jedra)

Polinomsko jedro, stopnja d

$$K(u, v) = (1 + \langle u, v \rangle)^d$$

Dimenzija transformiranega prostora $O(p^d)$.

Gaussovo (radialno) jedro $\gamma = 1/(2\sigma^2)$

$$K(u, v) = e^{-\gamma \|u - v\|^2}$$

Sorodnika: eksponentno in Lagrangeovo jedro. Dimenzija ∞.

Sigmoidno (nevronsko) jedro, κ_1, κ_2

$$K(u, v) = \tanh(\kappa_1 \langle u, v \rangle + \kappa_2)$$

Dimenzija transformiranega prostora ∞.

Jedrna matrika in Mercerjev izrek

Jedrna matrika K za podatkovno množico S

- Dimenzije matrike $|S| \times |S|$
- Elementi $K_{ij} = K(x_i, x_j) = \langle \phi(x_i), \phi(x_j) \rangle, x_i, x_j \in S$
- Zaradi komutativnosti skalarnega produkta je K simetrična

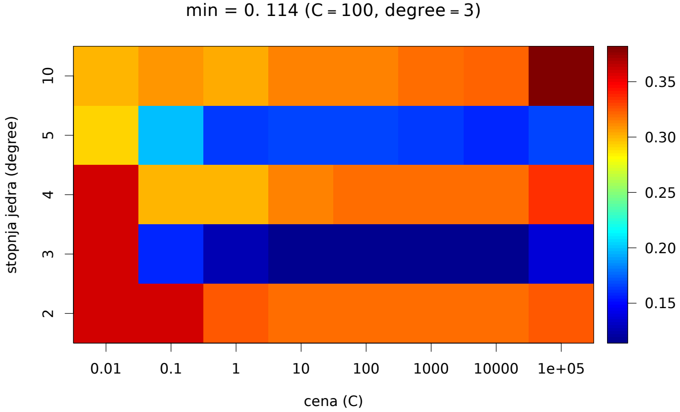
Mercerjev izrek

Nujni in zadostni pogoj zato, da matrika K določa veljavno jedro, je to da je K simetrična in pozitivno semi-definitna, t.j., $\forall z \in R^{|S|} : z^T K z \geq 0$.

Posledici

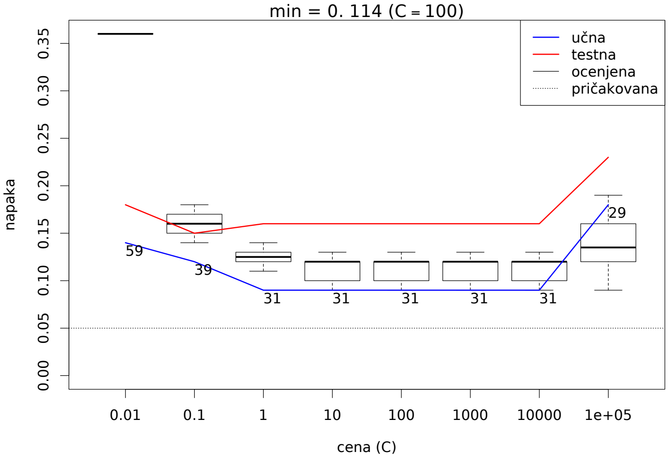
- K lahko obravnavamo kot matriko podobnosti med primeri
- Prostor možnih jeder zelo velik

Klasifikacija s polinomskim jedrom: *C* in *degree*

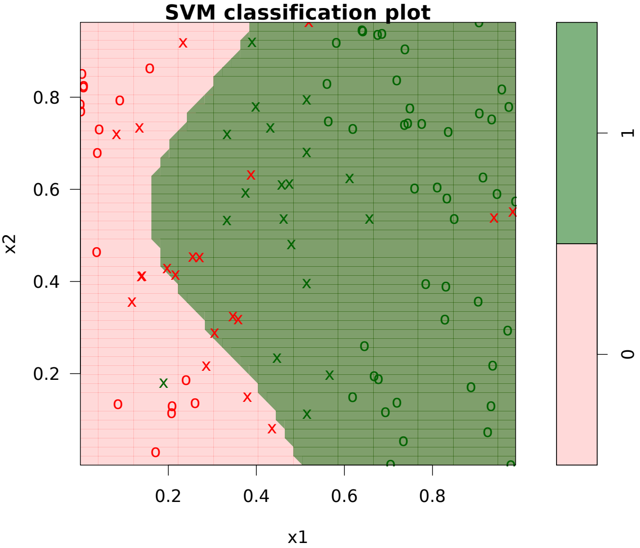


$Y = I((1 + X_1 + X_1X_2)/3 \geq 0.5), D_i = [0,1], i = 1..2, D_Y = \{0,1\}$
zamenjamo vrednost Y 0.05 naključno izbranim primerom

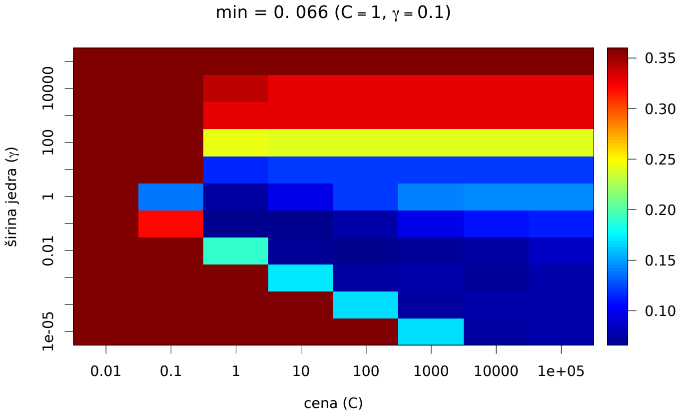
Klasifikacija s polinomskim jedrom: *degree* = 3



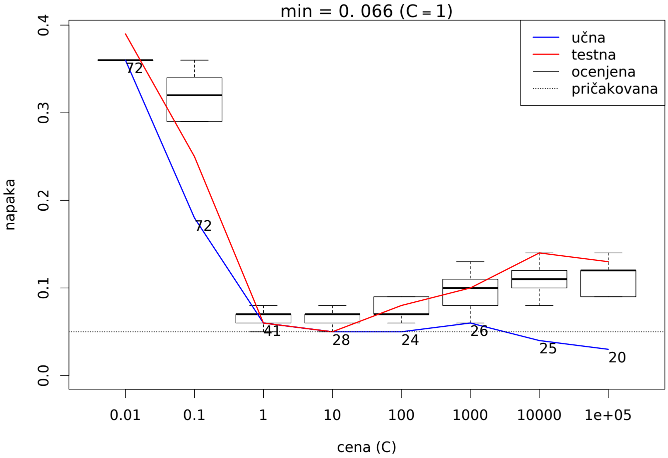
Klasifikacije s polinomskim jedrom: $C = 100$, $degree = 3$



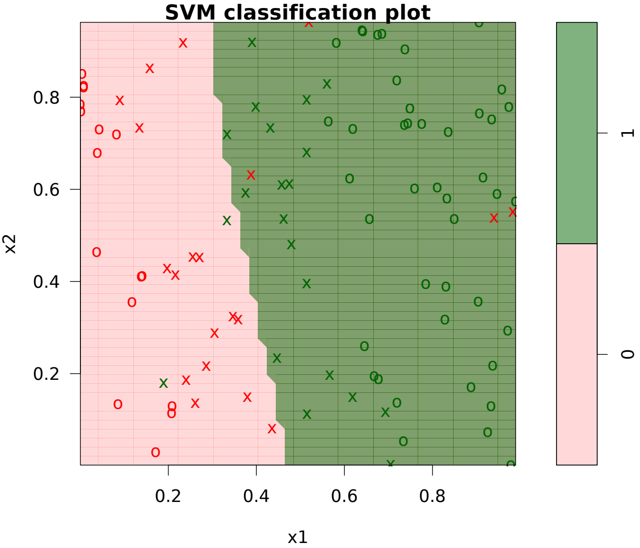
Klasifikacija z Gaussovim jedrom: C in γ



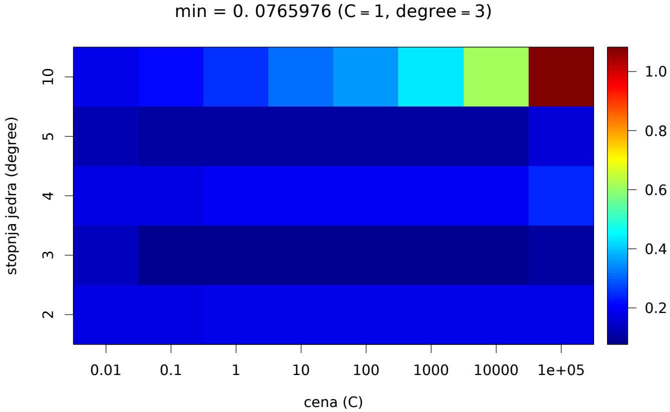
Klasifikacija z Gaussovim jedrom: $\gamma = 0.1$



Klasifikacije s Gaussovim jedrom: $C = 1$, $\gamma = 0.1$

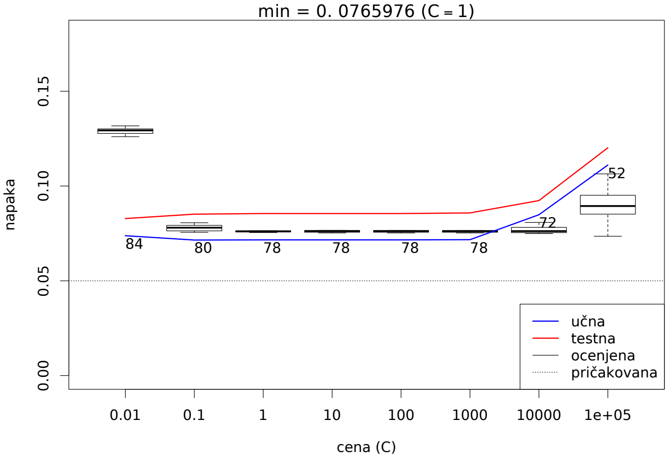


Regresija s polinomskim jedrom: *C* in *degree*

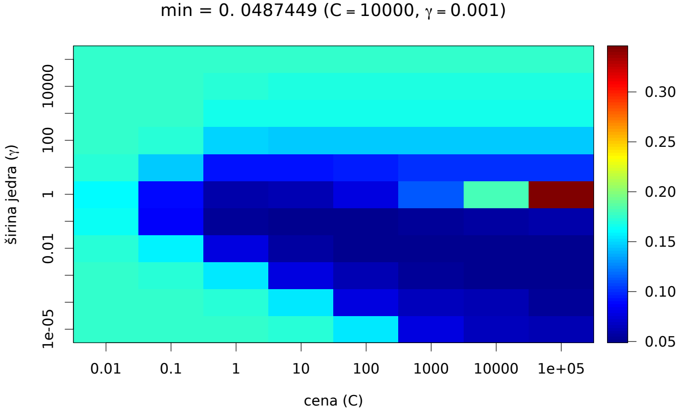


$$Y = (1 + X_1 + X_1 X_2)/3 + \mathcal{N}(0, 0.05),\ D_i = [0, 1], i = 1..2, D_Y = [0, 1]$$

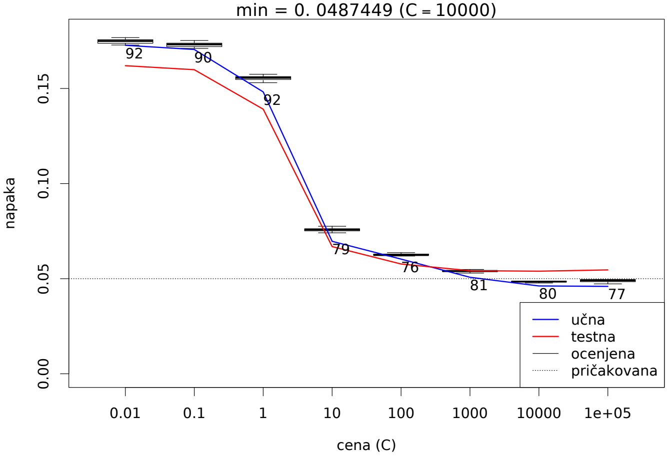
Regresija s polinomskim jedrom: *degree* = 3



Regresija z Gaussovim jedrom: C in γ



Regresija z Gaussovim jedrom: $\gamma = 0.1$



Primerjava napak

Klasifikacija, klasifikacijska napaka

- Linearno jedro, $C = 0.1$: 0.071
- Polinomsko jedro, $degree = 3$, $C = 100$: 0.114
- Gaussovo jedro, $\gamma = 0.1$, $C = 1$: **0.066**
- Drevesa 0.094, najbližji sosedi 0.065

Regresija, celotna napaka *RMSE*

- Linearno jedro, $C = 1000$: 0.0636
- Polinomsko jedro, $degree = 3$, $C = 1$: 0.0766
- Gaussovo jedro, $\gamma = 0.001$, $C = 10,000$: **0.0487**
- Drevesa 0.0679, najbližji sosedi 0.0548

Znani algoritmi in implementacije

Originalni predlog (Vapnik in Chervonenkis 1963)

Optimizacija roba za linearno ločljiva razreda.

Nadgradnja in posplošitev (Cortes in Vapnik 1995)

Knjižnica LIBSVM (C++ in Java), na voljo tudi v R (ovojnica *e1071*).