Uvod v strojno učenje in podatkovno rudarjenje

Ljupčo Todorovski
Univerza v Ljubljani, Fakulteta za upravo

Februar 2018

John Snow, London 1854

Vzorci

Skupine okužb

Smrtni primeri porazdeljeni **neenakomerno**

Hipoteze in modeli

Bližina vodnih virov

Skupine smrtnih primerov so v bližini vodne črpalke na ulici Broad (rdeči krog)

Terminologija strojnega učenja in podatkovnega rudarjenja

Podatki

- Seznam smrtnih primerov (pacientov dr. Snowa) in njihove lokacije
- Lokacije vodnih črpalk

Vzorci

Skupine smrtnih primerov v nekaj delih mesta

Hipoteza

Skupine primerov lokacijsko povezane z vodnimi črpalkami

Terminologija strojnega učenja in podatkovnega rudarjenja

Napovedni model

- Smrtni primeri najbolj pogosti v bližini vodne črpalke na ulici Broad
- Vhodni podatek je lokacija bivališča, napoved verjetnost preživetja

Odločitev

Zapreti vodno črpalko na ulici Broad

Cilji strojnega učenja in podatkovnega rudarjenja

Strojna analiza podatkov usmerjena v

1. Iskanje **zanimivih** vzorcev v podatkih
2. Gradnjo **točnih** napovednih modelov

Zanimivost in točnost pomembni lastnosti modelov in vzorcev

Med semestrom bomo spoznali različne metode
za vrednotenje oziroma računanje
zanimivosti vzorcev in **točnosti** modelov.

Uporabnost vzorcev in modelov

Boljše razumevanje podatkov

- Kratka in nazorna predstavitev podatkov
- Vzpostavitev lokacijske povezave med okužbami in vodnimi črpalkami

Pojasnjevanje in napovedovanje

- Voda kot vir oziroma medij za prenašanje bolezni
- Uporabniki vode iz črpalke na ulici Broad bodo zboleli

Odločanje in ukrepanje

- Zapiranje vodne črpalke na ulici Broad za zajezitev epidemije
- Pasterizacija vode za preprečevanje izbruha novih epidemij

Pregled vsebine

Formalne definicije

- Podatki, modeli in vzorci (*patterns*)
- Naloge strojnega učenja

Problemi strojnega učenja

- Prekletstvo dimenzionalnosti (*curse of dimensionality*)
- Pretirano in nezadostno prileganje (*overfitting* in *underfitting*)

Pravila igre

Izvedba semestra, obveznosti in ocenjevanje

Spremenljivke

Napovedne (vhodne, ~~neodvisne~~) spremenljivke (atributi) $X_i, i = 1..p$

- Urejena p-terica $\boldsymbol{X} = (X_1, X_2, \ldots X_p)$
- Zaloge vrednosti D_1, D_2, $\ldots D_p$
  - X_i je numerična (~~zvezna~~, kvantitativna), če $D_i \subseteq \mathbb{R}$
  - X_i je diskretna (kvalitativna), če je D_i **končna** in običajno **neurejena**

Ciljna (izhodna, ~~odvisna~~) spremenljivka Y

Zaloga vrednosti D_Y

Primeri

Nenadzorovano učenje

$e \in \bigtimes_{i=1}^{p} D_i$ oziroma $e = \boldsymbol{x} = (x_1, x_2, \ldots x_p), x_i \in D_i$

Nadzorovano učenje

$e \in \bigtimes_{i=1}^{p} D_i \times D_Y$ oziroma $e = (\boldsymbol{x}, y) = (x_1, x_2, \ldots x_p, y), x_i \in D_i, y \in D_Y$

Podatkovna množica $S \subseteq \mathcal{E}$

$\mathcal{E}$ označuje množico vseh možnih primerov e: $\bigtimes_{i=1}^{p} D_i$ oz. $\bigtimes_{i=1}^{p} D_i \times D_Y$

Opomba o notaciji

$$\bigtimes_{i=1}^{p} D_i = D_1 \times D_2 \times \ldots \times D_p$$

Ilustrativni primer: Kartica zvestobe

Primeri (vrstice) so kupci, spremenljivke (stolpci) lastnosti kupcev

Ime	Prihodki	Starost	Spol	Letna poraba	Dober kupec
Mojca	1,890	32	Ž	18,200	da
Janez	1,200	48	M	8,900	ne
Špela	900	63	Ž	9,200	da

Primeri so nakupi, spremenljivke kupec in produkti

Kupec	Spol	Pivo	Plenice	Voda	Kruh	Čokolada
Mojca	Ž	0	0	0	2	3
Janez	M	2	2	0	1	0

Modeli in vzorci

Modeli so funkcije

$$m : \bigtimes_{i=1}^{p} D_i \rightarrow D_Y$$

Za podane vrednosti neodvisnih spremenljivk $\boldsymbol{x} = (x_1, x_2, \ldots x_p)$ vrne model m ocenjeno (napovedano) vrednost ciljne spremenljivke
$\hat{y} = m(x_1, x_2, \ldots x_p) = m(\boldsymbol{x})$

Vzorci so običajno množice

- primerov, ki so si **podobni**
- vrednosti spremenljivk, ki se **pogosto pojavljajo skupaj**

Ilustrativni primer: Kartica zvestobe

Modeli

- $\textit{Dober kupec} = m(\textit{Starost}, \textit{Prihdoki}, \textit{Spol})$
- $\textit{Letna poroaba} = m(\textit{Starost}, \textit{Spol})$

Vzorci

- Skupine kupcev s podobnimi nakupi
- Moški ki kupujejo plenice, kupujejo tudi pivo

Nadzorovano strojno učenje

Definicija naloge

- Na osnovi podane učne podatkovne množice S_{train}
- Najdi model m, ki je **točen** in **splošno veljaven**

Točen in splošno veljaven model

Točen model doseže **minimalno napako** (maksimalno točnost) na učni množici S_{train}, splošno veljaven pa doseže **majhne napake** na poljubni podatkovni množici S.

Vrednotenje točnosti modela

Napaka na enem primeru, funkcija izgube $L : D_Y \times D_Y \to \mathbb{R}_0^+$

Vrne razliko $L(y, \hat{y})$ med opazovano (y) in napovedano $(\hat{y})$ vrednostjo ciljne spremenljivke Y na enem primeru.

Napaka na podatkovni množici $Err : (\bigtimes_{i=1}^{p} D_i \to D_Y) \times \mathcal{P}(\mathcal{E}) \to \mathbb{R}_0^+$

$$Err(m, S) = \frac{1}{|S|} \sum_{(\boldsymbol{x}, y) \in S} L(y, m(\boldsymbol{x}))$$

Vrne povprečno vrednost funkcije izgube podanega modela m na primerih iz podatkovne množice S.

Regresija: $D_Y \subseteq \mathbb{R}$

Običajna funkcija izgube je **kvadratna napaka**

$$L_{SE}(y, \hat{y}) = (y - \hat{y})^2$$

Napaka modela

S to funkcijo izgube vrednotimo srednjo kvadratno napako ($MSE = Err$) modela. Pogosto računamo tudi celotno napako $RMSE = \sqrt{MSE}$.

Razvrščanje: D_Y je končna, običajno neurejena množica

Običajna funkcija izgube

$$L_{01}(y, \hat{y}) = \begin{cases} 1; & y \neq \hat{y} \\ 0; & y = \hat{y} \end{cases}$$

Napaka modela

S to funkcijo izgube vrednotimo klasifikacijsko napako ($Err \in [0, 1]$) modela, običajno izraženo v odstotkih. Pogosto računamo tudi klasifikacijsko točnost Acc, ki je $1 - Err$.

Poseben primer: dvojiško razvrščanje ali binarna klasifikacija

V primeru, ko je $|D_Y| = 2$.

Algoritem za strojno učenje ali metoda strojnega učenja

Nadzorovano učenje (napovedno modeliranje)

$$\mathcal{A} : \mathcal{P}(\mathcal{E}) \rightarrow (\bigtimes_{i=1}^{p} D_i \rightarrow D_Y)$$

Na osnovi podane učne množice S_{train}, algoritem vrne model m za ocenjevanje (napovedovanje) vrednosti ciljne spremenljivke Y iz podanih vrednosti napovednih spremenljivk $X_1, X_2, \ldots X_p$, t.j., $\mathcal{A}(S_{train}) = m$.

Nenadzorovano učenje

Na osnovi podane učne množice vrne vzorce t.j. množice

- podobnih primerov
- vrednosti spremenljivk, ki se pogosto pojavljajo skupaj

Optimalen napovedni model

$m^*(\boldsymbol{x}) = E[Y|\boldsymbol{X} = \boldsymbol{x}]$ **minimizira** srednjo kv. napako $E[(Y - m(\boldsymbol{X}))^2]$

Ocena (približek) na učni množici S_{train}

$$m^*(\boldsymbol{x_0}) = \frac{1}{|S_0|} \sum_{(\boldsymbol{x},y)\in S_0} y, \quad S_0 = \{(\boldsymbol{x}, y) \in S_{train} : \boldsymbol{x} = \boldsymbol{x_0}\}$$

Problem točnosti ocene

V učni množici S_{train} je običajno premalo število primerov za katere velja $\boldsymbol{x} = \boldsymbol{x_0}$, t.j., je množica S_0 premajhna za dobro oceno y.

Rešitev: metoda najbližjih sosedov

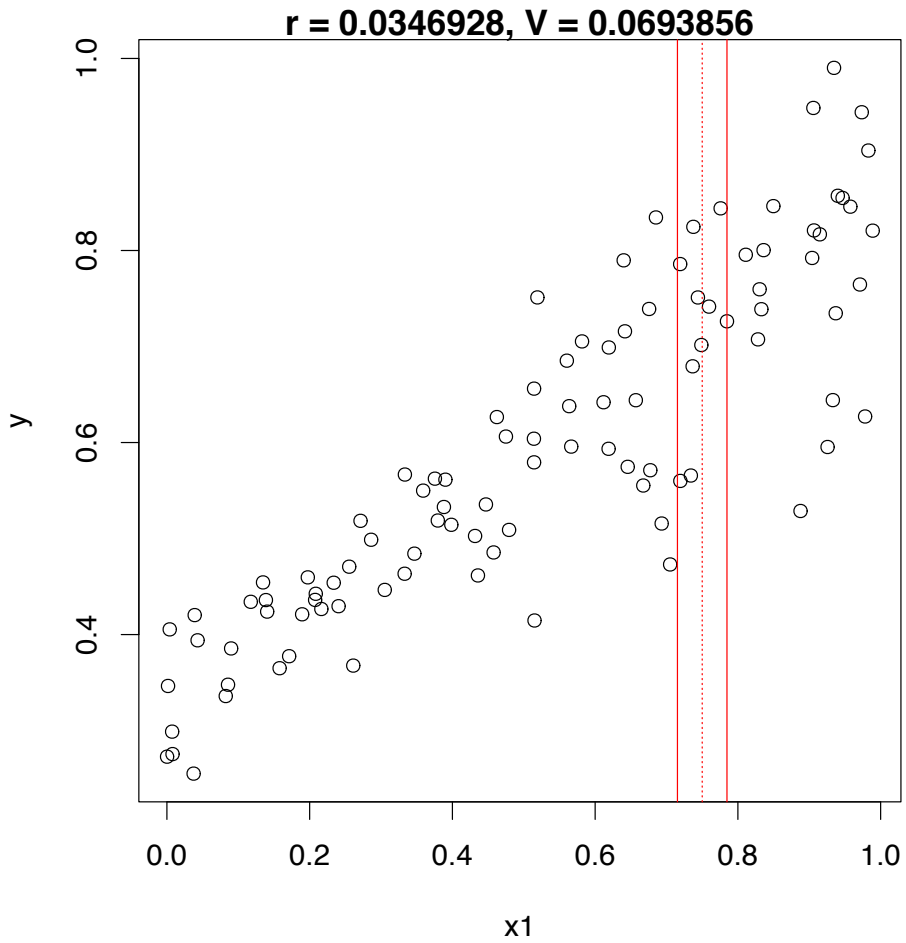
Spremenimo definicijo množice S_0

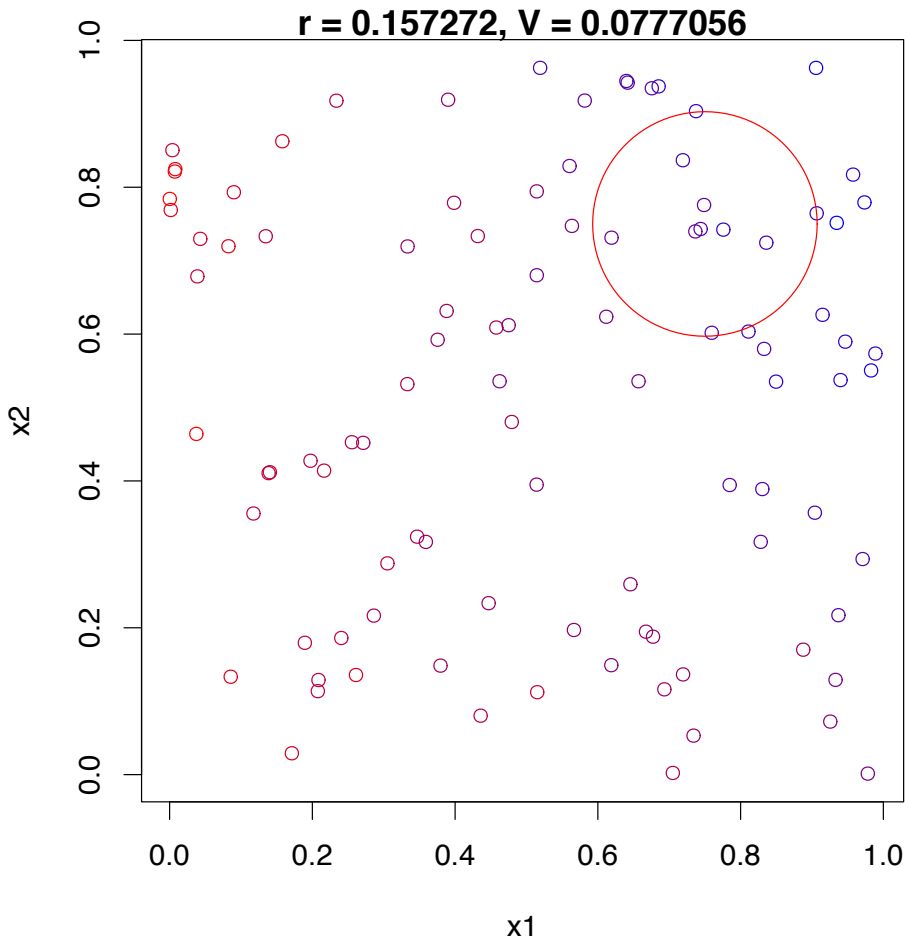
Tako, da vsebuje k primerov, ki so najbližji (imajo najmanjše Evklidske razdalje) primeru $\boldsymbol{x_0}$.

Koliko blizu so najbližji sosedi?

Če izberemo k najbližjih sosedov v S_0, nas potem zanima kako blizu so izbrani primeri izhodiščnemu primeru $\boldsymbol{x_0}$ oziroma kakšen je polmer in prostornina množice S_0.

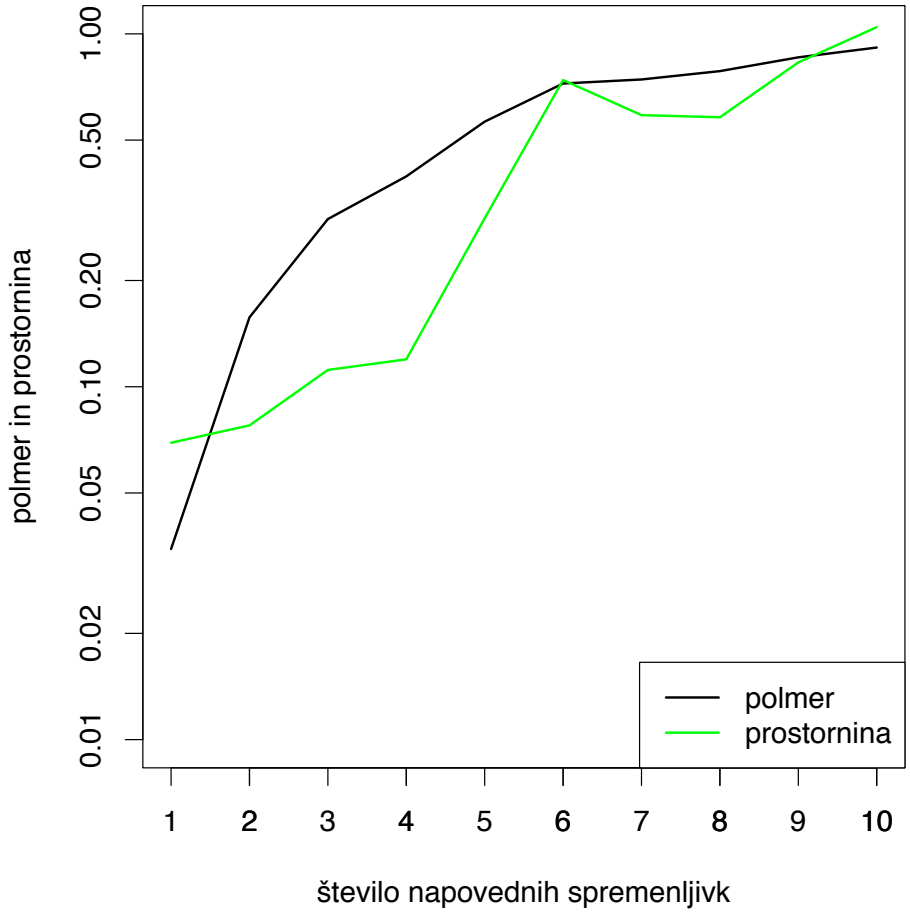
Polmer za $k = 10$ pri eni in dveh spremenljivkah

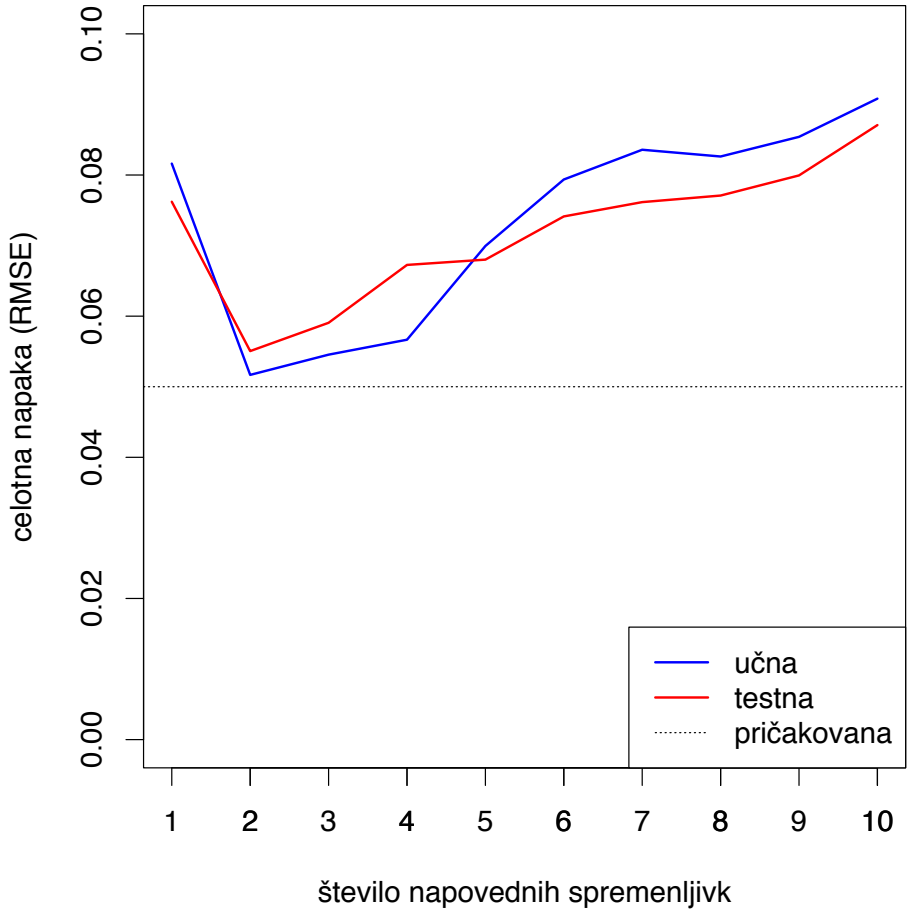




$$Y = (1 + X_1 + X_1X_2)/3 + \mathcal{N}(0, 0.05),\ D_i = [0, 1], i = 1..10, D_Y = [0, 1]$$

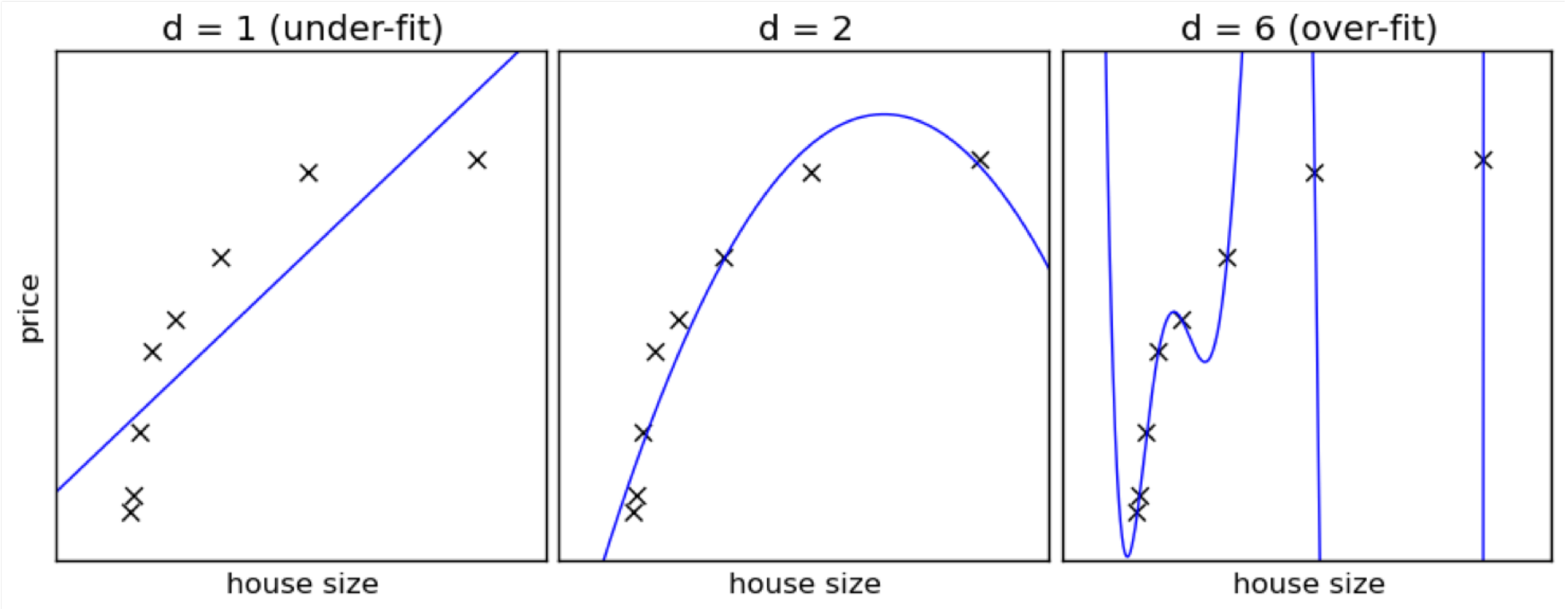
Odvisnost polmera od števila spremenljivk ($k = 10$)



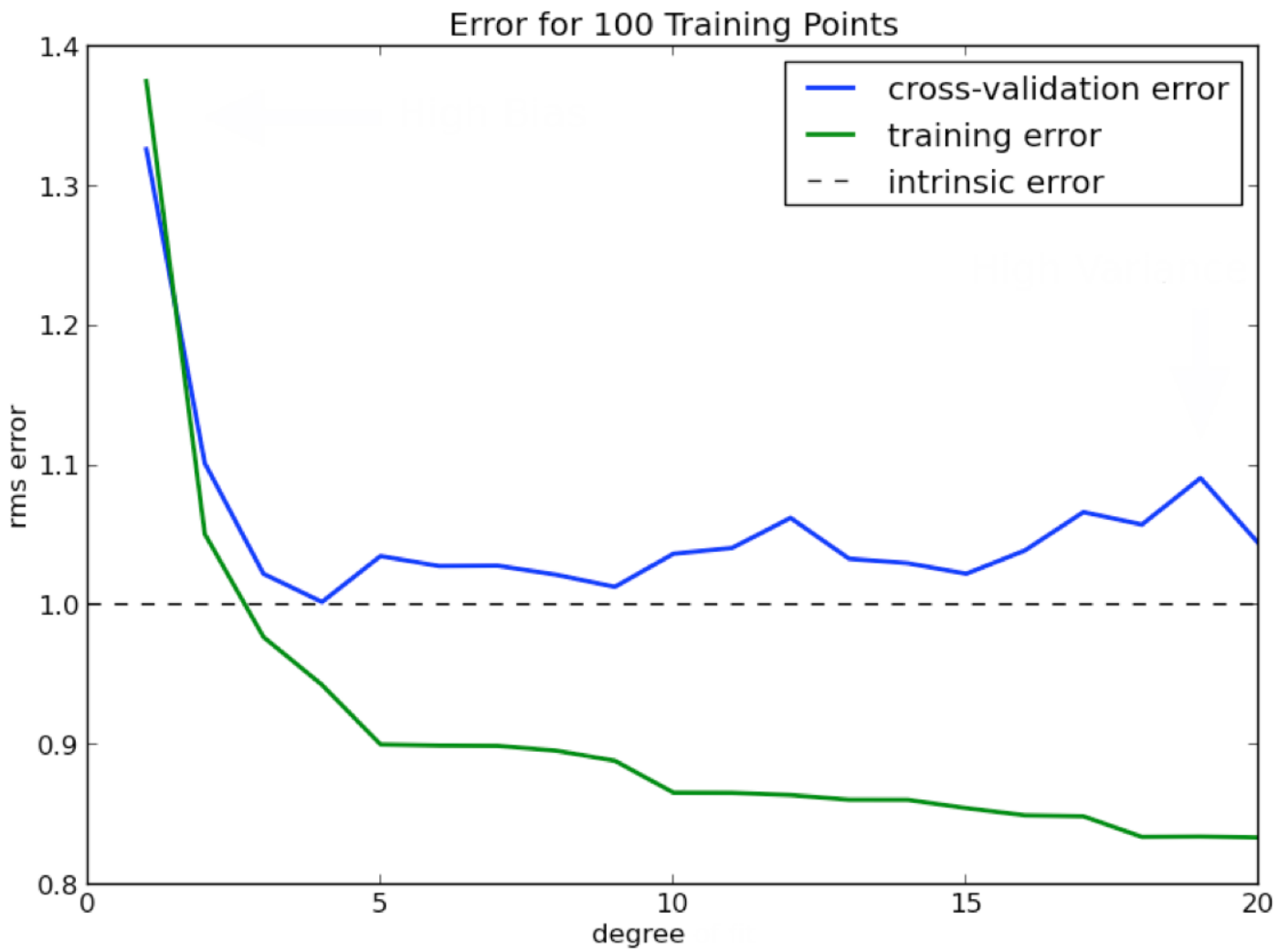


$$Y = (1 + X_1 + X_1X_2)/3 + \mathcal{N}(0, 0.05)$$

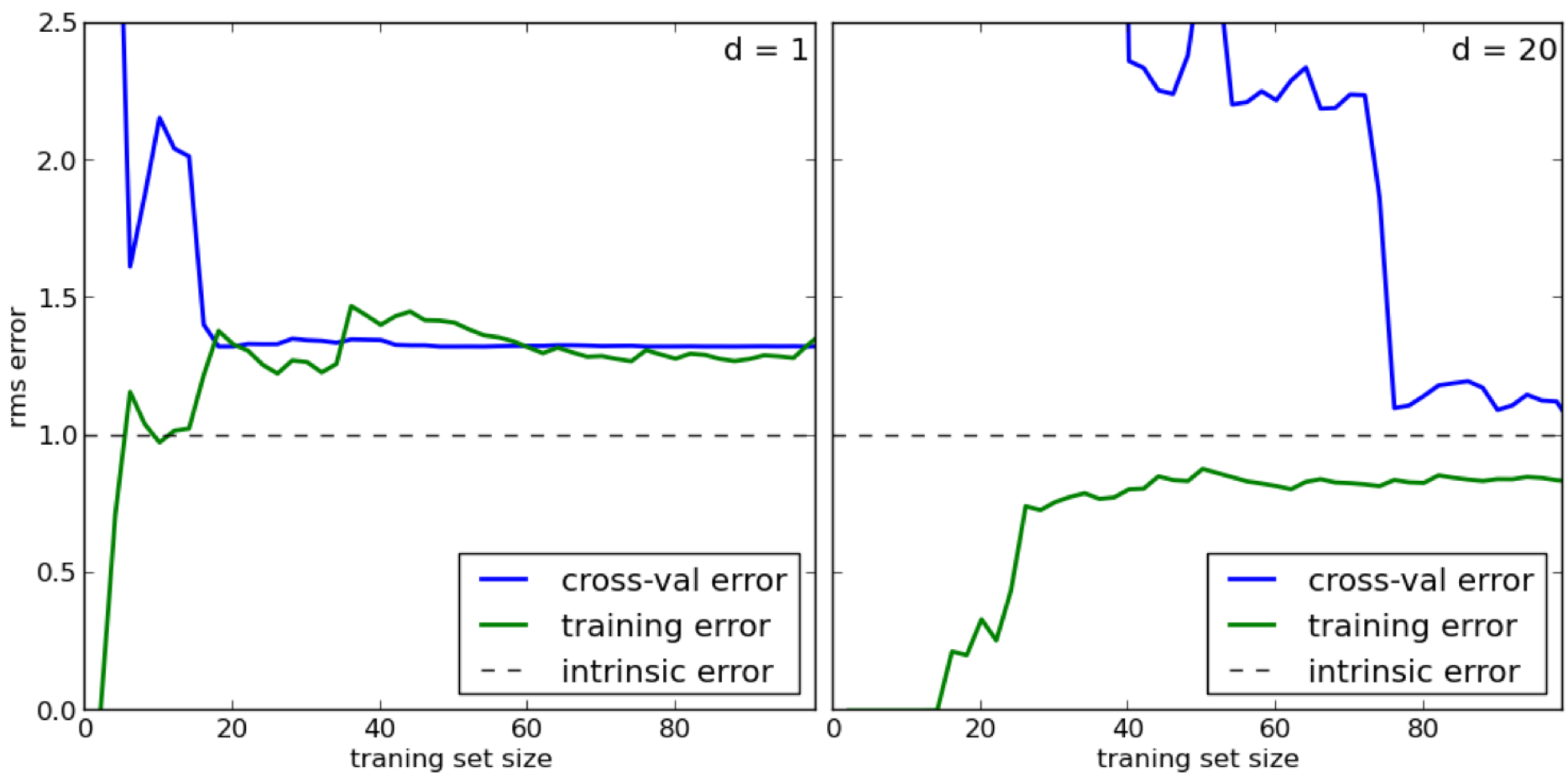
Polinomska regresija z eno spremenljivko



Učna in testna napaka



Učna in testna napaka: število primerov



Ključna vprašanja in problemi

Model z majhno napako ali enostaven model?

- Skrajnosti: enostaven linearen model ALI globoka nevronska mreža
- Kateri model izbrati?

Pretirano ali nezadostno prileganje?

- Skrajnosti: preveč enostaven model z veliko učno/testno napako ALI prezapleten model z majhno učno napako
- Dober model je nekje vmes: kako ga najti?

Model prozorne ali črne škatle?

- Skrajnosti: enostaven model z nekaj spremenljivkami ALI zapleten model z veliko dimenzijami
- Kako se izogniti prekletstvu večdimenzionalnosti?

Cilji in pridobljene kompetence

Razumevanje strojnega učenja in podatkovnega rudarjenja

- Razumevanje delovanja metod za strojno učenje
- Prepoznavanje prednosti in slabosti posamezne metode
- Nastavljanje parametrov metode za strojno učenje

Reševanje problemov s podatkovnim rudarjenjem

- Zanašati se le na napovedno napako ali izbrati bolj enostaven model?
- Kako se izogniti pretiranemu in nezadostnemu prilagajanju?
- Kako se izogniti prekletstvu večdimenzionalnosti?

Teorija: metode strojnega učenja

- Linearni modeli in metode najbližjih sosedov
- Merjenje napake in izbira napovednega modela
- Posebnosti binarne klasifikacije
- Odločitvena drevesa in pravila
- Metode podpornih vektorjev in jedrne funkcije
- Nevronske mreže
- Ansambli napovednih modelov
- Nenadzorovano učenje

Praksa: podatkovno rudarjenje

- Proces podatkovnega rudarjenja
- Obravnava različnih tipov podatkov in vložitve
- Izbira in konstrukcija napovednih spremenljivk
- Obravnava manjkajočih podatkov
- Obravnava neenakomerne porazdelitve vrednosti ciljne spremenljivke
- Napredne teme in vabljena predavanja

Obveznosti in ocenjevanje

Sprotne obveznosti na predavanjih

Pregled prosojnic (spletna učilnica) in samostojni študij literature

Sprotne obveznosti na vajah do 20 odstotnih točk

- Dve domači nalogi, vsaka prinese do 10 odstotnih točk
- Točke pridobite le, če nalogo oddate **pravočasno**

Seminarska naloga do 50 odstotnih točk

Od aprila do konca semestra oziroma ustnega izpita, oddaja **vsaj 5 dni pred** izpitnim rokom

Ustni izpit do 30 odstotnih točk

Uspešno opravljene obveznosti so pogoj za pristop k ustnemu izpitu

R in caret

Prosto-dostopna programska oprema za statistične obdelave

- www.r-project.org
- Uporabljali ga boste na vajah
- Prednost: na voljo imate impresivni nabor orodij za statistične obdelave podatkov in vizualizacijo
- Slabost: krmiljenje skozi ukazno vrstico; potrebno poznavanje programskega jezika

Številni dodatni paketi za strojno učenje in podatkovno rudarjenje

- caret: Classification and Regression Training, primerjava modelov
- randomForest, kernlab, neuralnet: različne metode strojnega učenja
- tm: Text Mining, obravnava besedilnih podatkov

Weka

Odprto-kodna zbirka algoritmov za strojno učenje v Javi

- www.cs.waikato.ac.nz/ml/weka/
- Ne boste ga spoznali na vajah
- Prednost: impresivno število implementiranih algoritmov, neposredna povezljivost s podatkovnimi bazami (JDBC)
- Slabost: počasne implementacije, nenavaden uporabniški vmesnik

Idealno izhodišče za razvoj novih metod

- Zelo hitra in enostavna primerjava z (tako rekoč vsemi) obstoječimi
- Dobra podpora in velika skupnost uporabnikov
- Podprta s knjigo, ki pa ni prosto dostopna (glej naprej)

Orange

Odprto-kodna programska oprema za strojno učenje v Pythonu

- orange.biolab.si, razvoj orodja na UL-FRI
- Ne boste ga spoznali na vajah
- Prednost: vizualno programiranje delotokov za obdelavo podatkov od priprave do gradnje in vrednotenja ter primerjave modelov
- Slabost: relativno omejen nabor razpoložljivih metod, ni možnosti nalaganja zunanjih implementacij algoritmov

Manjka nekaj pomembnih metod

- Odločitvena pravila
- Nevronske mreže
- Obravnava besedil

Vodič po literaturi

An Introduction to Statistical Learning with Applications in R

- Prosti dostop: www-bcf.usc.edu/~gareth/ISL/
- Iz spletne strani dostop do spletnih tečajev, priporočam tistega od Hastie in Tibshirani
- Obvezno branje: večina snovi predmeta pokrita s tem učbenikom

The Elements of Statistical Learning

- Prosti dostop: statweb.stanford.edu/~tibs/ElemStatLearn/
- Za tiste, ki imate radi teorijo
- Kakšna snov bo pokrita le v tem učbeniku, pregled v spletni učilnici

Vodič po literaturi

Data Mining: Practical Machine Learning Tools and Techniques

- Za ljubitelje prakse
- Referenčni vir za izkušene rudarje: lahko na hitro preverite delovanje algoritma strojnega učenja in nastavitve parametrov

Machine Learning: The Art and Science of Algs that Make Sense of Data

Odličen kompromis med teorijo in prakso

Applied Predictive Modeling

- Za eksperte praktičnega rudarjenja podatkov
- Zahtevni primeri uporabe, koristni triki v programskem orodju R, vodič po uporabi dodatnega paketa caret