Merjenje napovedne napake in izbira napovednega modela

Ljupčo Todorovski
Univerza v Ljubljani, Fakulteta za upravo

Marec 2018

Pregled predavanja

Predsodek (*bias*) in varianca

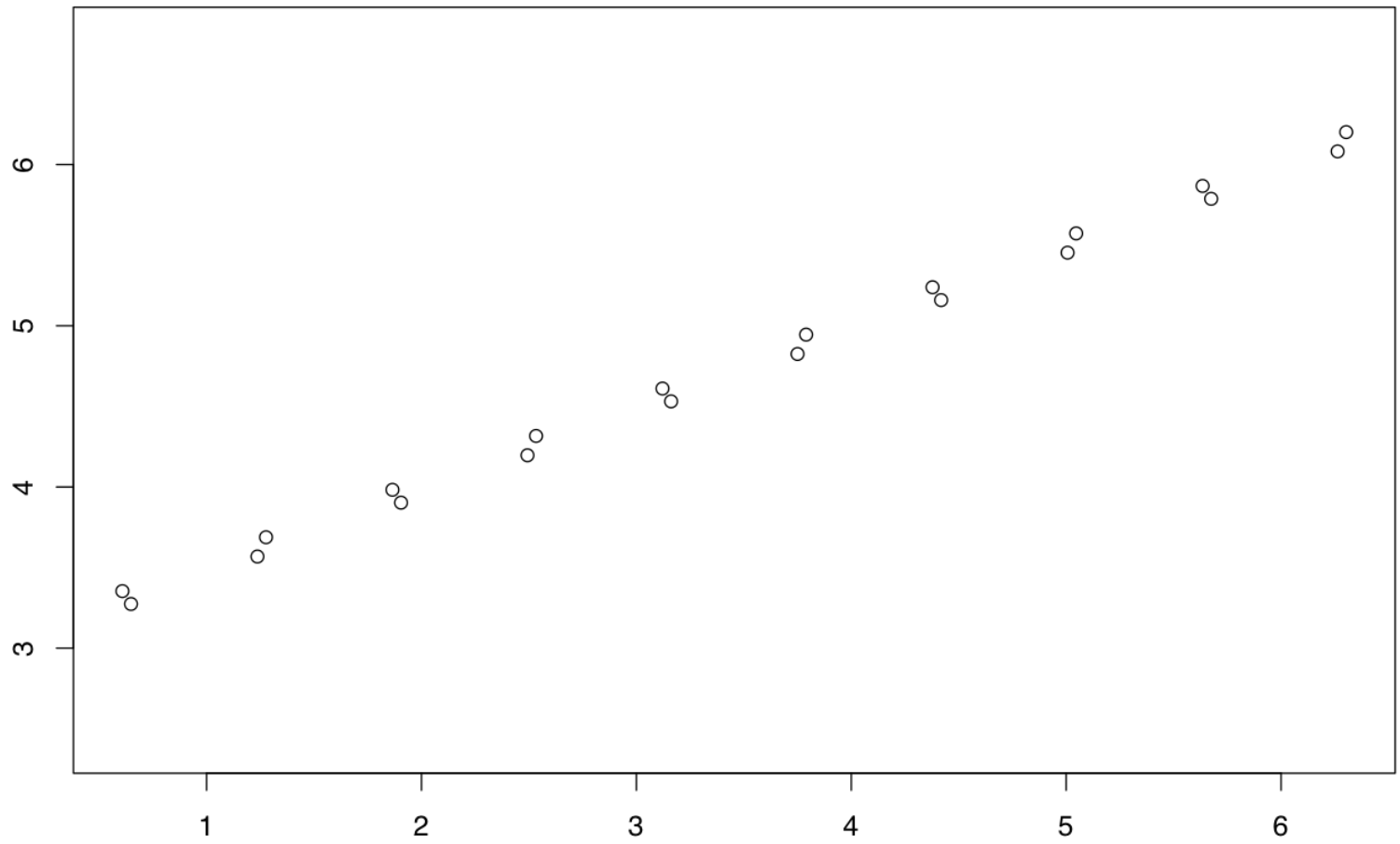
- Ilustrativni primer
- Dekompozicija napovedne napake na predsodek in varianco

Merjenje napovedne napake

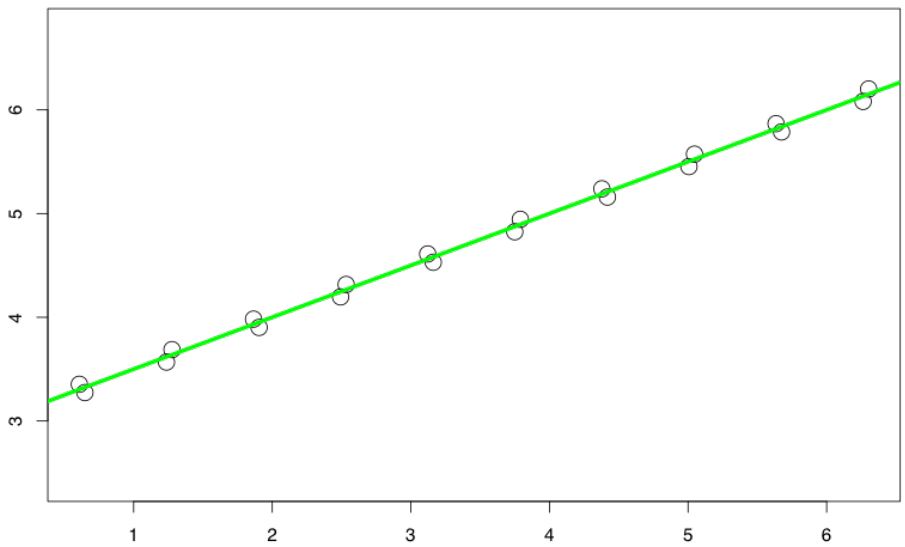
- Optimizem ocene napake na učni množici
- Razbitje podatkovne množice na učno in testno
- Vzorčenje podatkovne množice

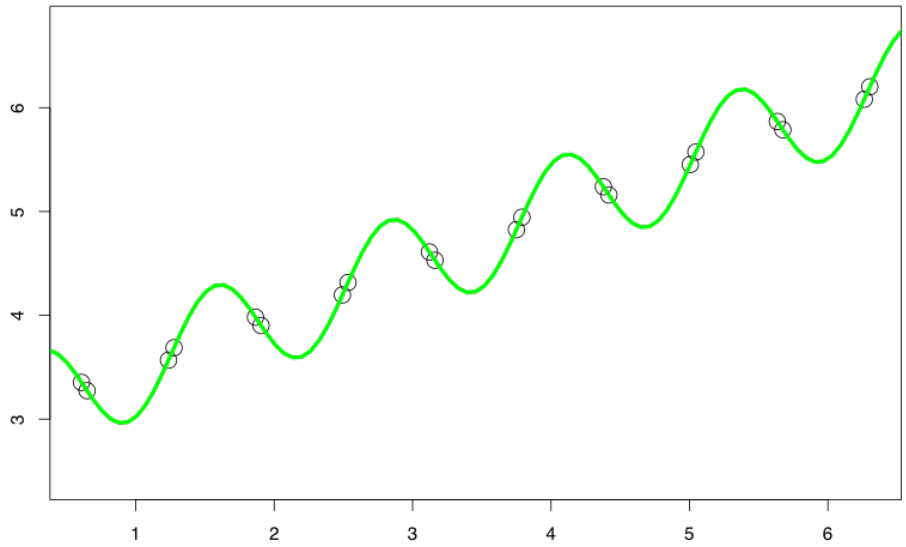
Izbira napovednega modela

Podatkovna množica: $p = 1, D_1, D_Y \subseteq \mathbb{R}$

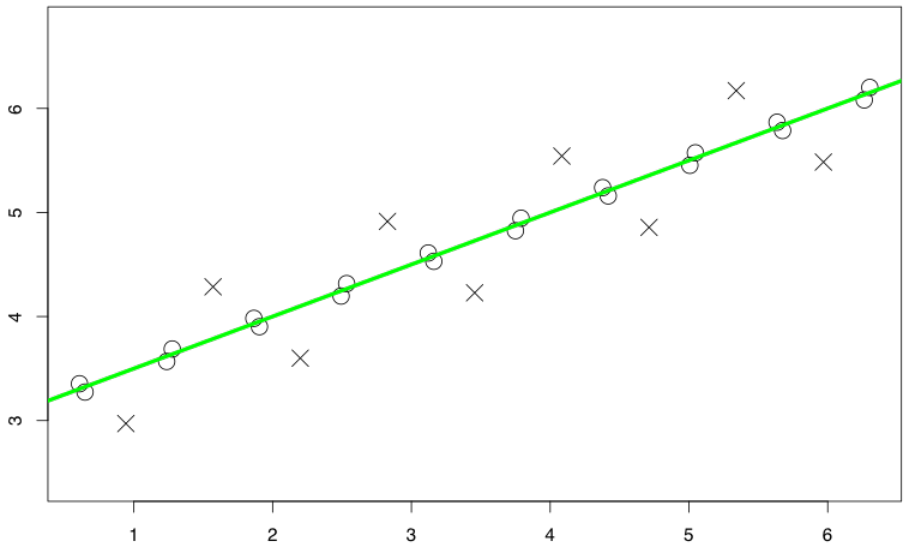


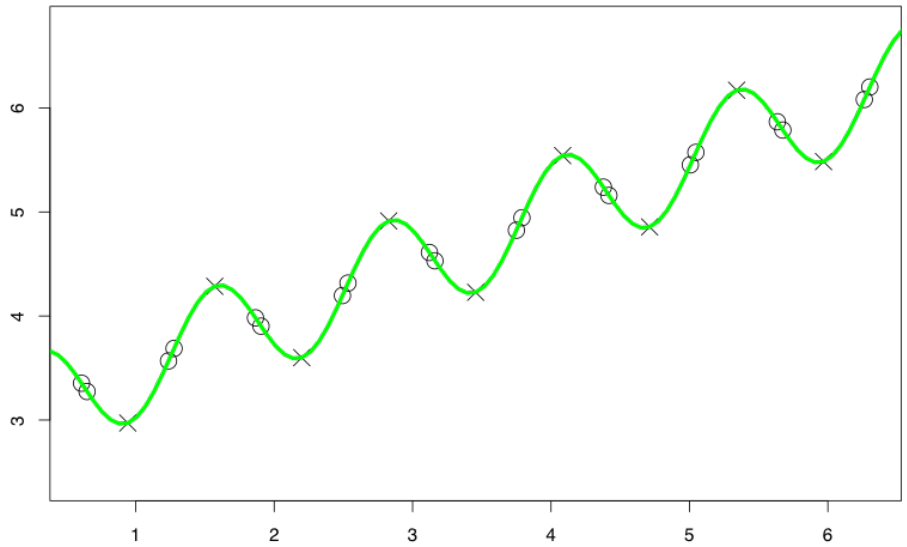
Izbira modela: linearni (William of Ockham)





Izbira modela: oscilatorski (testni podatki)

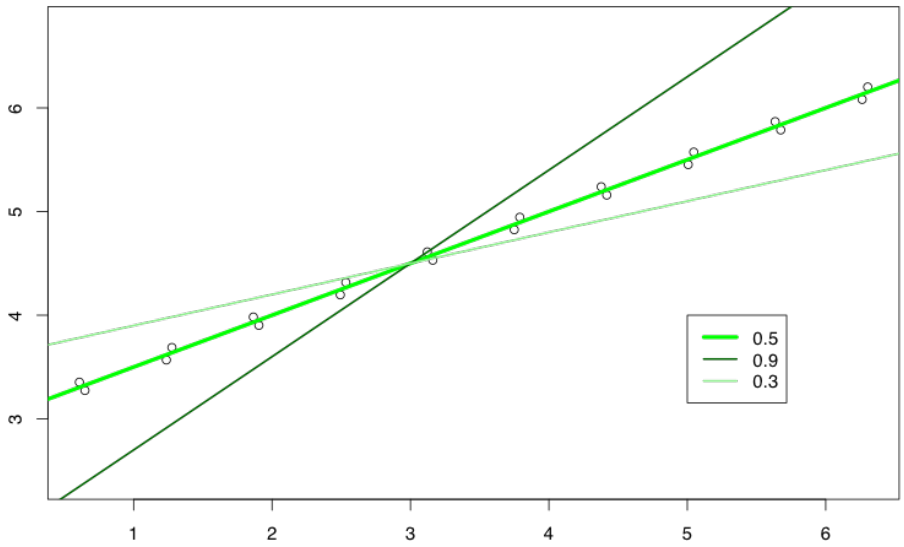


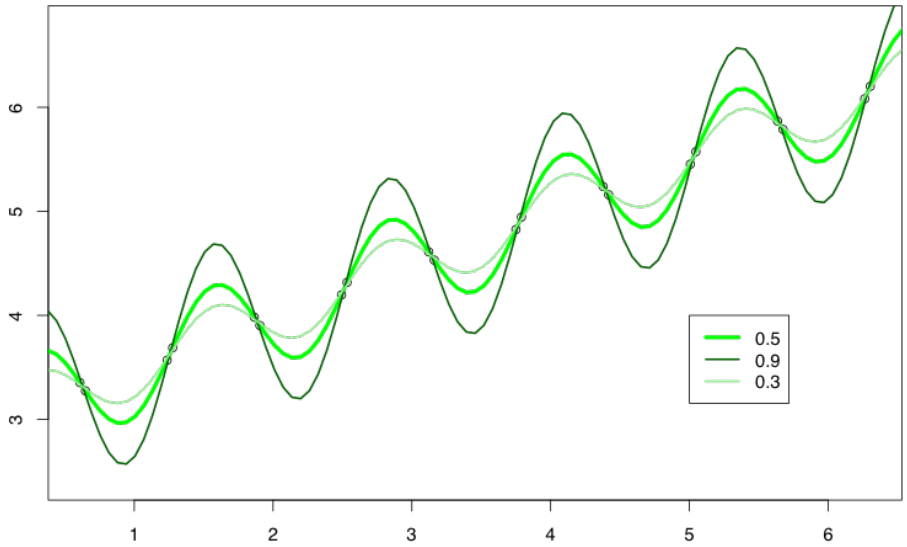


Prostora modelov

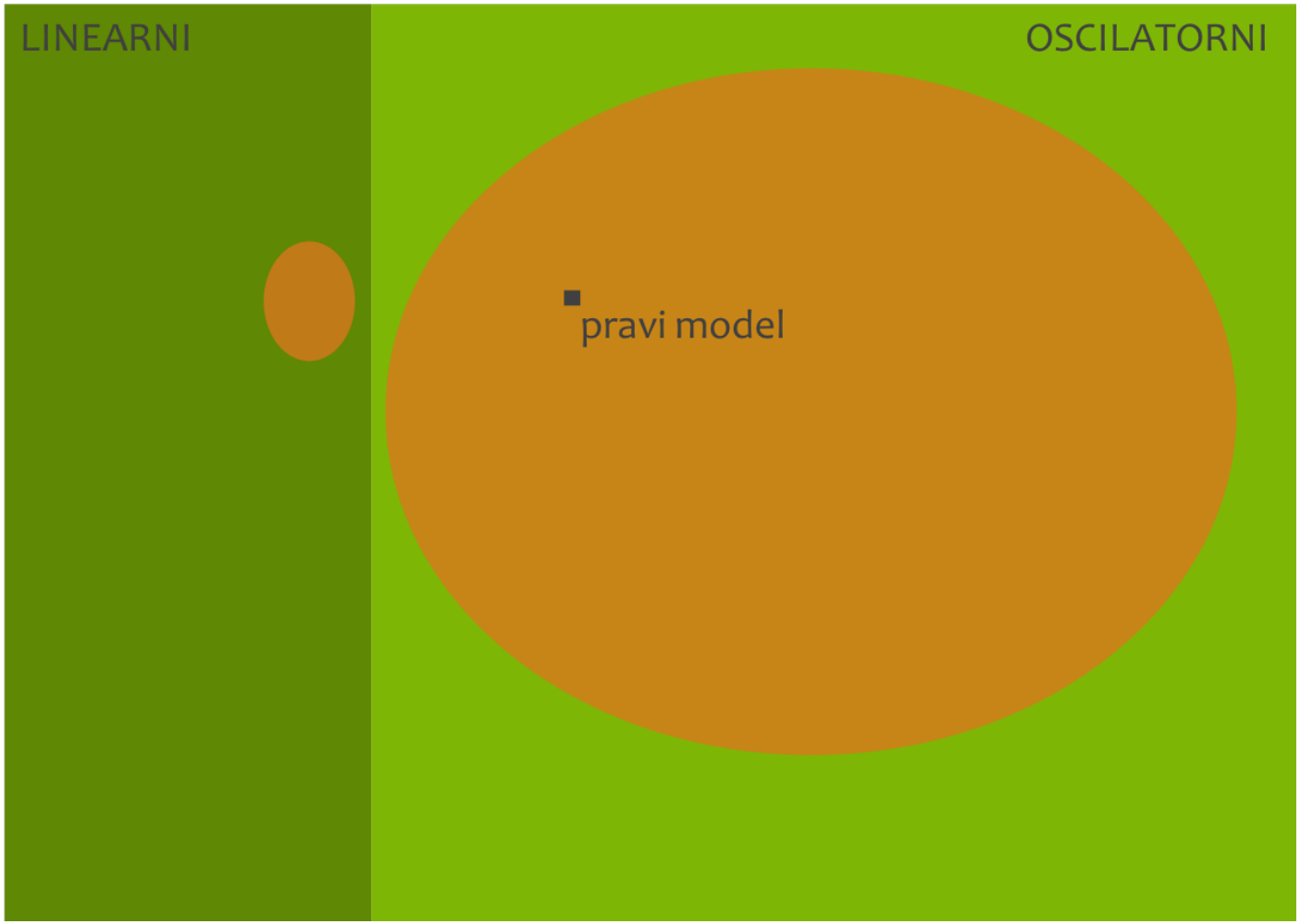
Primerjava varianc

- Nizka varianca linearnih modelov (levo): napaka modela zelo občutljiva na spremembe modela (naklon premice)
- Visoka varianca oscilatorskih modelov (desno): napaka modela neobčutljiva na spremembe modela (amplituda oscilacij)

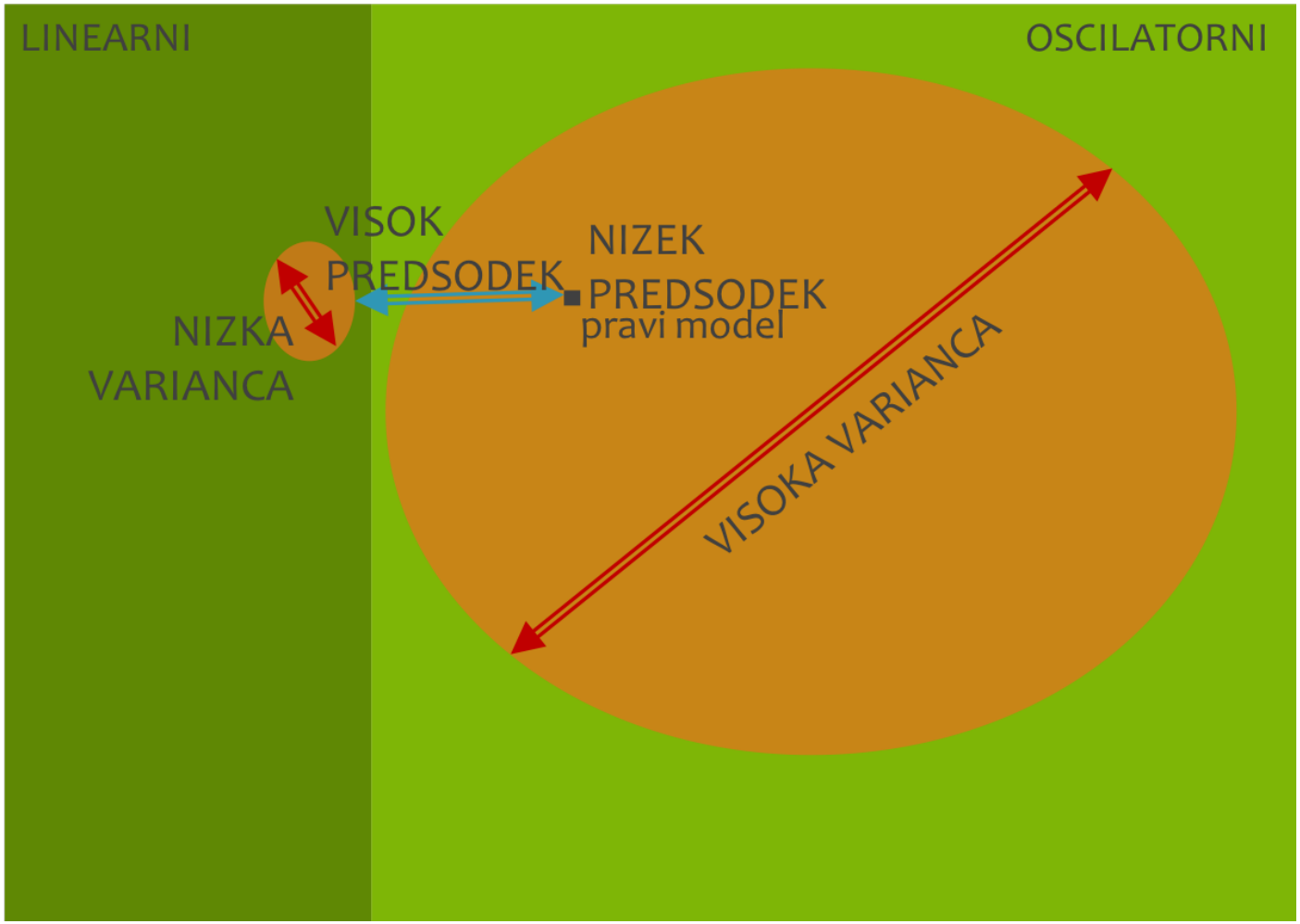




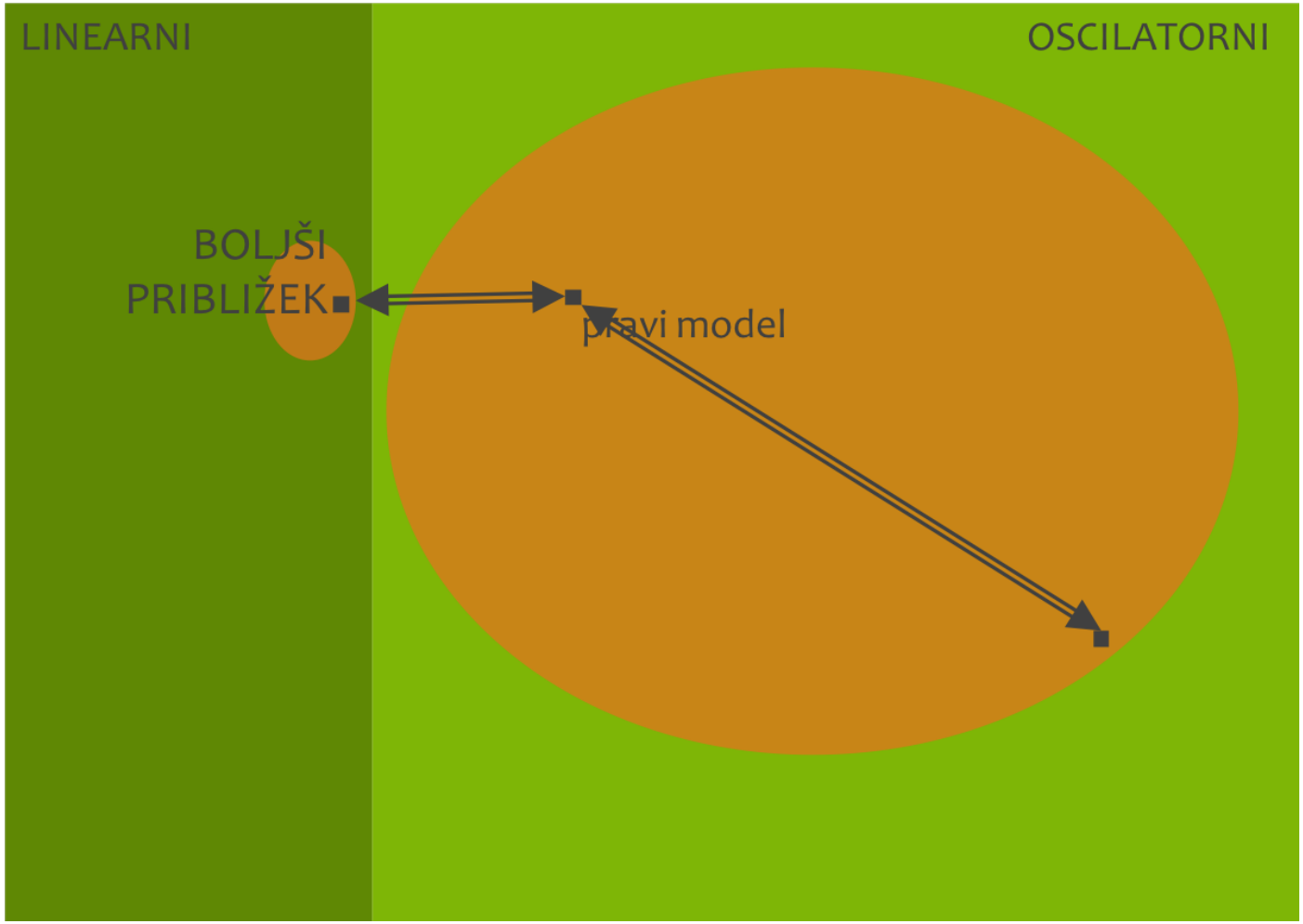
Prostora modelov in podatki



Prostora modelov: kompromis med predsodkom in varianco



Možna posledica: napačna izbira modela



Pravi model $m : Y = m(X) + \epsilon, E(\epsilon) = 0$

1. $E(m(\boldsymbol{x}_0)) = m(\boldsymbol{x}_0)$, ker je m deterministično
2. $E(Y|X = \boldsymbol{x}_0) = E(m(\boldsymbol{x}_0) + \epsilon) =_{[1]} m(\boldsymbol{x}_0)$
3. $Var(Y|X = \boldsymbol{x}_0) = E((Y - E(Y))^2|X = \boldsymbol{x}_0)$
   $= E((Y - m(\boldsymbol{x}_0))^2|X = \boldsymbol{x}_0) = E(\epsilon^2) = \sigma_\epsilon^2$
4. $E(Y^2|X = \boldsymbol{x}_0) = E((m(\boldsymbol{x}_0)+\epsilon)^2) = E(m(\boldsymbol{x}_0))^2 + 2E(m(\boldsymbol{x}_0)\epsilon) + E(\epsilon^2)$
   $=_{[2,3]} m(\boldsymbol{x}_0)^2 + \sigma_\epsilon^2$

Naučeni napovedni model $\hat{m}$

1. $E(\hat{m}(\boldsymbol{x}_0)^2) = E(\hat{m}(\boldsymbol{x}_0))^2 + Var(\hat{m}(\boldsymbol{x}_0))$,
   ker na splošno velja $Var(U) = E(U^2) - E(U)^2$

Dekompozicija napovedne napake modela *Err*

$$
\begin{aligned}
Err(\boldsymbol{x}_0) &= E((Y-\hat{m}(\boldsymbol{x}_0))^2|X=\boldsymbol{x}_0) \\
&= E(Y^2+\hat{m}(\boldsymbol{x}_0)^2-2Y\hat{m}(\boldsymbol{x}_0)|X=\boldsymbol{x}_0) \\
&= E(Y^2|X=\boldsymbol{x}_0)+E(\hat{m}(\boldsymbol{x}_0)^2)-2E(Y\hat{m}(\boldsymbol{x}_0)|X=\boldsymbol{x}_0) \\
&=_{[m2,m4]} m(\boldsymbol{x}_0)^2+\sigma_\epsilon^2+E(\hat{m}(\boldsymbol{x}_0)^2)-2m(\boldsymbol{x}_0)E(\hat{m}(\boldsymbol{x}_0)) \\
&=_{[\hat{m}1]} \sigma_\epsilon^2+m(\boldsymbol{x}_0)^2+E(\hat{m}(\boldsymbol{x}_0))^2+Var(\hat{m}(\boldsymbol{x}_0)) \\
&\quad -2m(\boldsymbol{x}_0)E(\hat{m}(\boldsymbol{x}_0)) \\
&= \sigma_\epsilon^2+(E(\hat{m}(\boldsymbol{x}_0))-m(\boldsymbol{x}_0))^2+Var(\hat{m}(\boldsymbol{x}_0)) \\
&= \sigma_\epsilon^2+Predsodek^2+Varianca
\end{aligned}
$$

Komponente napovedne napake

σ_ϵ^2 je fiksna komponenta

Na to komponento nimamo vpliva.

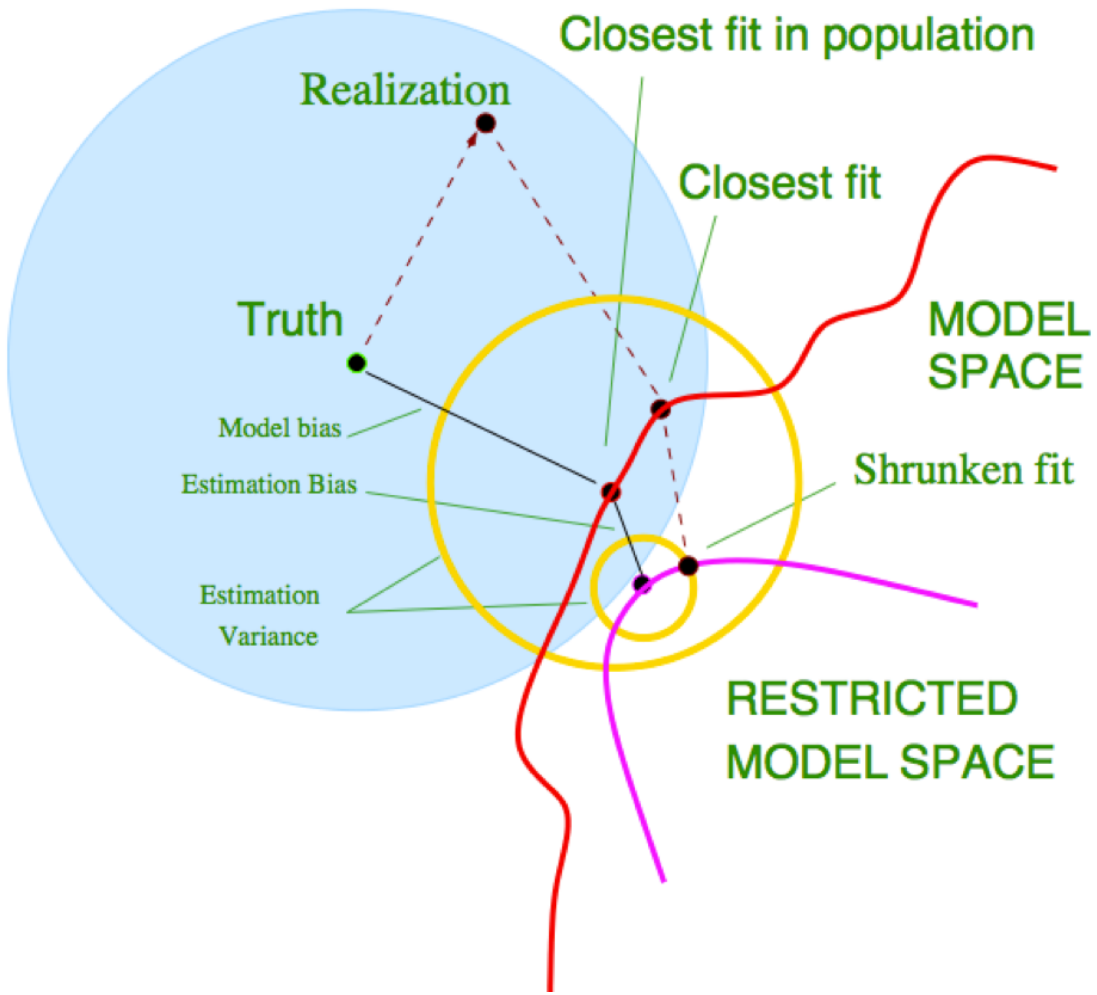
Predsodek $(E(\hat{m}(\boldsymbol{x}_0))-m(\boldsymbol{x}_0))^2$

- Razlika med pričakovano vrednostjo napovedi $\hat{Y}$ in vrednostjo Y
- Za podani prostor modelov nam pove koliko se lahko $\hat{Y}$ približa Y

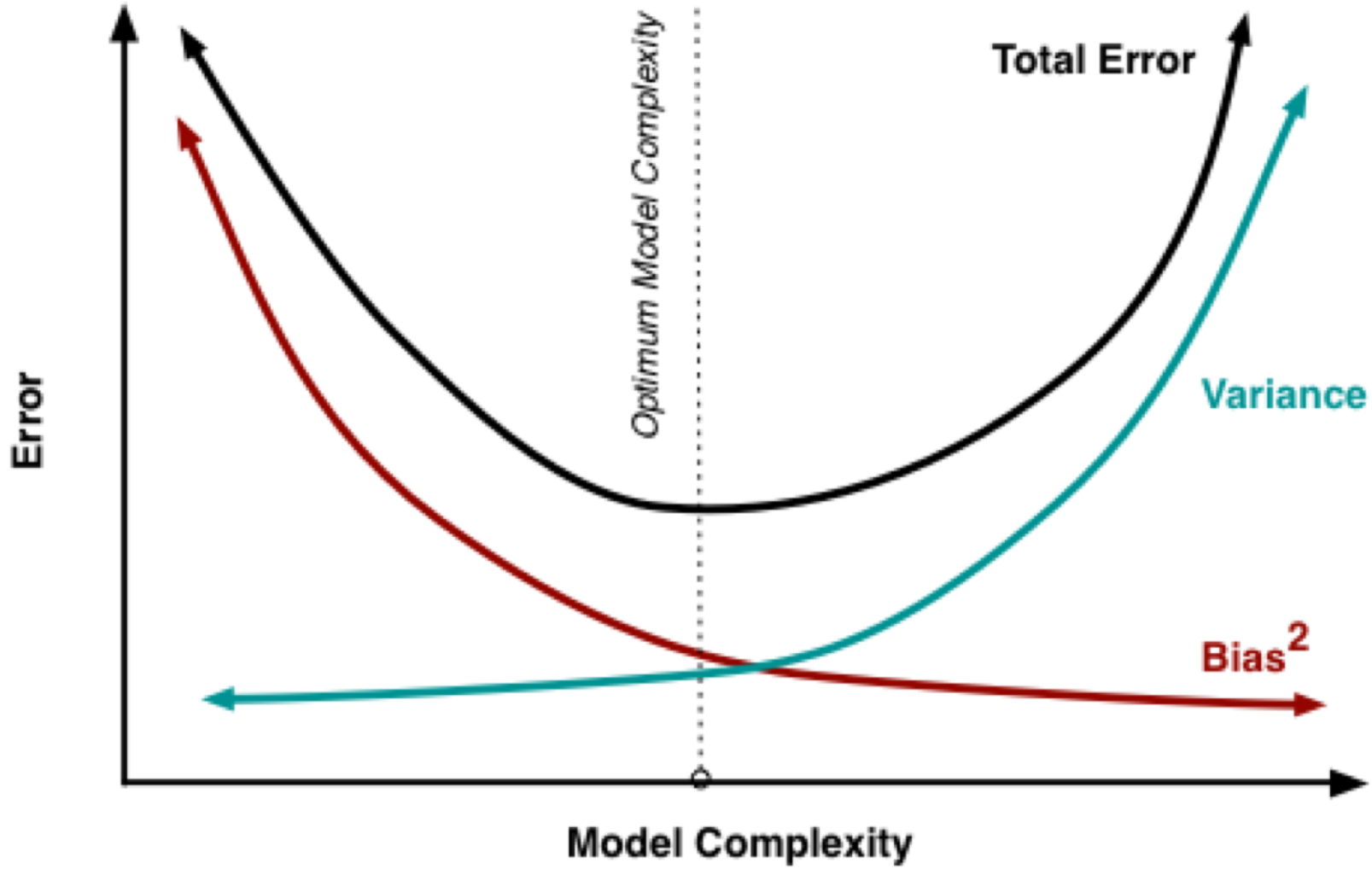
Varianca $Var(\hat{m}(\boldsymbol{x}_0))$

- Je varianca napovedi $\hat{Y}$ okoli pričakovane vrednosti $E(\hat{Y})$
- Za napoved $\hat{Y}$ in model $\hat{m}$ nam pove kakšna je njena stabilnost

Predsodek in varianca: grafična ponazoritev



Kompromis med predsodkom in varianco



Prileganje, predsodek in varianca

Preprileganje (*overfitting*)

- Situacija ko se soočamo z nizkim predsodkom in visoko varianco
- Kompleksni prostor modelov: najbližji sosedi z nizkim k
- Desna polovica grafa na prejšnji prosojnici

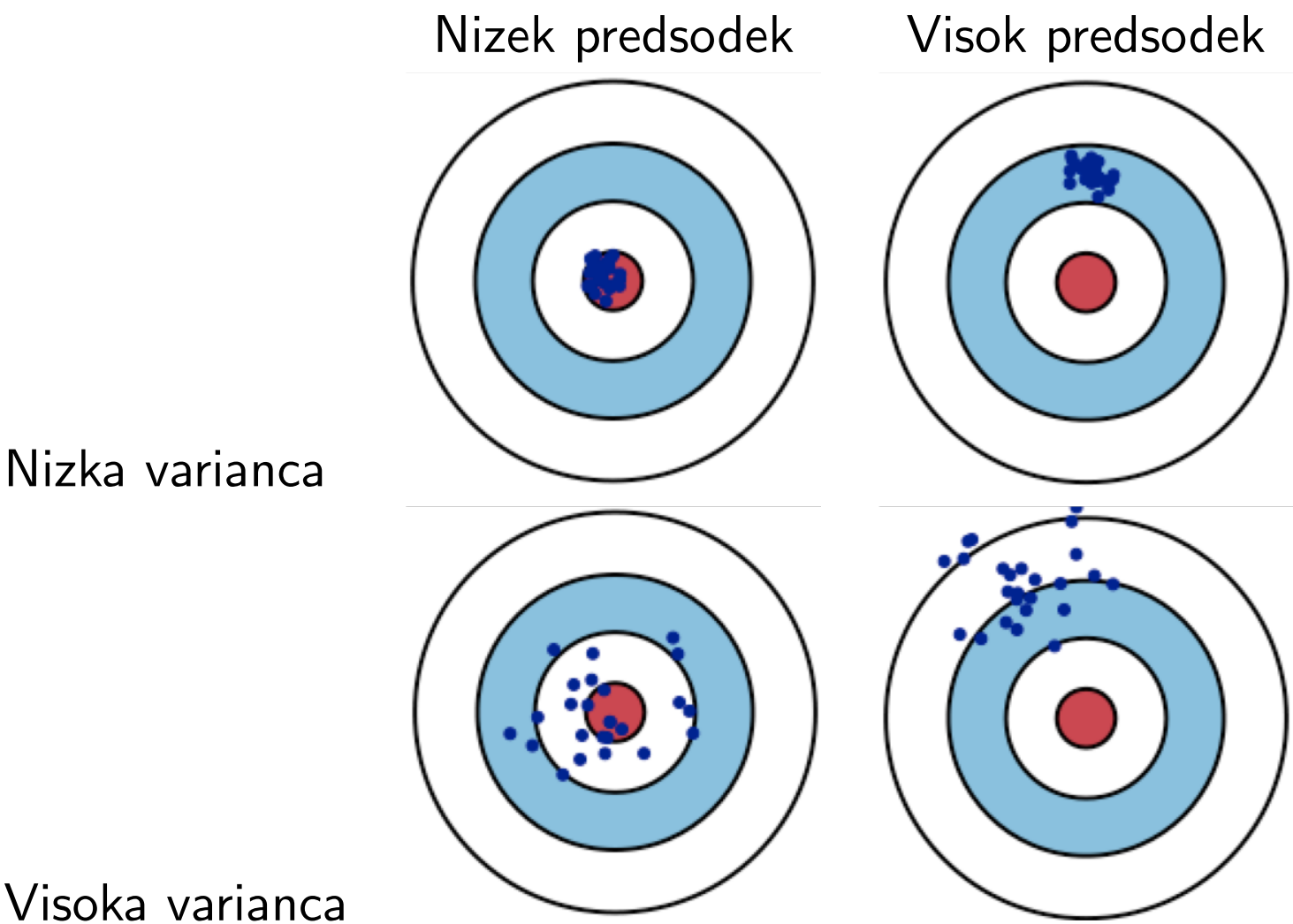
Podprileganje (*underfitting*)

- Situacija, ko se soočamo z visokim predsodkom in nizko varianco
- Enostaven prostor modelov: linarna oz. logistična regresija
- Leva polovica grafa na prejšnji prosojnici

Optimalni model

- Je idealna rešitev kompromisa med predsodkom in varianco
- Ravno prav zapleten prostor modelov

Predsodek in variance metode strojnega učenja



Priporočila

Zmanjševanje predsodka

- Linearni modeli: dodajanje nelinearnih členov
- Najbližji sosedi: zmanjševanje števila sosedov

Zmanjševanje variance

- Linearni modeli skoraj nikoli nimajo težav z visoko varianco
- Najbližji sosedi: povečevanje števila sosedov in zmanjševanje množice vhodnih spremenljivk (izbira spremenljivk)
- Generalna metoda: ansambli modelov

V večini primerov zmanjševanje predsodka povečuje varianco in obratno: zato ocenjevanje celotne napovedne napake

Učna in realna napaka

Funkcija izgube $L : D_Y \times D_Y \to \mathbb{R}_0^+$

- Za regresijo $L_{SE}(y, \hat{y}) = (y - \hat{y})^2$
- Za klasifikacijo $L_{01}(y, \hat{y}) = 1 - I(y, \hat{y})$

Učna napaka: napaka m izmerjena na učni množici S

$Err_{train}(m, S) = \frac{1}{|S|} \sum_{(\boldsymbol{x},y)\in S} L(y, m(\boldsymbol{x}))$
Optimistična: izmerjena na primerih uporabljenih za učenje modela

Realna napaka: napaka m ocenjena na primerih izven S

$Err_{real}(m, S) = E(L(y, m(\boldsymbol{x}))|(\boldsymbol{x}, y) \notin S)$

Teoretični rezultati in ...

Optimizem učne napake $o = Err_{real} - Err_{train}$

Če upoštevamo negotovost Y v učni množici, nas zanima $\omega = E_Y(o)$

Najbolj splošen rezultat

$$\omega = \frac{2}{|S|} \sum_{(\boldsymbol{x},y)\in S} Cov(m(\boldsymbol{x}), y)$$

Optimizem je obratno sorazmeren številu učnih primerov $|S|$

Rezultat za linearne modele $Cov(m(\boldsymbol{x}), y) = p\sigma_\epsilon^2$

Optimizem je premo sorazmeren številu napovednih spremenljivk p

Praksa

Uporabna rešitev

Ocenjevanje napake na posebni testni množici oziroma na primerih, ki niso bili uporabljeni za učenje modela.

Praktični pristop

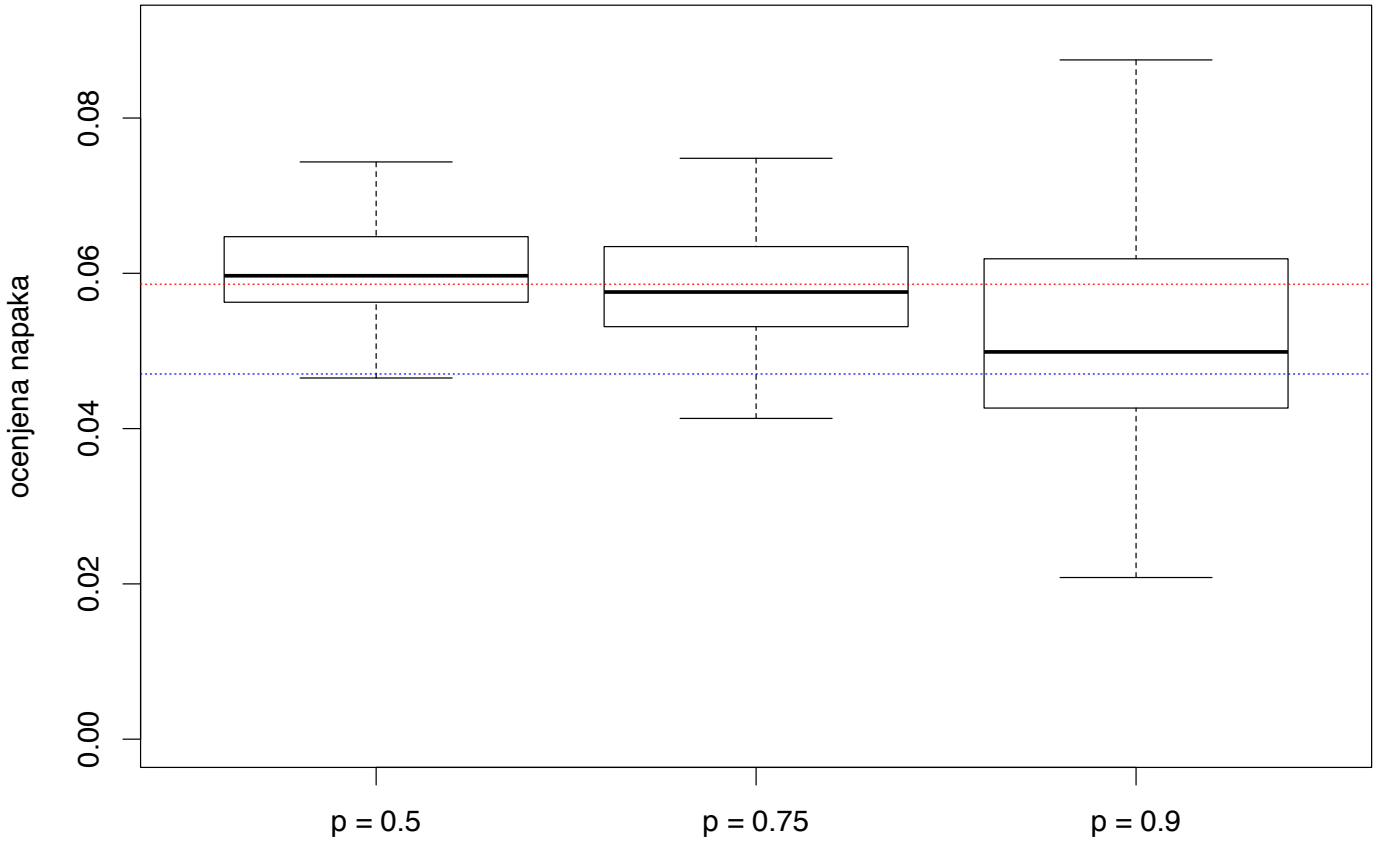
Razbitje učne množice S na

- podmnožico primerov za učenje modela S_{train}
- podmnožico primerov za ocenjevanje napake S_{test}
- $S_{train} \cup S_{test} = S, \quad S_{train} \cap S_{test} = \emptyset$

Izbor primerov v S_{train} naključen

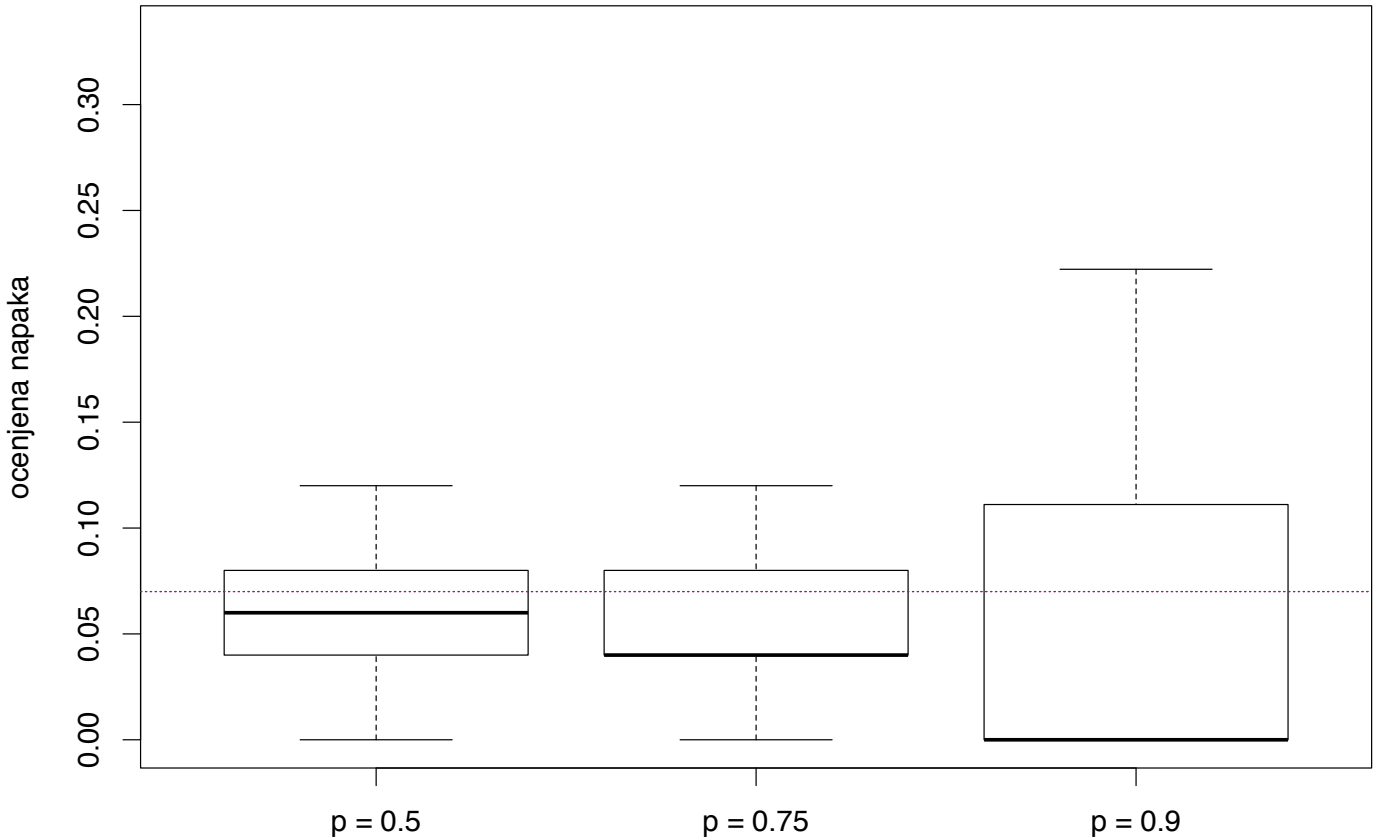
Odločimo se o deležu učnih primerov $|S_{train}|/|S|$: oznaka p v grafih

Ocena napake regresijskega modela kNN: $k = 5$



$$Y = (1 + X_1 + X_1X_2)/3 + \mathcal{N}(0, 0.05),\ D_i = [0, 1], i = 1..2, D_Y = [0, 1]$$

Ocena napake klasifikacijskega modela kNN: $k = 7$



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

Zakaj vzorčenje

Nižanje variance ocene

Grafi razkrijejo visoko varianco ocene napovedne napake

Dva pristopa

- Prečno preverjanje (*cross validation*)
- Zankanje (*bootstrap*)

Razbitje učne množice S

k podmnožic $S_i : i = 1..k$ (k-kratno prečno preverjanje)

- S_i: $|S_i| \approx |S|/k$
- $S = \bigcup_{i=1}^{k} S_i, \quad \forall i \neq j : S_i \cap S_j = \emptyset$
- Stratifikacija: vsaka podmnožica ima približno enako porazdelitev Y

Od razbitja do ocene napake

Za vsako S_i, $i = 1..p$ opravimo koraka

1. Naučimo se napovedni model m_i na primerih iz $S \setminus S_i$
2. Izmerimo napako Err_i na primerih iz S_i: $Err_i = Err(m_i, S_i)$

Ocena napovedne napake Err_{CV} in alternativa

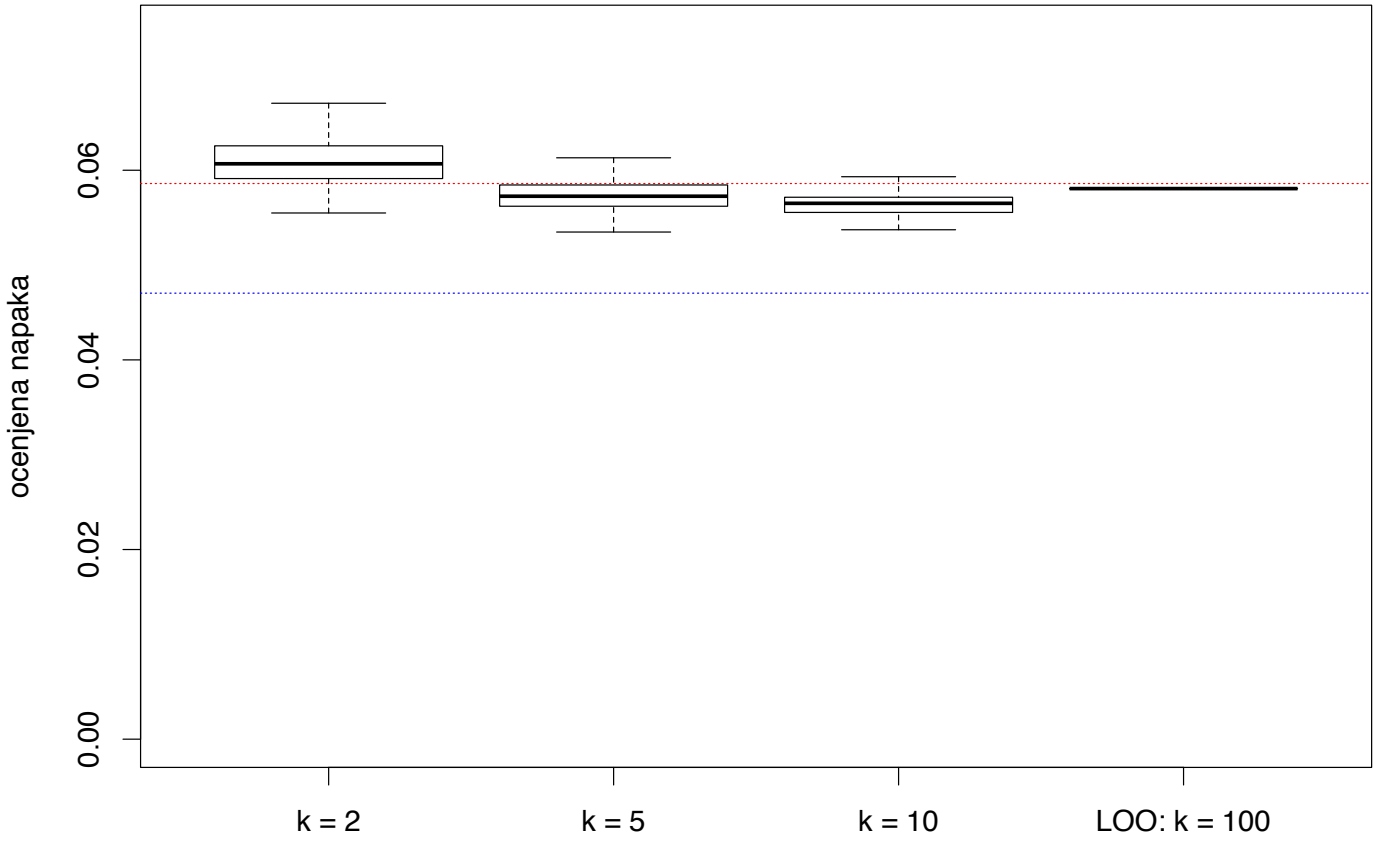
$$Err_{CV} = \frac{1}{k} \sum_{i=1}^{k} Err_i$$

$Err_{CV1} = \frac{1}{|S|} \sum_{e=(\boldsymbol{x},y) \in S} L(y, m_e(\boldsymbol{x}))$, kjer je $m_e = m_i : e \in S_i$

Običajni vrednosti k: 5 ali 10

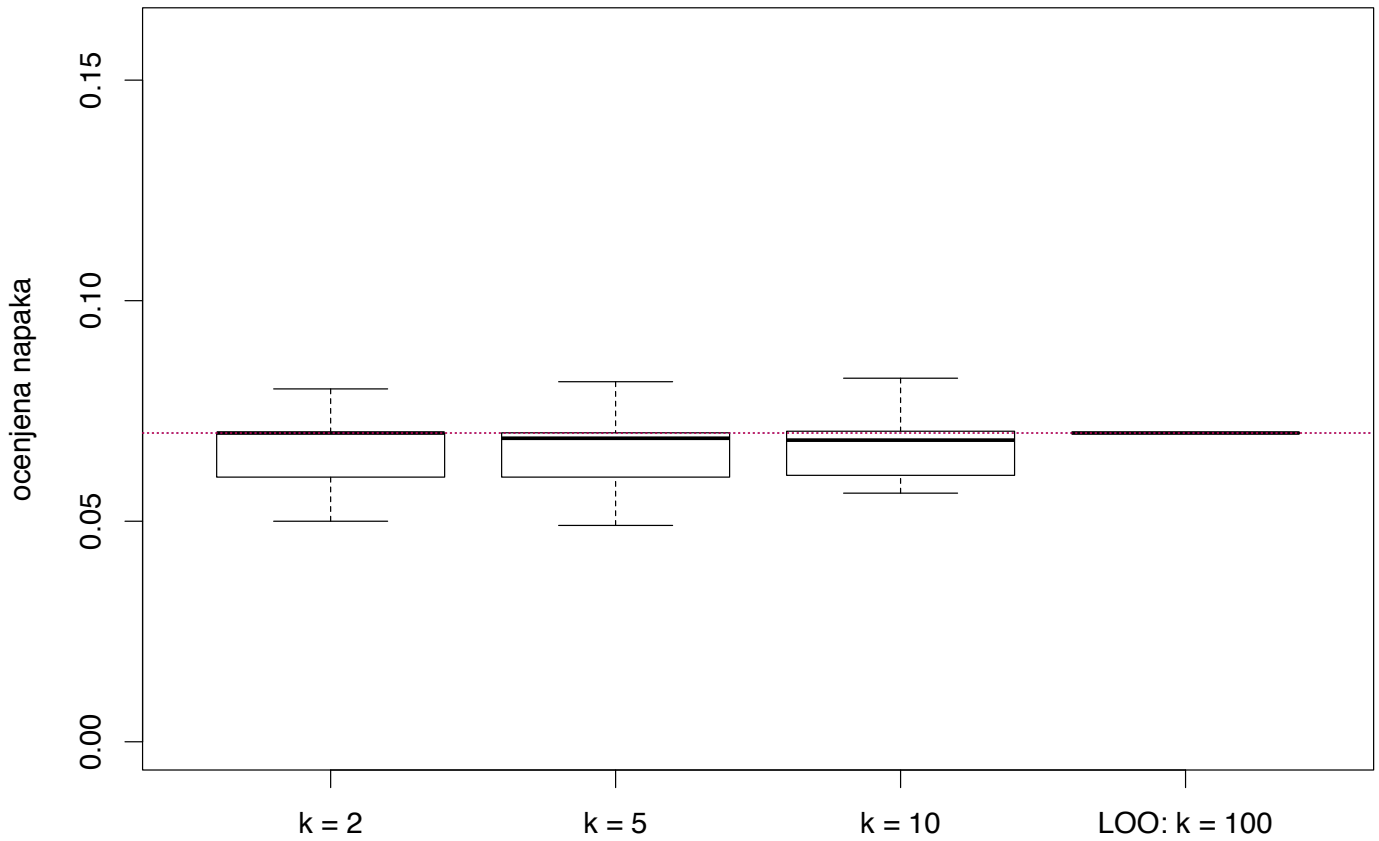
Poseben primer $k = |S|$: **izpusti enega** (*leave-one-out, LOO*)

Ocena napake regresijskega modela kNN: $k = 5$



$$Y = (1 + X_1 + X_1X_2)/3 + \mathcal{N}(0, 0.05),\ D_i = [0, 1], i = 1..2, D_Y = [0, 1]$$

Ocena napake klasifikacijskega modela kNN: $k = 7$



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

Vzorčenje učne množice S

B vzorcev $V_i : i = 1..B$ množice S

- $V_i : |V_i| = |S|$, vzorčenje s ponavljanjem
- Verjetnost $p(e \notin V_i) = (1 - 1/|S|)^{|S|}$

$$\lim_{|S| \to \infty} p(e \notin V_i) = \frac{1}{e} = 0.368$$

Od vzorčenja do ocene napake

Za vsak $V_i, i = 1..B$ opravimo koraka

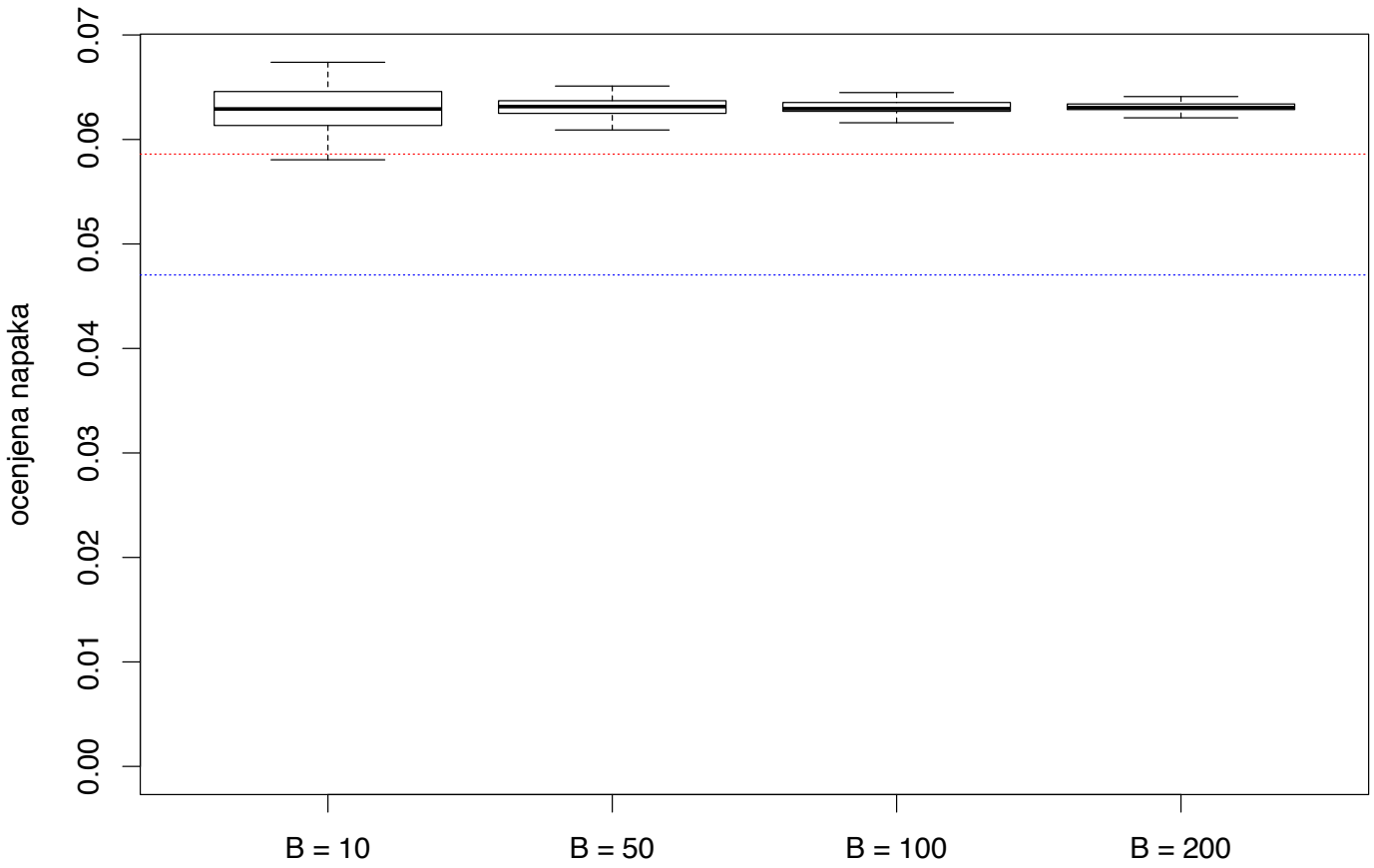
1. Naučimo se napovedni model m_i na primerih iz V_i
2. Izmerimo napako Err_i na primerih iz $S \setminus V_i$: $Err_i = Err(m_i, S \setminus V_i)$

Ocena napovedne napake Err_{BS}: običajna in alternativi

$$Err_{BS} = \frac{1}{B} \sum_{i=1}^{B} Err_i$$

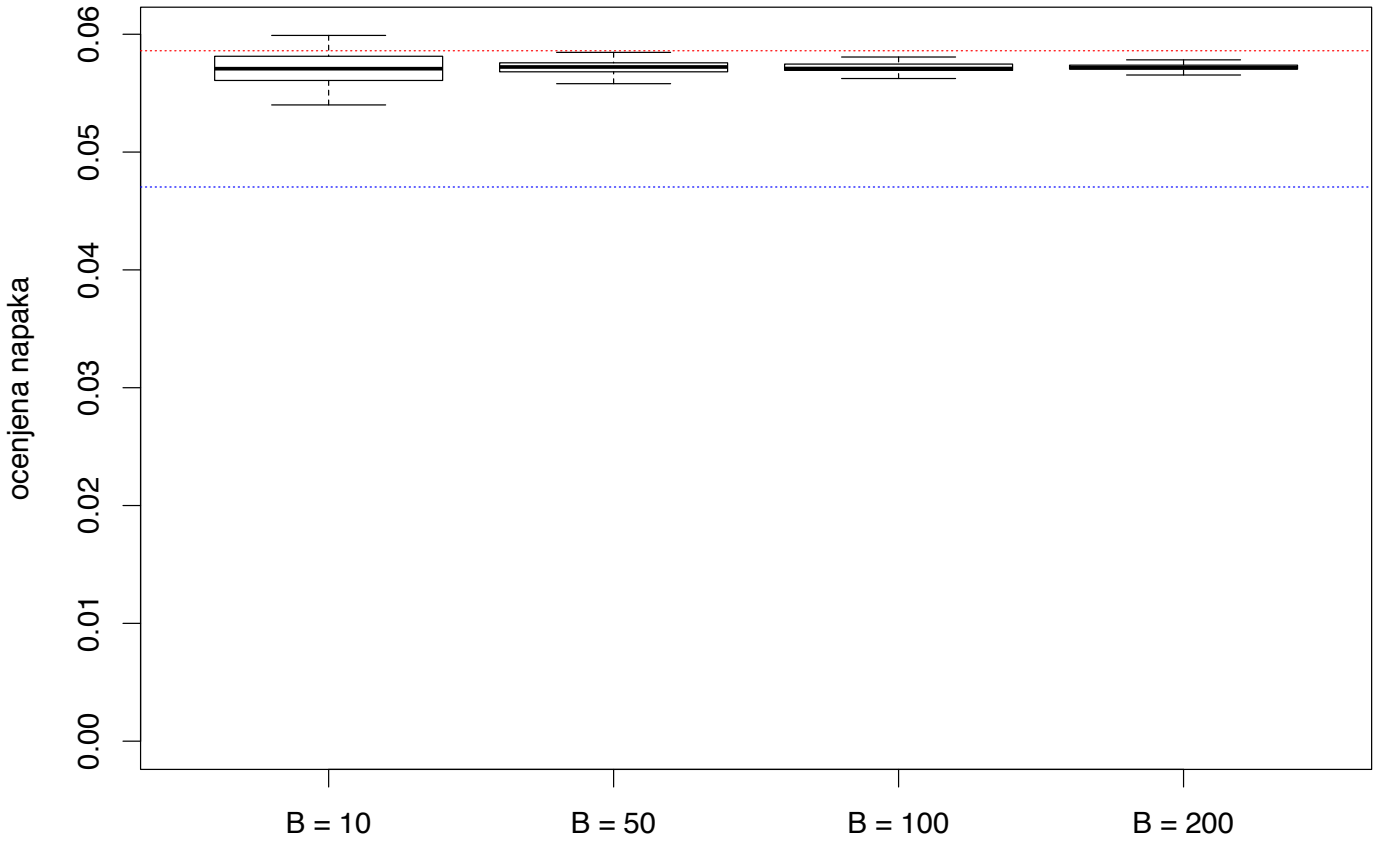
- $Err_{BS1} = \frac{1}{|S|} \sum_{e=(\boldsymbol{x},y)\in S} \frac{1}{|M_e|} \sum_{m\in M_e} L(y, m(\boldsymbol{x}))$, kjer je $M_e = \{m_i : e \notin V_i\}$
- $Err_{BS632} = 0.368 Err_{train} + (1 - 0.368) Err_{BS1}$

Ocena napake regresijskega modela kNN: $k = 5$



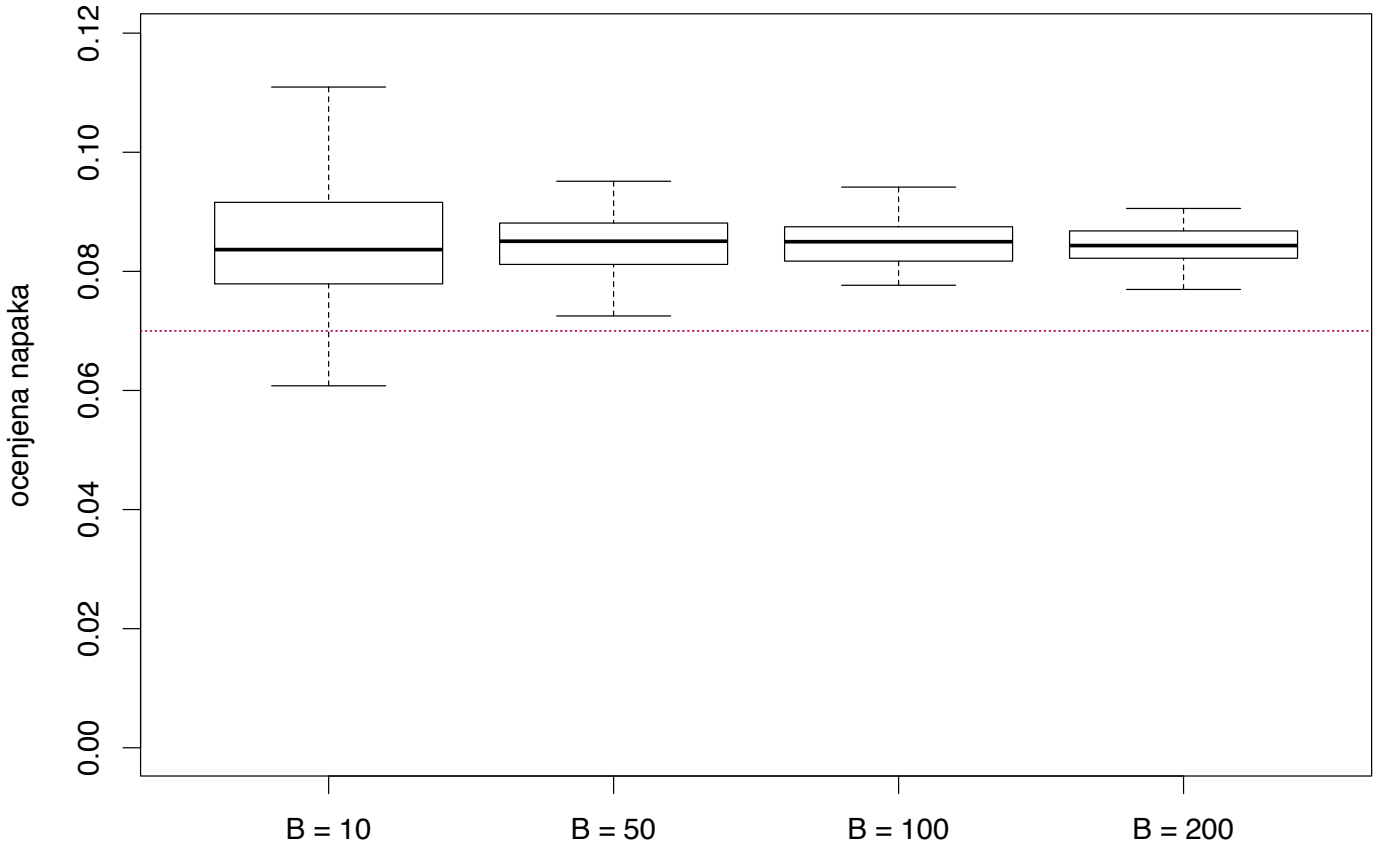
$$Y = (1 + X_1 + X_1X_2)/3 + \mathcal{N}(0, 0.05),\ D_i = [0, 1], i = 1..2, D_Y = [0, 1]$$

Ocena napake regresijskega modela kNN: $k = 5$, popravek



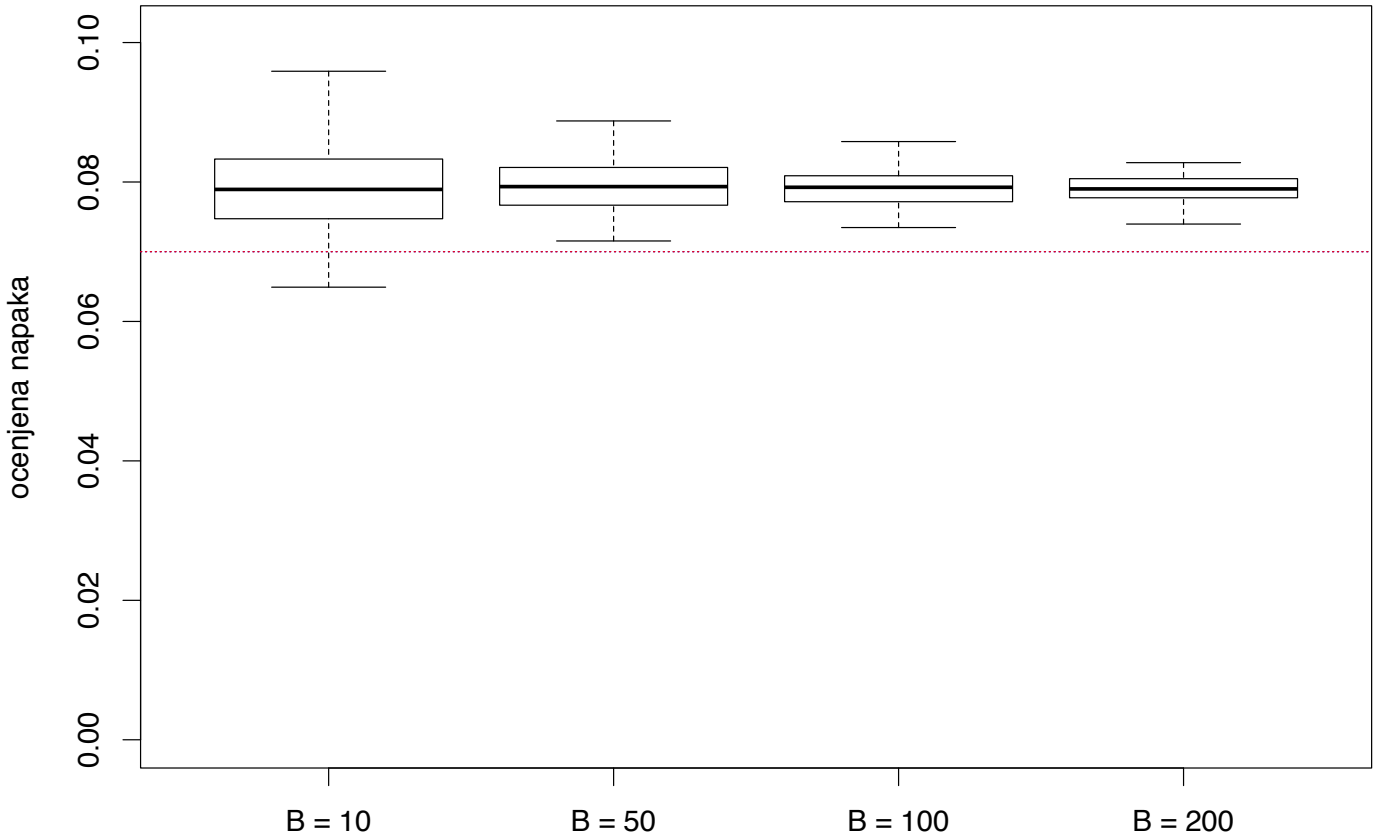
$$Y = (1 + X_1 + X_1X_2)/3 + \mathcal{N}(0, 0.05),\ D_i = [0, 1], i = 1..2, D_Y = [0, 1]$$

Ocena napake klasifikacijskega modela kNN: $k = 7$



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

Ocena napake klasifikacijskega modela kNN: $k = 7$, popr.



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

Izbira modela

Za podano podatkovno množico S zberemo

- Nabor algoritmov $\mathcal{A}_1, \mathcal{A}_2, \ldots \mathcal{A}_q$ za strojno učenje
- Za vsak algoritem $\mathcal{A}_i : i = 1..q$ izberemo nabor nastavitev parametrov $N_{i1}, N_{i2}, \ldots N_{ir_i}$

Za vsak par $(\mathcal{A}_i, N_{ij})$

1. Uporabimo zankanje na množici S za oceno napake Err_{ij}
2. Ugotovimo kateri par ima minimalno napako Err
3. Uporabimo ta par za gradnjo končnega modela na celi množici S

Izbira algoritmov in nastavitev

Kateri algoritmi?

Lastnosti podatkovne množice: število primerov, spremenljivk, domene

Kateri parametri?

Tisti, ki spreminjajo kompleksnost modela
oziroma vplivajo na odnos med predsodkom in varianco.

Priporočila za enostavne metode

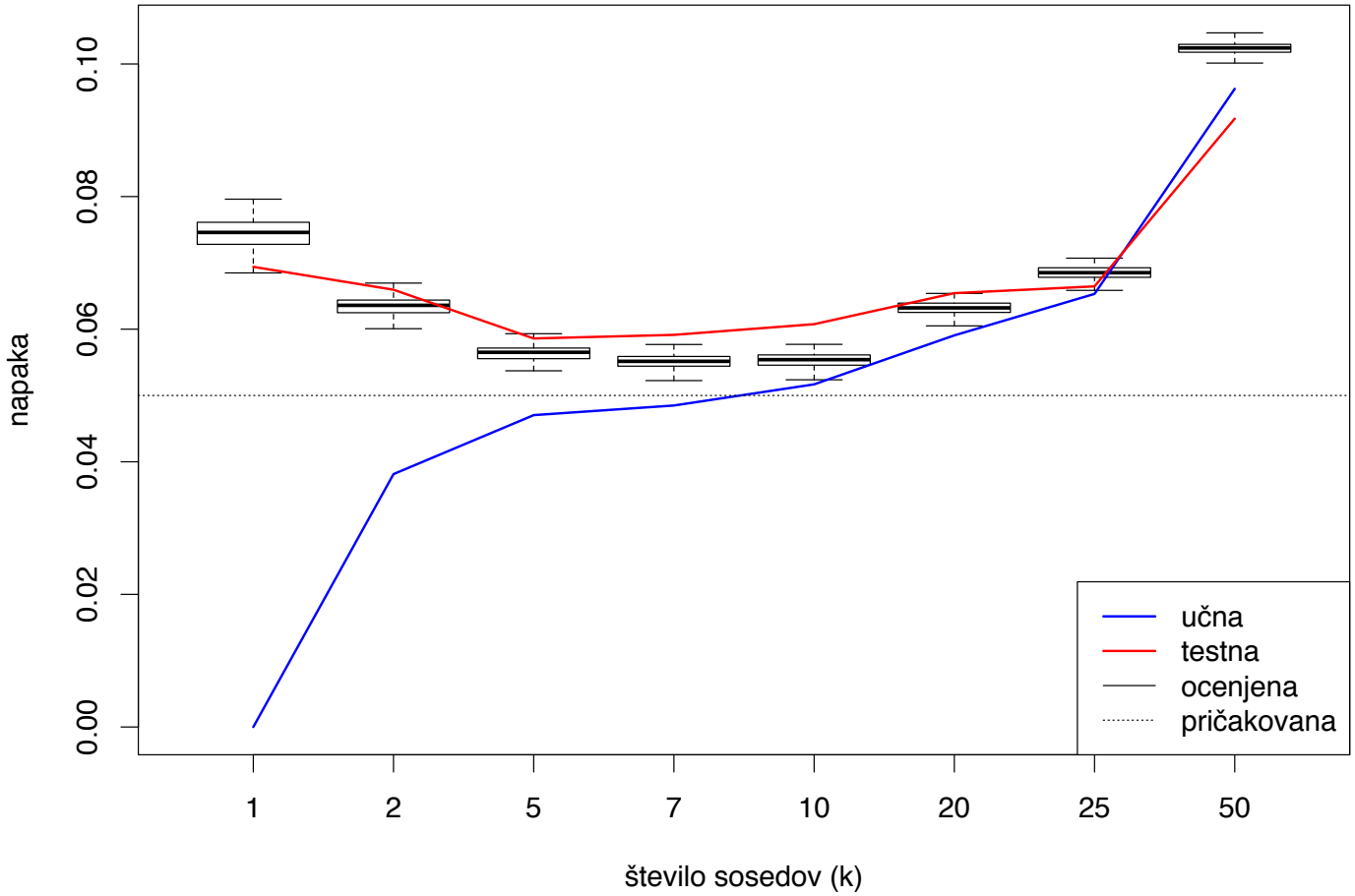
Linearni modeli

- Manjši predsodek: nelinearne kombinacije napovednih spremenljivk
- Manjša varianca: izbira podmnožice vhodnih spremenljivk
- Manjša varianca: regularizacija

Najbližji sosedi

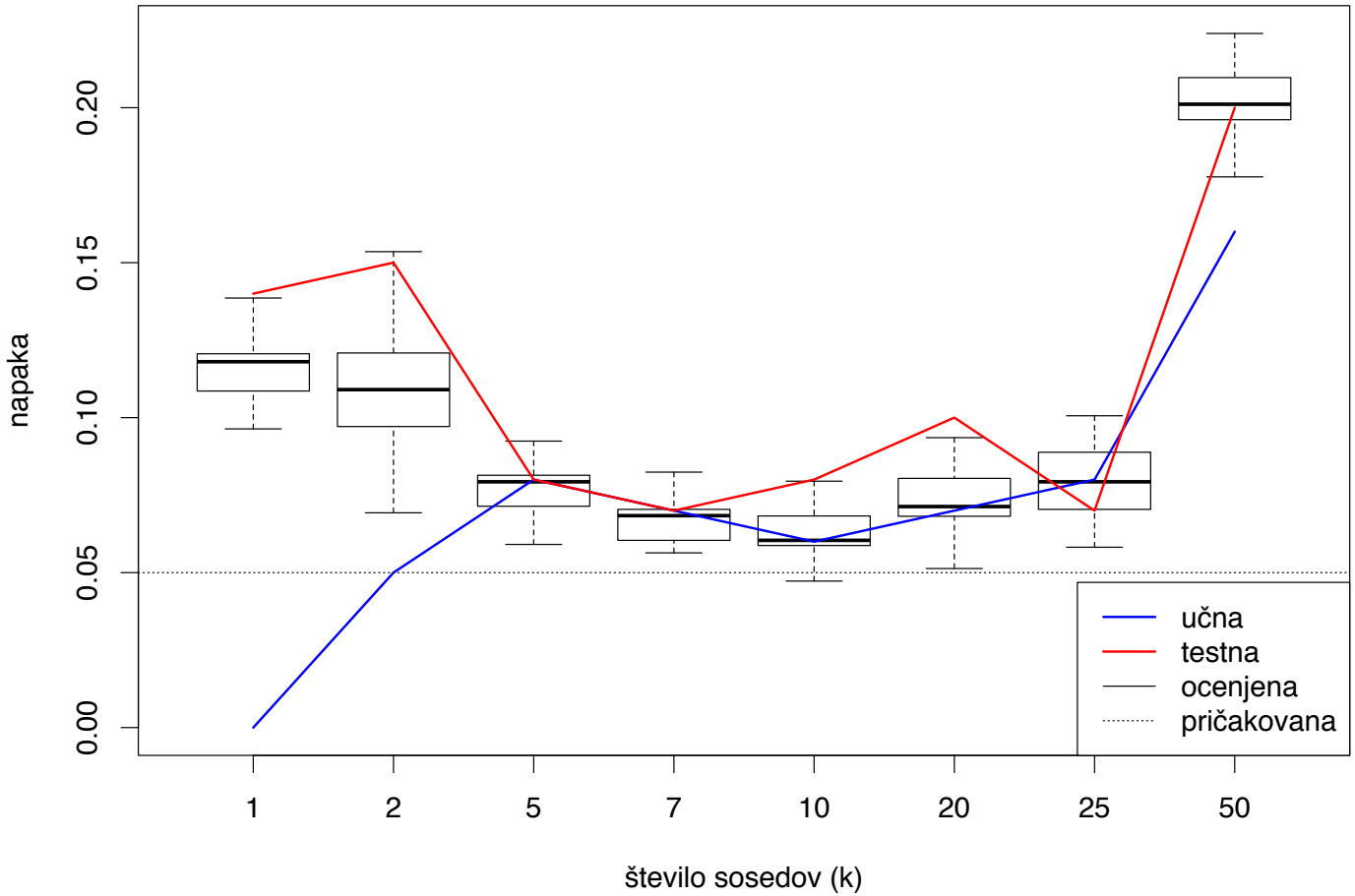
- Manjši predsodek: manjše število sosedov k
- Manjša varianca: izbira podmnožice vhodnih spremenljivk

Izbira števila sosedov za regresijski model: CV-10



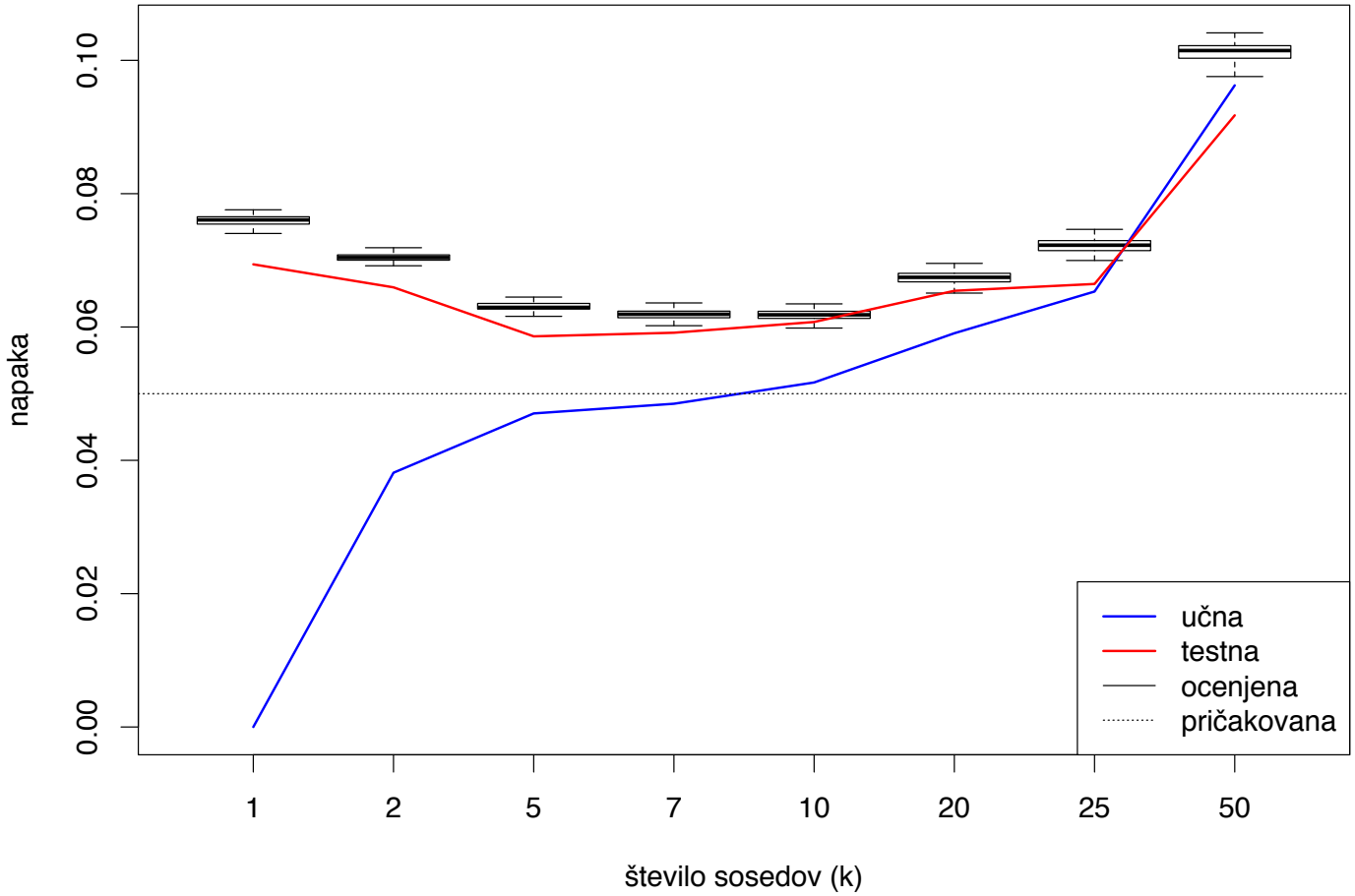
$$Y = (1 + X_1 + X_1X_2)/3 + \mathcal{N}(0, 0.05),\ D_i = [0, 1], i = 1..2, D_Y = [0, 1]$$

Izbira števila sosedov za klasifikacijski model: CV-10



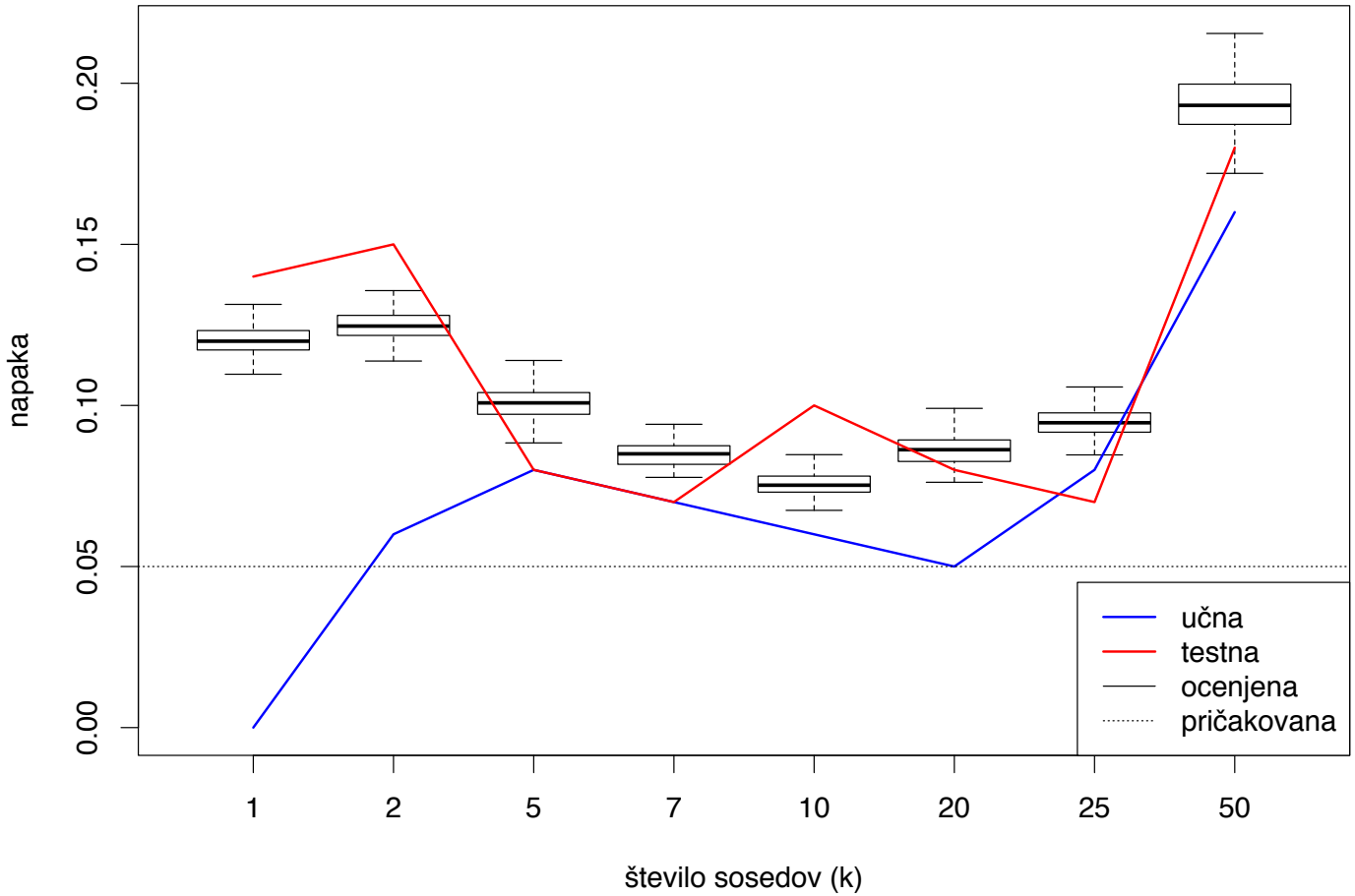
$Y = I((1 + X_1 + X_1X_2)/3 \geq 0.5), D_i = [0, 1], i = 1..2, D_Y = \{0, 1\}$
zamenjamo vrednost Y 0.05 naključno izbranim primerom

Izbira števila sosedov za regresijski model: BS-100



$$Y = (1 + X_1 + X_1X_2)/3 + \mathcal{N}(0, 0.05),\ D_i = [0, 1], i = 1..2, D_Y = [0, 1]$$

Izbira števila sosedov za klasifikacijski model: BS-100



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

Izbira modela za podatkovno množico Iris: BS-100