Enostavne metode napovednega modeliranja

Ljupčo Todorovski
Univerza v Ljubljani, Fakulteta za upravo

Februar 2018

Pregled predavanja

Linearna regresija

- Metoda najmanjših kvadratov
- Intervali zaupanja in statistična značilnost

Logistična regresija

- Metoda največjega verjetja
- Razlike med linearno in logistično regresijo

Metoda najbližjih sosedov

- Soseščina in razdalje
- Vpliv števila najbližjih sosedov

Naloga

Podatkovna množica $S \in \mathbb{R}^{p+1}$

- Numerične napovedne spremenljivke $\boldsymbol{X} = \{X_1, X_2, \ldots X_p\}$, $D_i = \mathbb{R}$
- Numerična ciljna spremenljivka Y, $D_Y = \mathbb{R}$

Model

$$\hat{Y} = \beta_0 + \sum_{i=1}^{p} \beta_i X_i$$

$\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots \beta_p) \in \mathbb{R}^{p+1}$ je vektor konstantnih koeficientov modela: izsek β_0 in p smernih koeficientov β_i, $i = 1..p$

Vektorska predstavitev linearnega modela

$$\hat{Y} = \beta^T \boldsymbol{X}$$

$\boldsymbol{X} = (X_0, X_1, \ldots X_p)$ je razširjen vektor napovednih spremenljivk; dodana spremenljivka $X_0 = \mathbf{1}$ ima domeno $D_0 = \{1\}$, t.j., vrednost X_0 je v vseh primerih podatkovne množice 1.

Kako do vrednosti koeficientov?

Iščemo optimalne vrednosti koeficientov β za učno množico S

$$\min_{\beta} \sum_{(\boldsymbol{x},y)\in S} (y - \beta^T \boldsymbol{x})^2 = \min_{\beta} (S_Y - S_{\boldsymbol{X}}\beta)^T (S_Y - S_{\boldsymbol{X}}\beta)$$

- $S_{\boldsymbol{X}}$ je matrika (dimenzij $|S| \times p$) vrednosti spremenljivk napovednih spremenljivk $\boldsymbol{X}$ v primerih iz S
- S_Y je vektor (dolžine $|S|$) vrednosti spremenljivke Y v S

Opozorilo o spremembi notacije $S_X \to X, S_Y \to Y$

Za matriko S_X bomo uporabljali oznako X (običajno napovedna spremenljivka), za vektor S_Y pa Y (običajno ciljna spremenljivka)

$$X : S_X = \begin{array}{c|cccc} & X_1 & X_2 & \dots & X_p \\ \hline e_1 & x_{11} & x_{12} & \dots & x_{1p} \\ e_2 & x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ e_n & x_{n1} & x_{n2} & \dots & x_{np} \end{array} \qquad Y : S_Y = \begin{array}{c|c} & Y \\ \hline e_1 & y_1 \\ e_2 & y_2 \\ \vdots & \vdots \\ e_n & y_n \end{array}$$

Metoda najmanjših kvadratov

Prva dva odvoda ciljne funkcije za minimizacijo *RSS*

- $\frac{\partial RSS}{\partial \beta} = -2X^T(Y - X\beta)$
- $\frac{\partial^2 RSS}{\partial \beta^2} = 2X^T X$

Minimum *RSS* dosežemo pri

$$X^T(Y - X\beta) = 0$$

Rešitev

$$\beta \quad = \quad (X^T X)^{-1} X^T Y$$

Geometrijska interpretacija

$$\hat{Y} = X(X^T X)^{-1} X^T Y = HY$$

Projekcijska matrika $H = X(X^T X)^{-1} X^T$

$\hat{Y}$ je projekcija vektorja Y na hiper-ravnino, ki jo razpenja X

Intervali zaupanja in statistična značilnost koeficientov

Intervali zaupanja za parametre β

$$Var(\beta) = (X^T X)^{-1} Var(Y)$$

- $Var(\beta_j) = v_j Var(Y)$, kjer je $v_j = (X^T X)^{-1}_{jj}$
- Ocena $Var(Y)$ na množici S: $\sigma^2 = \sum_{(\boldsymbol{x},y)\in S}(y - \beta^T \boldsymbol{x})^2/(n-p-1)$

Ali je vpliv X_j na Y statistično značilen?

- Ničelna $H_0 : \beta_j = 0$ in alternativna hipoteza $H_A : \beta_j \neq 0$
- $z_j = \beta_j/(\sigma\sqrt{v_j})$ je porazdeljena po Studentu $t(n–p–1)$

Napaka in statistična značilnost modela

Residualna standardna napaka

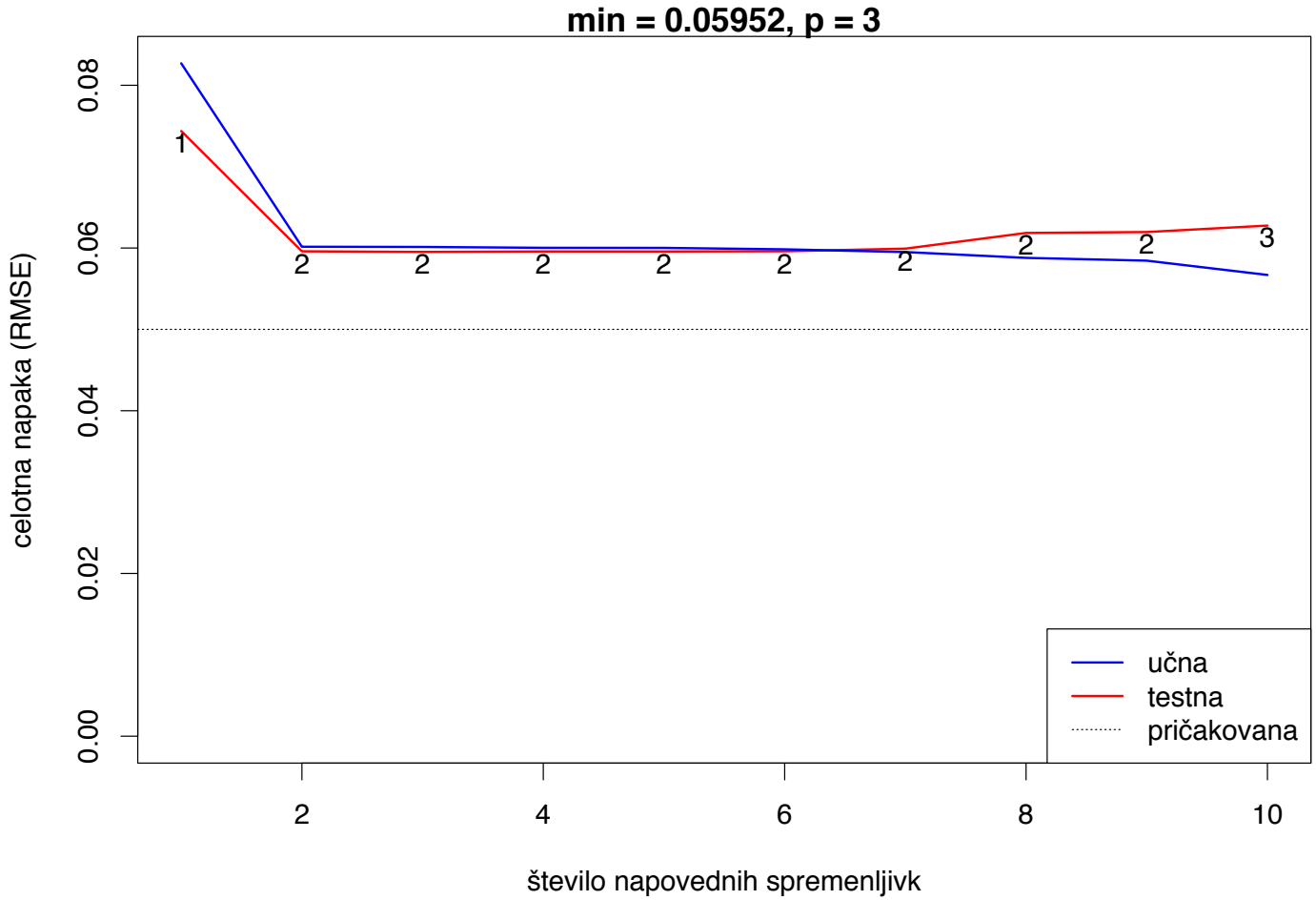
$$RSE = \sqrt{\frac{RSS}{n-p-1}}, RSS = \sum_{(\boldsymbol{x},y)\in S} (y - \boldsymbol{\beta}^T \boldsymbol{x})^2$$

Pogosto računamo tudi $R^2 = 1–RSS/TSS$, $TSS = \sum_{(\boldsymbol{x},y)\in S}(y–\bar{y})^2$, t.j., delež variance Y, ki ga lahko pojasnimo s pomočjo linearnega modela.

Ali je vpliv vsaj ene napovedne spremenljivke X_j na Y statistično značilen?

- Ničelna $H_0 : \forall j : \beta_j = 0$ in alternativna hipoteza $H_A : \exists j : \beta_j \neq 0$
- $F = ((TSS–RSS)/p)/(RSS/(n–p–1))$ je porazdeljena po Fišerju $F(p, n–p–1)$

Koliko spremenljivk ima statistično značilen vpliv?



$$Y = (1 + X_1 + X_1X_2)/3 + \mathcal{N}(0, 0.05),\ D_i = [0, 1], i = 1..10, D_Y = [0, 1]$$

Izbira napovednih spremenljivk

Katero podmnožico vhodnih spremenljivk izbrati, da bomo dobili najboljši linearni model?

Preverjanje vseh množic nemogoče: tri **požrešne** strategije

- **Izbira naprej**: Začnemo s praznim modelom, ki vključuje le β_0 nato iterativno dodajamo po eno spremenljivko X_i, ki najbolj zmanjša RSS.
- **Izbira nazaj**: Začnemo s polnim modelom, ki vključuje vse napovedne spremenljivko, nato iterativno odstranjujemo spremenljivke z najnižjo statistično značilnost vpliva na Y (uporabna le takrat, ko velja $p \leq n$).
- **Izbira naprej-nazaj**: Začnemo z izbiro naprej in ko se značilnost vpliva spremenljivk zniža pod neko ravnijo, začnemo izbiro nazaj; ponavljamo dokler se množica še spreminja.

Kvalitativne napovedne spremenljivke

Spremenljivko X_j z domeno vrednosti $D_j = \{v_1, v_2, \ldots v_m\}$

- Spremenimo v $U_{jk}, k = 1..m-1$ z domenami $D_{jk} = \{0, 1\}$

$$U_{jk} = \begin{cases} 1 & : X_j = v_k \\ 0 & : X_j \neq v_k \end{cases}$$

- Koeficient β_{jk} je enak spremembi vrednosti ciljne spremenljivke Y, ki jo povzroči sprememba vrednosti X_j iz v_m v v_l.

Potencialne težave

Soodvisnost napovednih spremenljivk

- Posledica $rang(X^T X) < p$, torej ne moremo izračunati $(X^T X)^{-1}$
- Alternativni načini ocenjevanja β

Nelinearni vpliv spremenljivk X na Y

- Predpostavke o možnih interakcijah in nelinearnostih
- Dodatne spremenljivke, npr. $X_1 X_2$ ali X_1^2: v obeh primerih model ostaja linearen glede na koeficiente β

Korelacija med reziduali modela za posamezne primere

- Omejena aplikacija linearnega modela za obravnavo časovnih vrst
- Oziroma časovno odvisnih primerov v podatkovni množici

Kaj če je ciljna spremenljivka Y diskretna?

Rešitev: klasifikacijski model kot množica regresijskih modelov

- Zamenjava ene ciljne spremenljivke z $|D_Y|$ ciljnih spremenljivk
- Za vsako vrednost $v \in D_Y$, definiramo

$$Y_v = \begin{cases} 1 & ; Y = v \\ 0 & ; Y \neq v \end{cases}$$

- Za vsako novo ciljno spremenljivko Y_v zgradimo regresijski model
- Napoved $\hat{Y}_v$ je verjetje, da je pravilna napoved za podan primer v
- Klasifikacijski model napove v, kjer $\hat{Y}_v$ doseže maksimum

Dve težavi in bolj korektna rešitev

Težavi

- Napovedana verjetja so lahko tudi negativna
- Običajna predpostavka, da je napaka regresijskega modela normalno porazdeljena je za te model očitno neveljavna

Korektna rešitev: logistična regresija

- Namesto da modeli napovedovali $p = p(Y = v|\boldsymbol{X} = \boldsymbol{x})$
- Napovedujejo logaritem obetov $log(p/(1-p))$
- Pogosto uporabljena pri $|D_Y| = 2$, kjer rabimo en regresijski model

Model logistične regresije

Od regresije k klasifikaciji pri $|D_Y| = 2, D_Y = \{0, 1\}$

$$p(Y = 1|\boldsymbol{X}) = \frac{e^{\hat{Y}}}{1 + e^{\hat{Y}}}$$

$$p(Y = 0|\boldsymbol{X}) = 1 - p(Y = 1|\boldsymbol{X}) = \frac{1}{1 + e^{\hat{Y}}}$$

$\hat{Y}$ je napoved regresijskega modela

Linearni klasifikacijski model

Odločitvena ali klasifikacijska meja $p(Y = 1|X) = p(Y = 0|X) = 0.5$

- To se zgodi pri $e^{\hat{Y}} = 1$, t.j., pri $\hat{Y} = 0$ oziroma $\boldsymbol{\beta}^T \boldsymbol{X} = 0$
- Meja je torej (linearna) hiper-ravnina

Zakaj ravno logaritem obetov?

Obet dogodka A

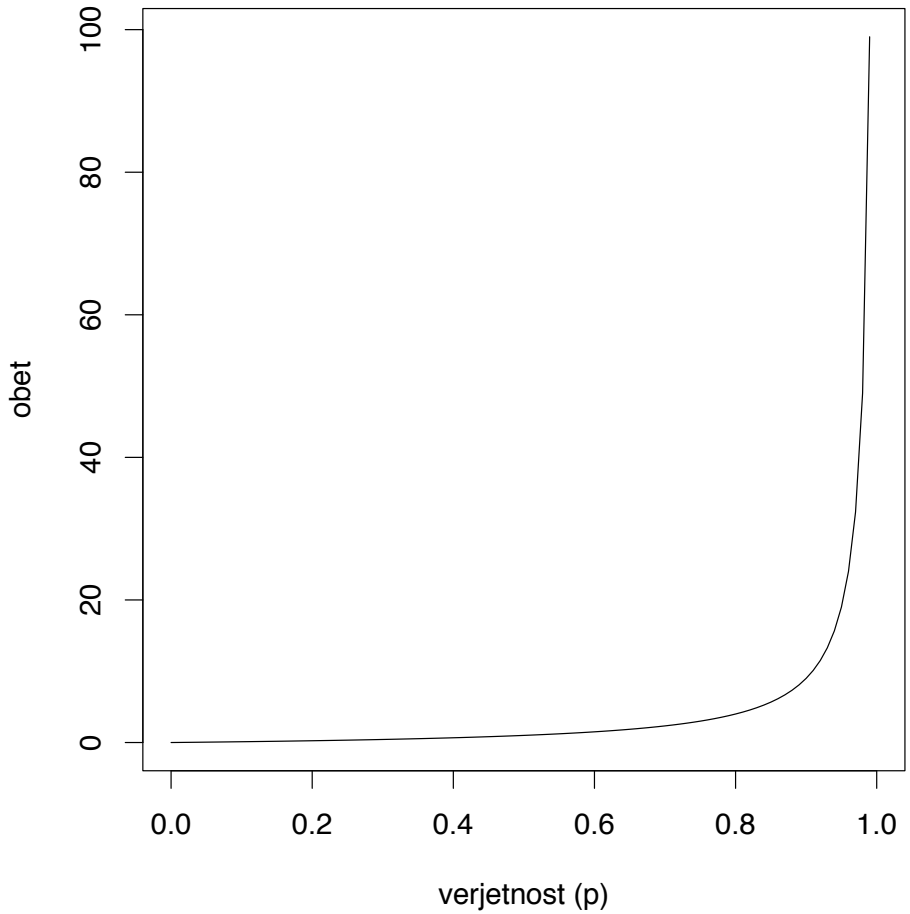
$$obet(A) = \frac{p(A)}{1 - p(A)}$$

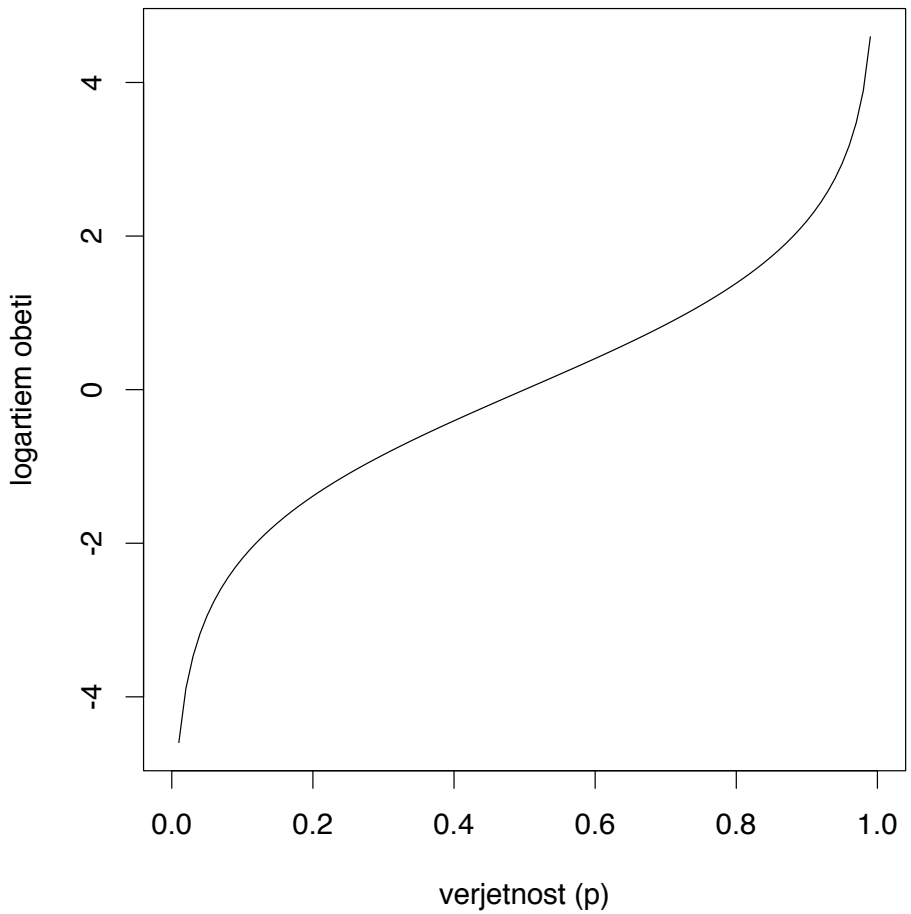
- Kolikokrat je verjetnost A večja od verjetnosti, da se A ne zgodi
- V obliki razmerja pogosta uporaba pri stavah (npr. 1:10, 1:100)
- Domena možnih vrednosti $\mathbb{R}_0^+$
- Regresijski model ne more zagotoviti nenegativnih vrednosti

Zato: logaritem obetov

Domena možnih vrednosti $\mathbb{R}$

Obet in logaritem obetov





Kako do vrednosti koeficientov?

Metoda največjega verjetja (*maximum likelihood*) L

$$L(\boldsymbol{X}, Y; \boldsymbol{\beta}) = \prod_{(\boldsymbol{x},y)\in S} p(Y = y|\boldsymbol{X} = \boldsymbol{x}; \beta)$$

$$= \prod_{(\boldsymbol{x},y)\in S:y=1} \frac{e^{\boldsymbol{\beta}^T \boldsymbol{x}}}{1 - e^{\boldsymbol{\beta}^T \boldsymbol{x}}} \prod_{(\boldsymbol{x},y)\in S:y=0} \frac{1}{1 - e^{\boldsymbol{\beta}^T \boldsymbol{x}}}$$

Kako do vrednosti koeficientov?

Logaritem verjetja

$$\begin{aligned}
\log L &= \sum_{(\boldsymbol{x},y)\in S: y=1} (\boldsymbol{\beta}^T \boldsymbol{x} - \log(1 + e^{\boldsymbol{\beta}^T \boldsymbol{x}})) \quad [\cdot y] \\
&+ \sum_{(\boldsymbol{x},y)\in S: y=0} - \log(1 + e^{\boldsymbol{\beta}^T \boldsymbol{x}}) \quad [\cdot (1-y)] \\
&= \sum_{(\boldsymbol{x},y)\in S} (\boldsymbol{\beta}^T \boldsymbol{x} y - \log(1 + e^{\boldsymbol{\beta}^T \boldsymbol{x}}))
\end{aligned}$$

Iščemo rešitev

$$\max_{\boldsymbol{\beta}} \sum_{(\boldsymbol{x},y)\in S} (\boldsymbol{\beta}^T \boldsymbol{x} y - \log(1 + e^{\boldsymbol{\beta}^T \boldsymbol{x}}))$$

Razlike med linearno in logistično regresijo

Drugačna interpretacija koeficientov

Če se vrednost X_j spremeni za eno enoto, se

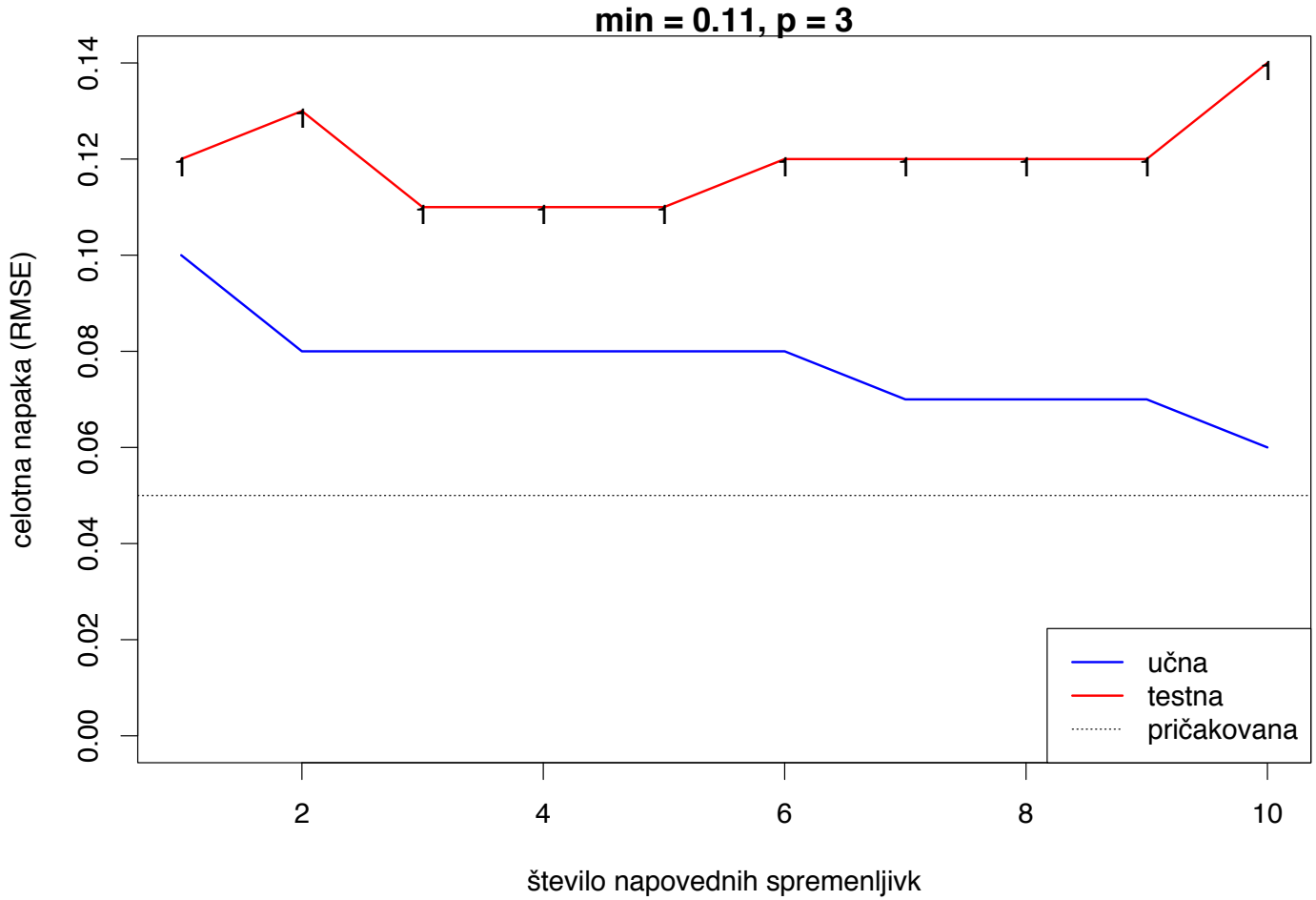
- logaritem obeti za napoved Y spremeni za β_j oz.
- obet za napoved Y se spremeni za e^{β_j} krat

Pozor: sprememba **verjetnosti** napovedi Y je odvisna od trenutne vrednosti X_j, še vedno predznak β_j nakazuje smer spremembe.

Nelinearna optimizacija

- $\log L(X, Y; b)$ je nelinearna funkcija glede na $\boldsymbol{\beta}$
- Izpeljava intervalov zaupanja in statistične značilnosti bolj zapletena

Katere spremenljivke so pomembne?



$Y = I((1 + X_1 + X_1X_2)/3 \geq 0.5), D_i = [0, 1], i = 1..10, D_Y = \{0, 1\}$
zamenjamo vrednost Y 0.05 naključno izbranim primerom

Izbira napovednih spremenljivk in težave

Problem hujši kot pri linearni regresiji

Uporabljamo splošne metode za izbiro spremenljivk, ki jih bomo obravnavali proti koncu semestra.

Ostale težave in njih rešitve enake kot pri linearni regresiji

- Obravnava kvalitativnih spremenljivk s pretvorbo v numerične
- Soodvisnost vhodnih spremenljivk povzroča probleme
- Nelinearni vpliv spremenljivk X na Y
- Povezava/korelacija med reziduali modela za posamezne primere

Najbližji sosedi

Soseščina $N_k(\boldsymbol{x}_0) \subseteq S$ točke $\boldsymbol{x}_0$

$$|N_k(\boldsymbol{x}_0)| = k,\ \forall((\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2)) \in N_k(\boldsymbol{x}_0) \times S : d(\boldsymbol{x}_1, \boldsymbol{x}_0) \leq d(\boldsymbol{x}_2, \boldsymbol{x}_0)$$

je množica k primerov iz S, ki so najbližji sosedi točke $\boldsymbol{x}_0$,
$d : \bigtimes_{i=1}^{p} D_i \times \bigtimes_{i=1}^{p} D_i \to \mathbb{R}_0^+$ je mera razdalje med primeri

Napovedni model

- Regresija: $\hat{y}_0 = \frac{1}{k} \sum_{(\boldsymbol{x},y) \in N_k(\boldsymbol{x}_0)} y$
- Klasifikacija: $\hat{y}_0$ vrne vrednost, ki se največkrat pojavi v množici $\{y : (\boldsymbol{x}, y) \in N_k(\boldsymbol{x}_0)\}$

Neparametrična metoda

- Ne zastavi (močnih) predpostavk glede oblike odvisnosti Y od $\boldsymbol{X}$
- Odločitvena meja pri razvrščanju je lahko poljubne oblike

Linearna in logistična regresija so parametrične metode

- Predpostavijo linearno odvisnost Y od $\boldsymbol{X}$
- Odločitvena meja pri klasifikaciji je linearna

Mere razdalje

Razdalja med primeri $d : \bigtimes_{i=1}^{p} D_i \times \bigtimes_{i=1}^{p} D_i \to \mathbb{R}_0^+$

$d(\boldsymbol{x}_1, \boldsymbol{x}_2) = \sum_{i=1}^{p} \delta(x_{1i}, x_{2i})$, kjer x_{ki} predstavlja vrednost X_i za primer x_k

Razdalja med vrednostmi $\delta : D_i \times D_i \to \mathbb{R}_0^+$

Če je X_i numerična, sta običajni izbiri

- Evklidska razdalja $\delta(v_1, v_2) = (v_1 - v_2)^2$
- Manhatnska razdalja $\delta(v_1, v_2) = |v_1 - v_2|$

Če je X_i diskretna, je običajna izbira $\delta(v_1, v_2) = I(v_1 \neq v_2)$

Normalizacija

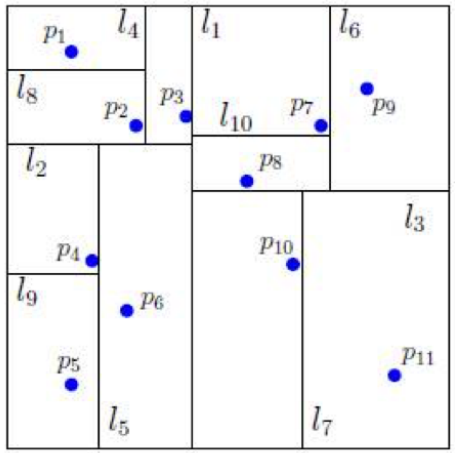
Pred uporabo metode najbližjih sosedov običajno normaliziramo (ali standardiziramo) vrednosti napovednih spremenljivk.

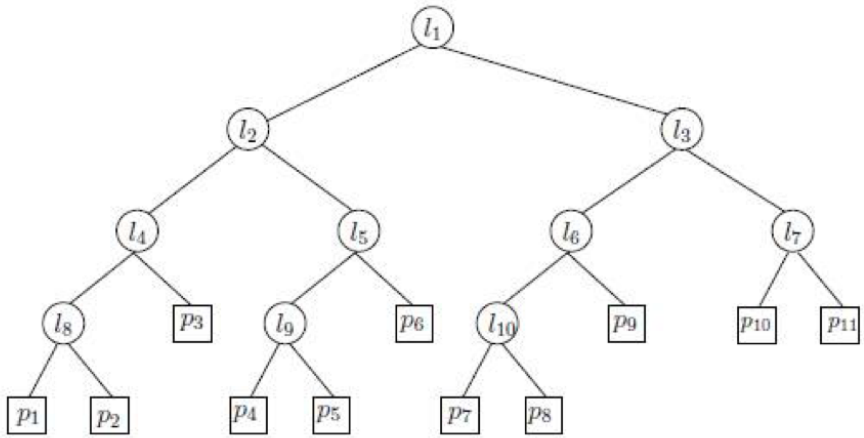
Učinkovito iskanje najbližjih sosedov

Iskanje najbližjih sosedov računsko zahteven problem

Vsaka napoved zahteva $O(pn)$ časa: lahko težavno pri velikem n

Učinkovita rešitev $O(p \log n)$: podatkovna struktura drevo kD

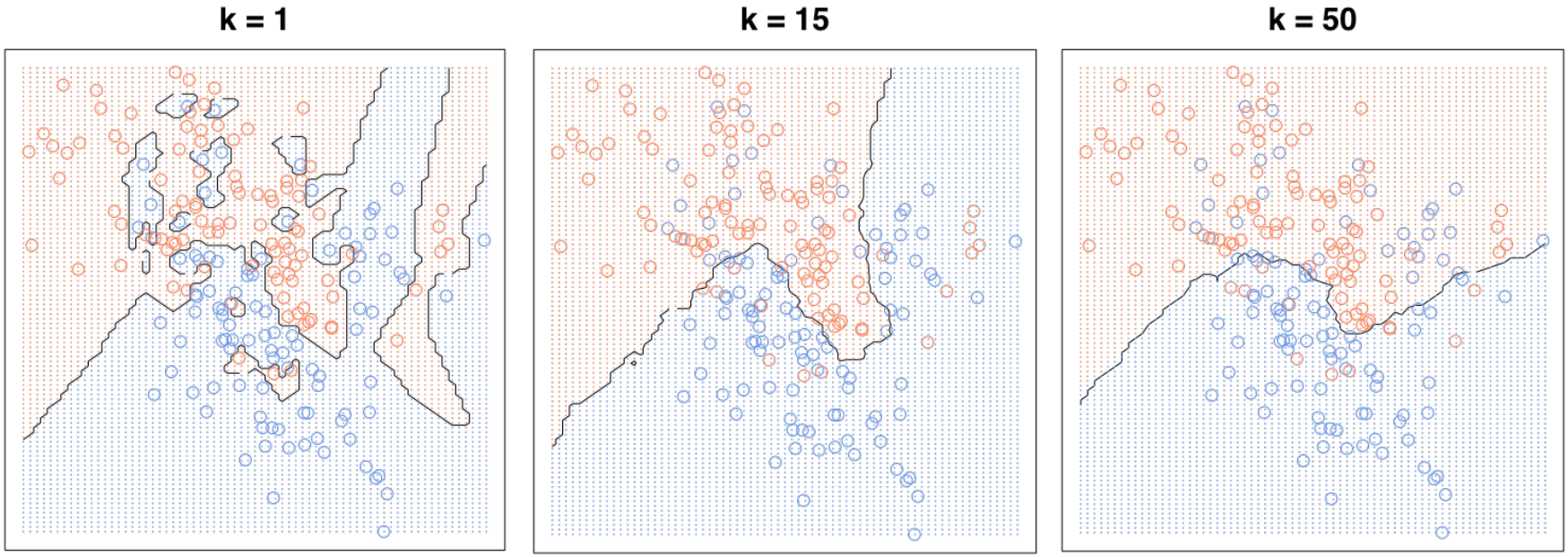




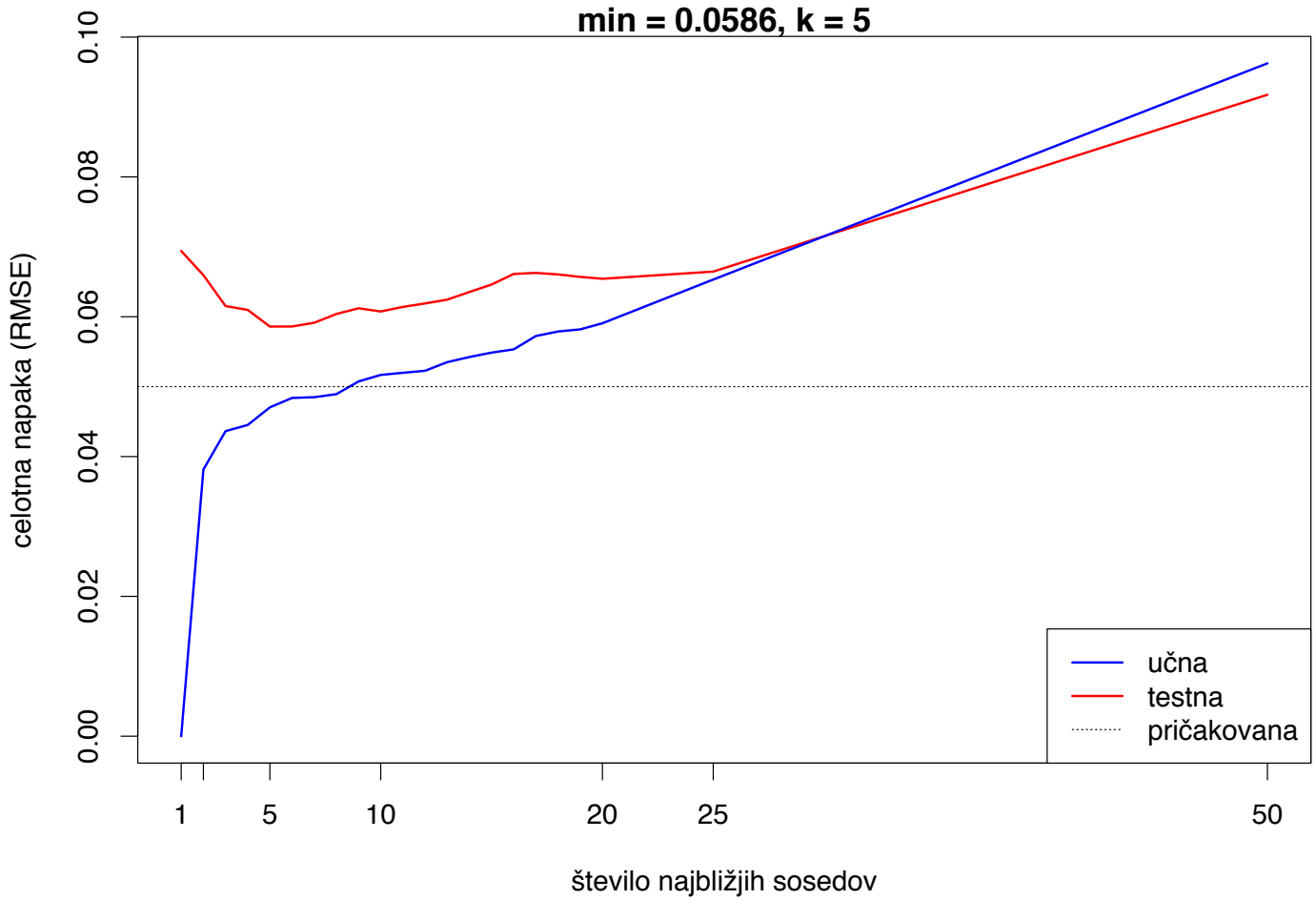
Vpliv parametra k

Število sosedov k vpliva na prilagodljivost modela

- Pri $k = 1$ se model popolnoma prilagodi učni množici
- Za večje vrednosti k se model vedno manj prilagaja učni množici

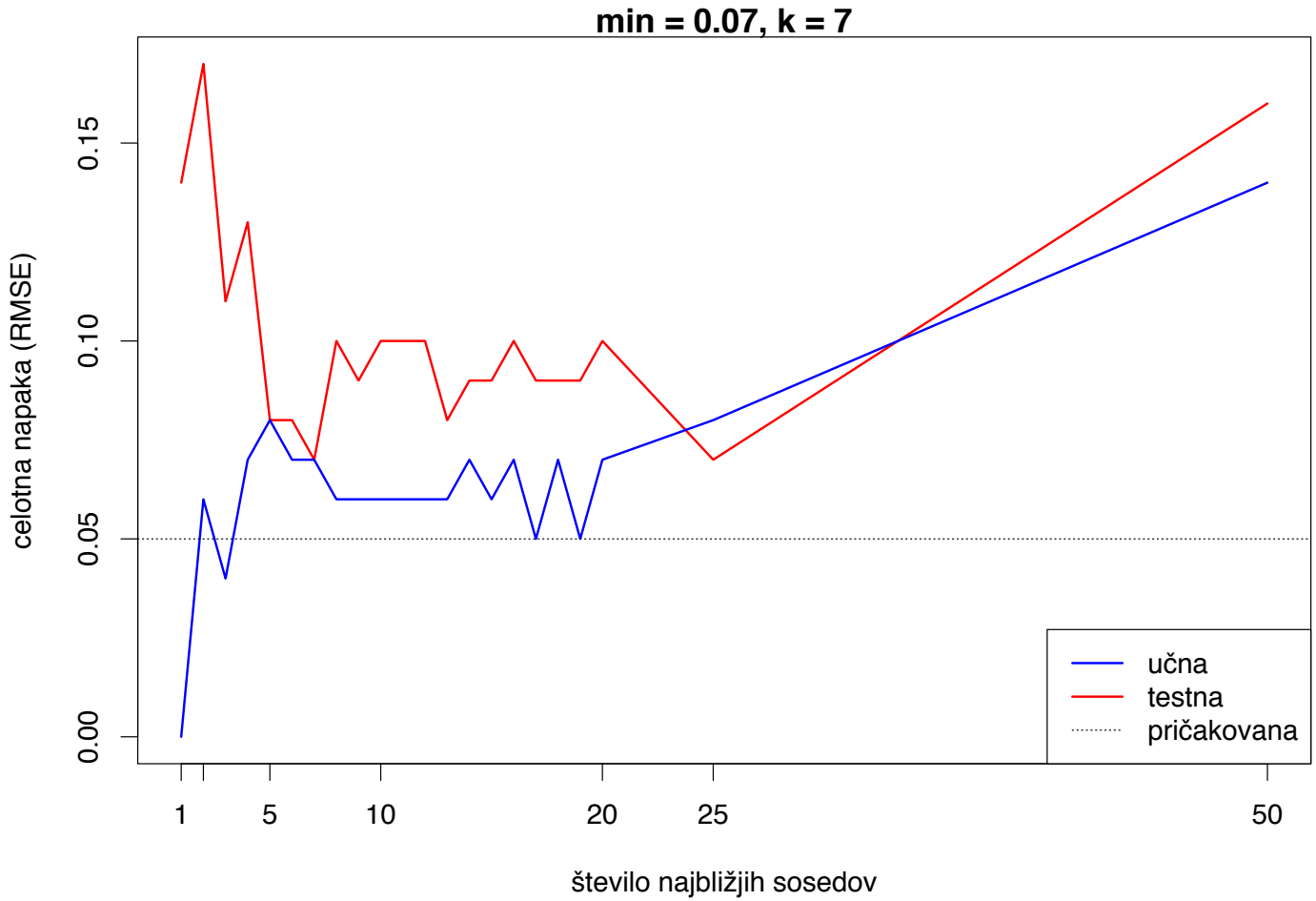


Vpliv k: regresijski model (X_1 in X_2)



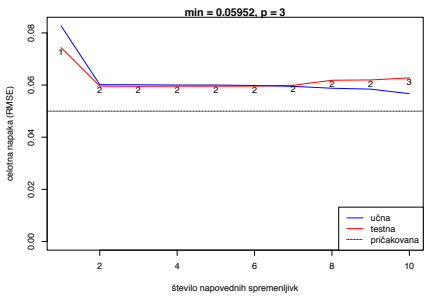
$$Y = (1 + X_1 + X_1X_2)/3 + \mathcal{N}(0, 0.05),\ D_i = [0, 1], i = 1..10, D_Y = [0, 1]$$

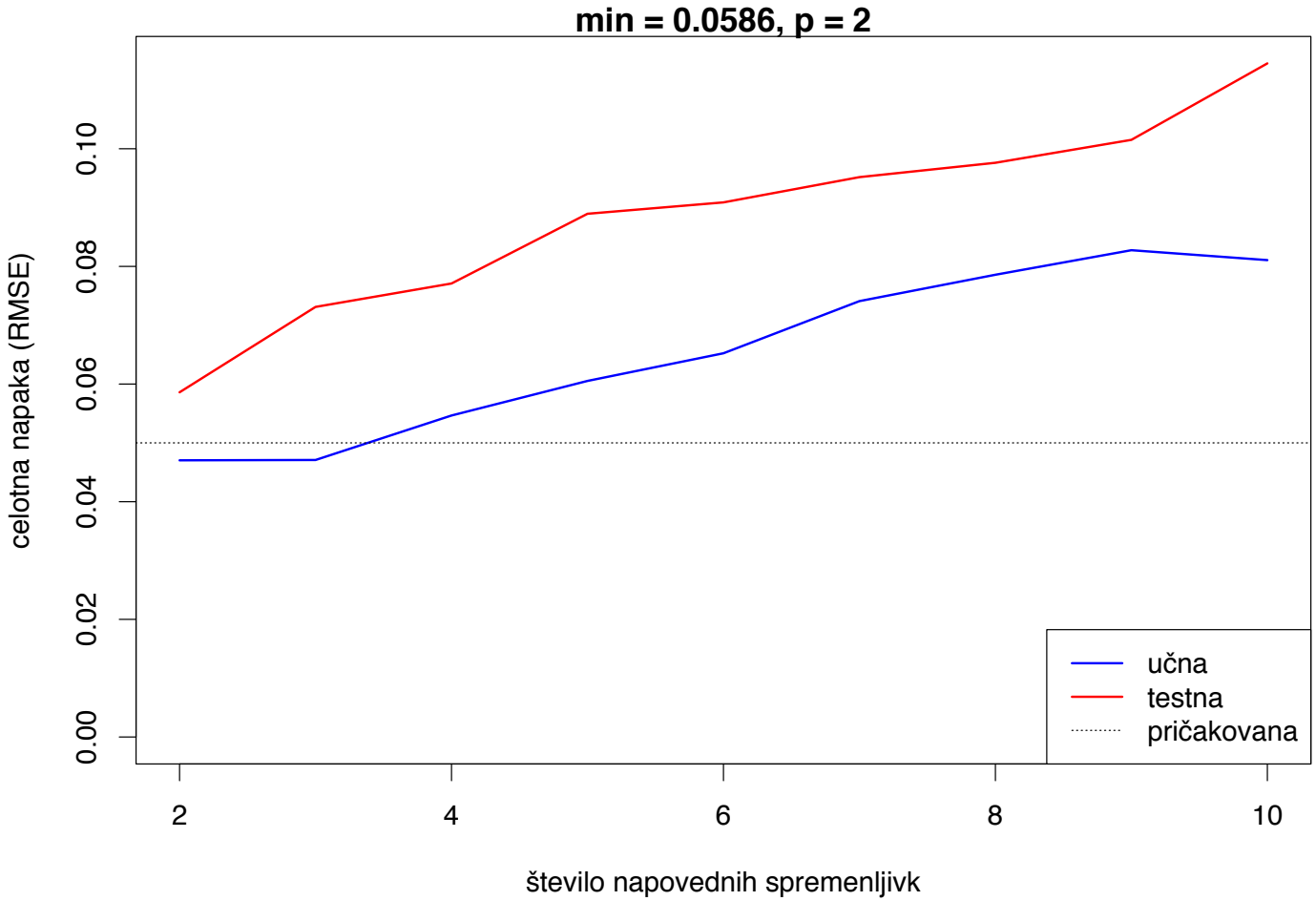
Vpliv k: klasifikacijski model (X_1 in X_2)



$Y = I((1 + X_1 + X_1X_2)/3 \geq 0.5), D_i = [0, 1], i = 1..10, D_Y = \{0, 1\}$
zamenjamo vrednost Y 0.05 naključno izbranim primerom

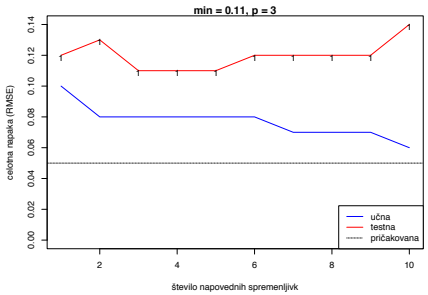
Vpliv števila napovednih spremenljivk: regresija $k = 5$

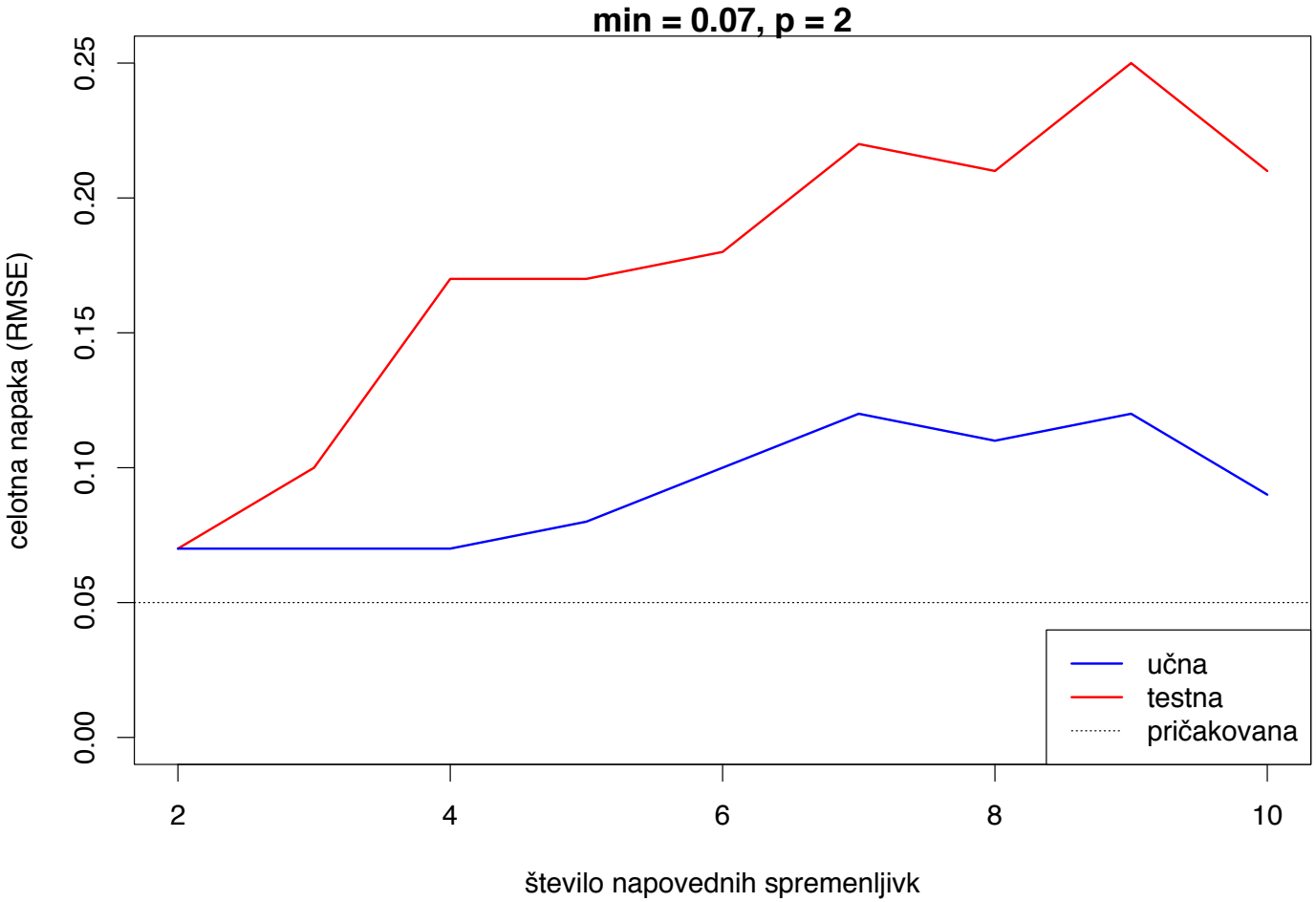




$$Y = (1 + X_1 + X_1X_2)/3 + \mathcal{N}(0, 0.05),\ D_i = [0, 1], i = 1..10, D_Y = [0, 1]$$

Vpliv števila napovednih spremenljivk: klasifikacija $k = 7$



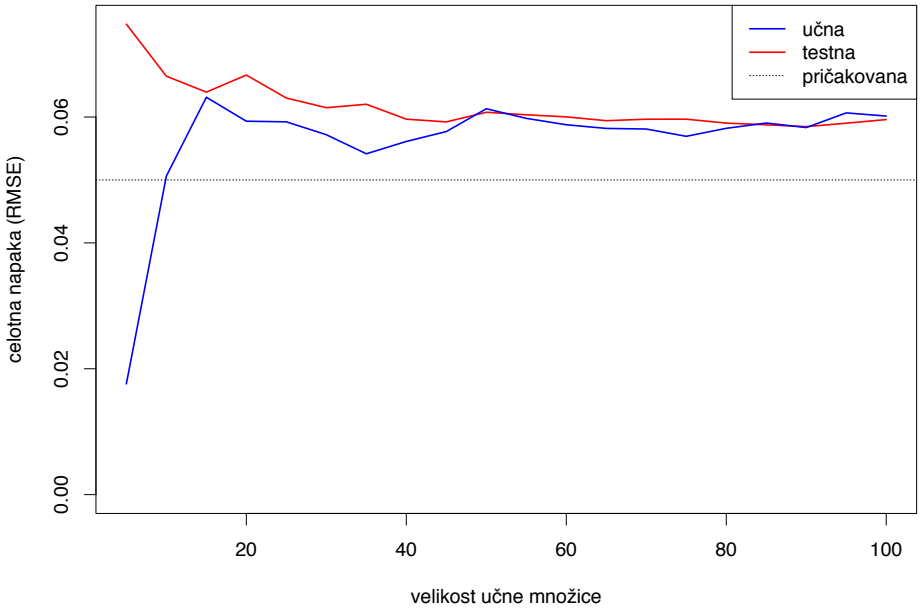


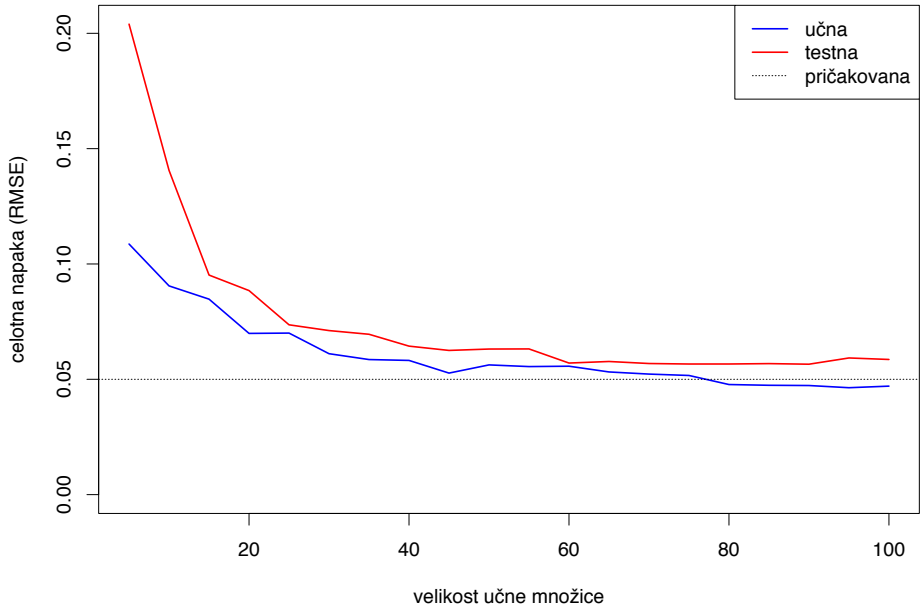
$Y = I((1 + X_1 + X_1X_2)/3 \geq 0.5), D_i = [0, 1], i = 1..10, D_Y = \{0, 1\}$

zamenjamo vrednost Y 0.05 naključno izbranim primerom

Primerjava enostavnih modelov: regresija

Skoraj identični napaki: RMSE(LM) = 0.060, RMSE(kNN) = 0.059

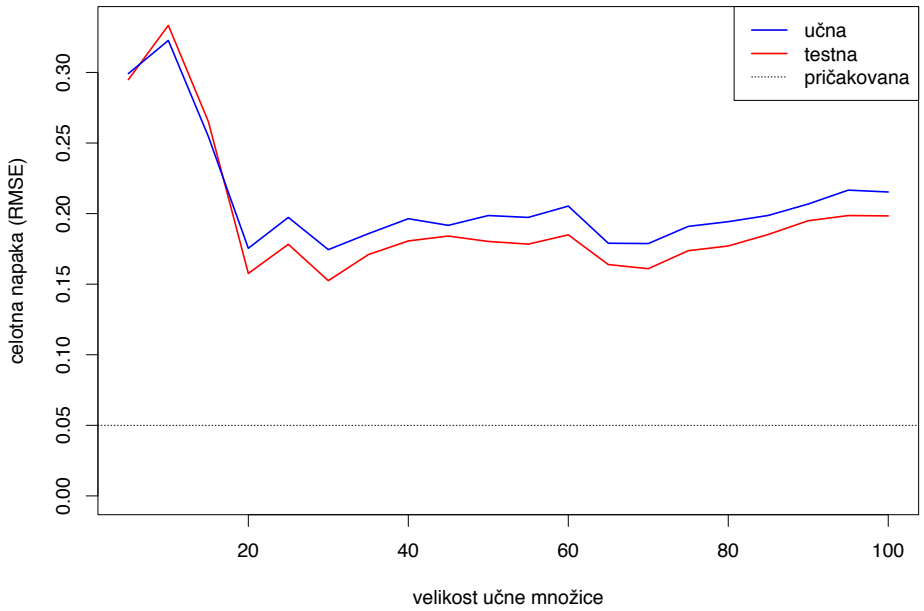


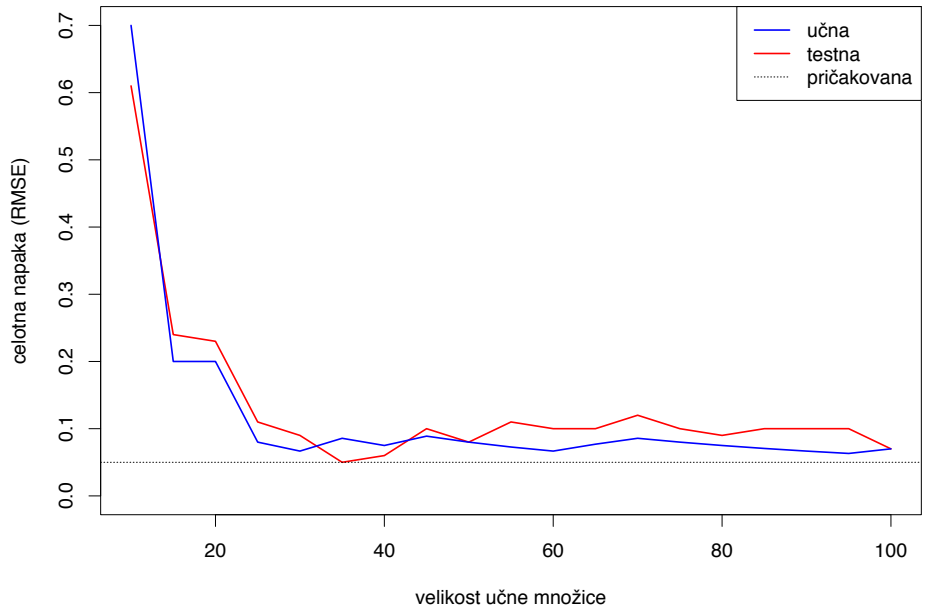


$$Y = (1 + X_1 + X_1X_2)/3 + \mathcal{N}(0, 0.05),\ D_i = [0, 1], i = 1..10, D_Y = [0, 1]$$

Primerjava enostavnih modelov: klasifikacija

Model najbližjih sosedov boljši: Err(LM) = 13%, Err(kNN) = 7%





$Y = I((1 + X_1 + X_1X_2)/3 \geq 0.5), D_i = [0, 1], i = 1..10, D_Y = \{0, 1\}$
zamenjamo vrednost Y 0.05 naključno izbranim primerom