# Zmanjševanje razsežnosti podatkov: izbira in tvorba napovednih spremenljiv

Ljupčo Todorovski

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

Maj 2020

# Pregled predavanja

## Izbira napovednih spremenljivk

- Relevantnost in rangiranje
- Mere relevantnosti: enostavne in na modelih osnovane
- Od relevantnosti, preko rangiranja do izbire

## Tvorba novih spremenljivk

- Tvorba novih spremenljivk in zmanjševanje razsežnosti
- Metoda glavnih komponent (*Principal Component Analysis, PCA*)

# Naloga rangiranja spremenljivk

## Vhodi

- Množica napovednih spremenljivk $X = \{X_1, X_2, \ldots X_p\}$ z zalogami vrednosti $D_i : i = 1 \ldots p$
- Ciljna spremenljivka $Y$ z zalogo vrednosti $D_Y$
- Podatkovna množica $S \subseteq D_1 \times D_2 \times \ldots \times D_p \times D_Y$

## Izhod

Rangiranje $\pi$ napovednih spremenljivk, kjer velja $\pi(X_i) < \pi(X_j)$ natanko takrat ko je $X_i$ bolj relevantna od $X_j$ za napovedovanje ciljne spremenljivke $Y$ v podatkovni množici $S$.

# Relevantnost napovedne spremenljivke

**Funkcija relevantnosti $r_S : X \to \mathbb{R}$**

- Za podano spremenljivko $X_i \in X$
- Izračuna relevantnost spremenljivke $X_i$ za napoved $Y$
- V podatkovni množici $S$

**Centralno vprašanje: kako izračunamo vrednost $r_S$?**

Ko izračunamo relevantnost, postane problem rangiranja trivialno rešljiv.

# Taksonomija mer relevantnosti

## Mere relevantnosti brez modelov

- Nenadzorovane, izračunane iz $S_X = \{x : (x, y) \in S\}$
- Nadzorovane, izračunane iz $S$

## Relevantnost na osnovi napovednih modelov

Izračun relevantnosti na osnovi napovednega modela $m : D_1 \times D_2 \times \ldots D_p \rightarrow D_Y$.

# Primer: podatkovna množica *Arcene*

## Napovedovanje raka na osnovi masne spektrometrije

- Napovedne spremenljivke: koncentracije proteinov določene mase
- Ciljna spremenljivka: zdrava oseba ali pacient z rakom
- Problem dvojiškega razvrščanja

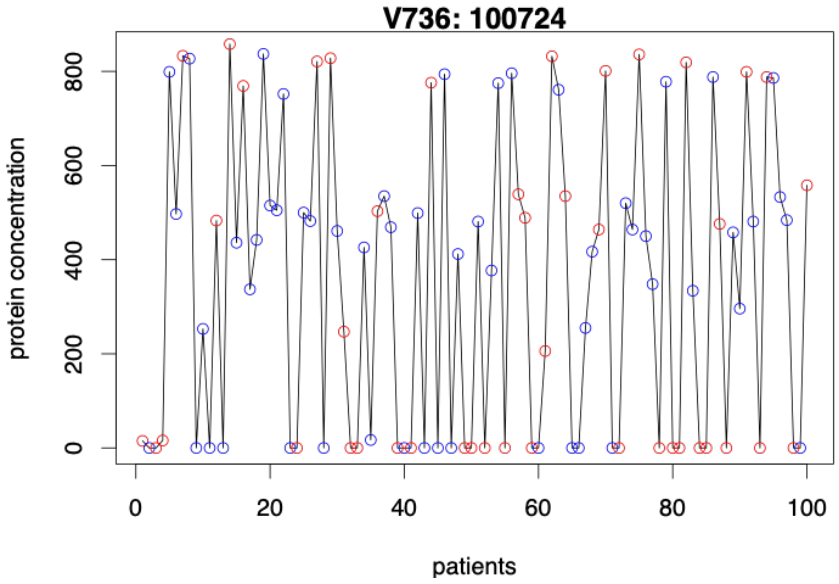
## Velikost podatkovne množice

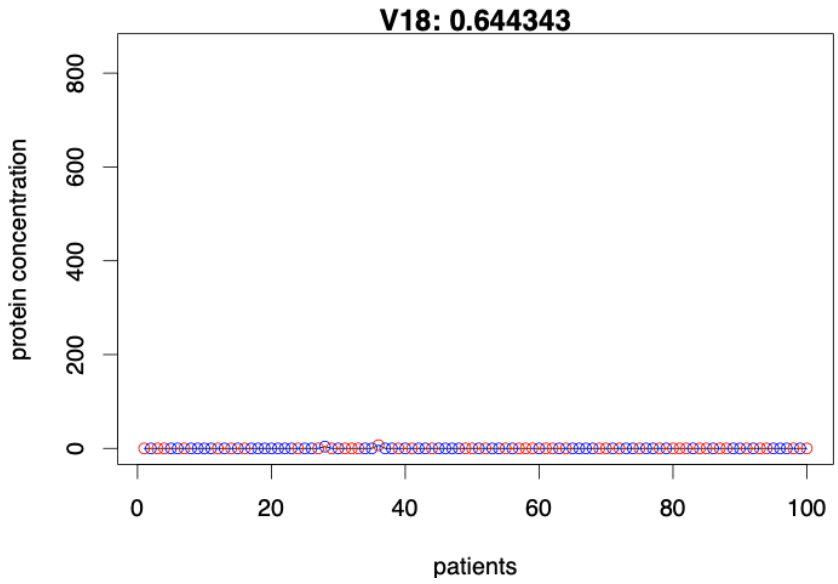
- 10.000 napovednih spremenljivk (7.000 realnih, 3.000 umetnih)
- 100 primerov, dodatnih 100 primerov v tesni množici

# Varianca spremenljivke

## Intuitivna razlaga

- Spremenljivke z majhno varianco nimajo napovedne moči
- Varianca torej pozitivno povezana z relevantnostjo
- POZOR: ista merska lestvica (ali pa vsaj normalizacija)

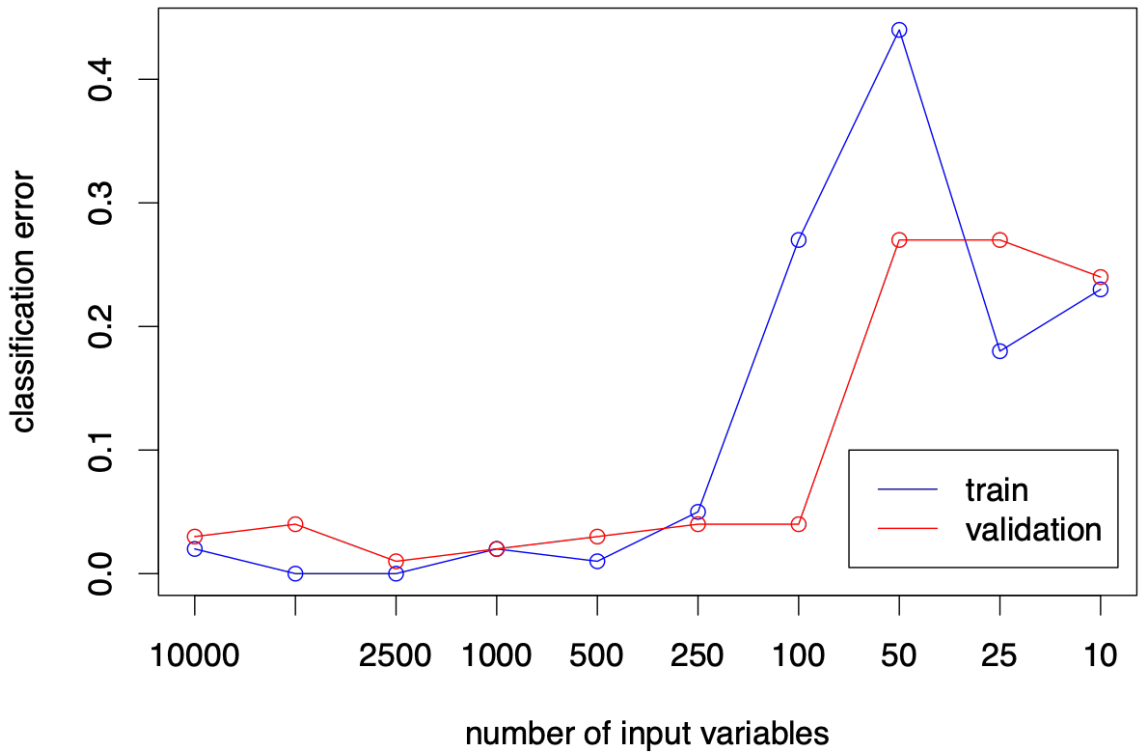


# Izbira spremenljivk na osnovi variance: *Arcene*

V učni množici izberemo $q$ najbolj relevantnih spremenljivk, $q < p$

- Iz tako skrčene učne množice $S_q$ se naučimo drevo $T_q$
- Na testni množici izmerimo napovedno točnost drevesa $T_q$



# Entropija spremenljivke

## Varianta variance za diskretne spremenljivke

$$H(X_i) = -\sum_{v \in D_i} p_v \log_2 p_v, \; p_v = \frac{|S_{i,v}|}{|S|}$$

- $S_{i,v}$ je množica učnih primerov za katere velja $X_i = v$
- Po analogiji z varianco: višja entropija ustreza večji relevantnosti

# Uni-variantne metode

## Fokus na povezavi/asociaciji med $X_i$ in $Y$

- Ker uporabljajo podatke o $Y$ jim rečemo nadzorovane
- Ker ignorirajo vrednosti $X_j : j \neq i$ so uni-variantne

## Mere in statistični testi povezanosti

Izbira odvisna od tipov spremenljivk $X_i$ in $Y$

# $X_i$ in $Y$ diskretni: $\chi^2$ in vzajemna informacija

## Vrednost $\chi^2$ statistike

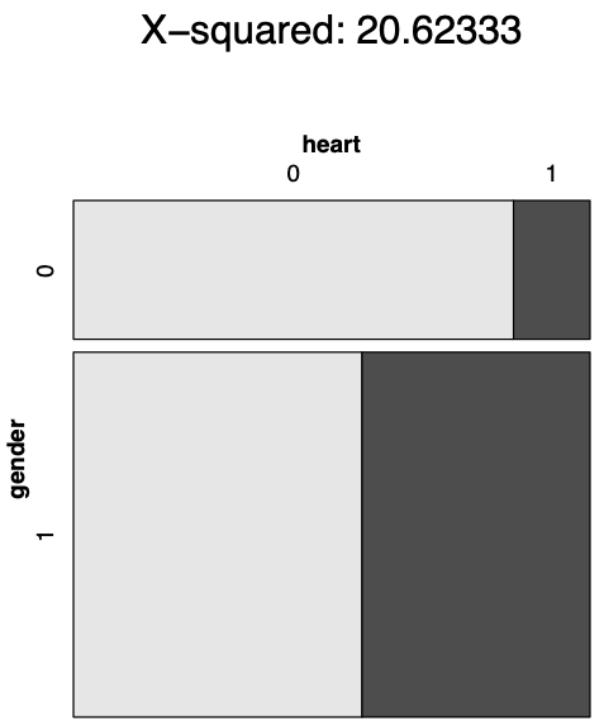
- Kontingenčna tabela $T$ dimenzij $|D_i| \times |Y|$
- $T_{ij}$ je število primerov z določeno kombinacijo vrednosti $X_i$ in $Y$
- Statistični test za merjenje povezanosti med $D_i$ in $Y$

## Vzajemna informacija $I(X_i; Y)$

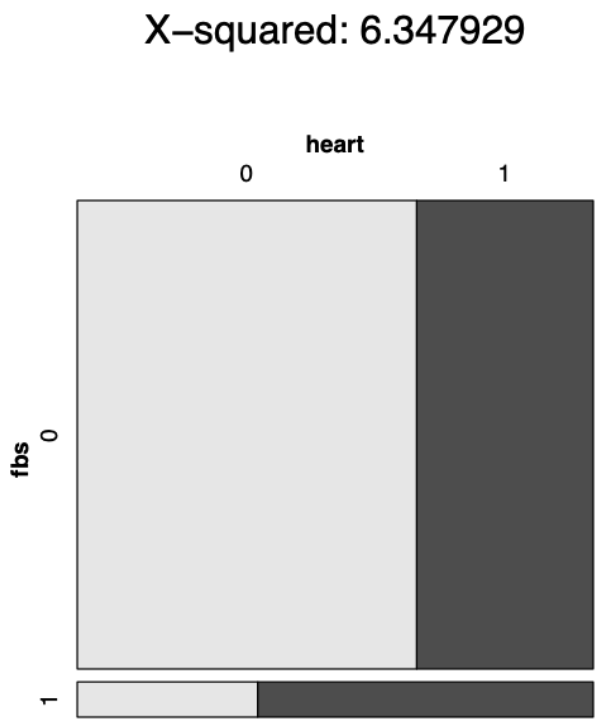
$$I(X_i; Y) = H(X_i) + H(Y) - H(X_i, Y)$$

- $H(X_i, Y)$ je vzajemna entropija $X_i$ in $Y$
- $H(X_i)$ in $H(Y)$ pa entropiji $X_i$ in $Y$

# Primera na množici *heart*




$I = 0.05912517$




$I = 0.01580802$

# $X_i$ in $Y$ različnih tipov: Welchov $t$–test ali ANOVA

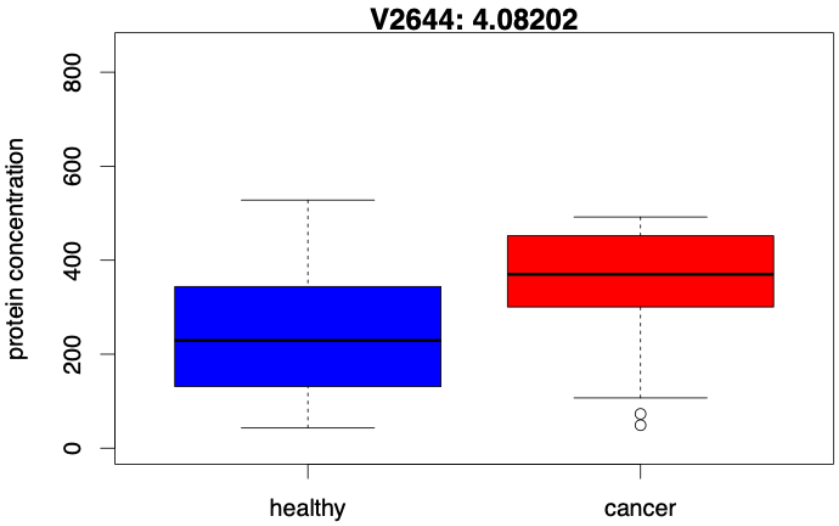
## Welchov $t$-test

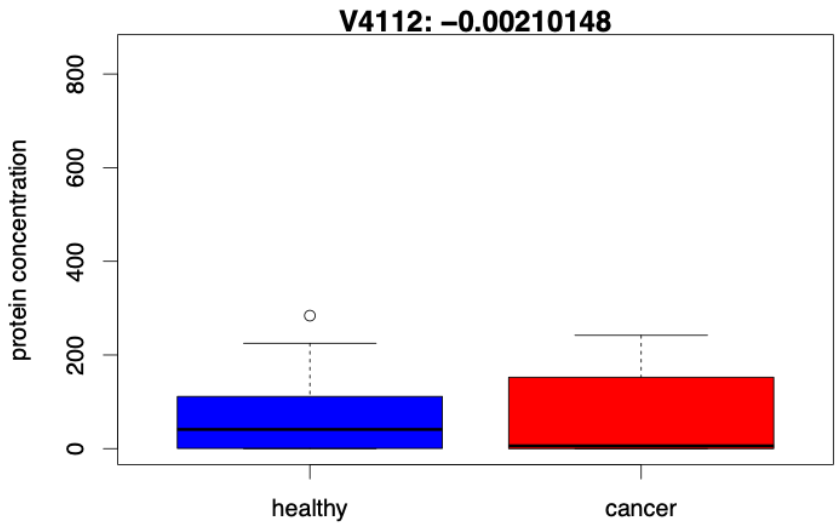
- Primerjamo povprečji numerične spremenljivke v dveh skupinah
- Vrednosti numerične spremenljivke razdelimo v dve skupini glede na vrednost diskretne spremenljivke
- Uporaben le v primeru, ko ima diskretna spremenljivka dve vrednosti

## Test ANOVA

V primerih, ko ima diskretna spremenljivka več kot dve možni vrednosti.

# Primera na množici *Arcene*

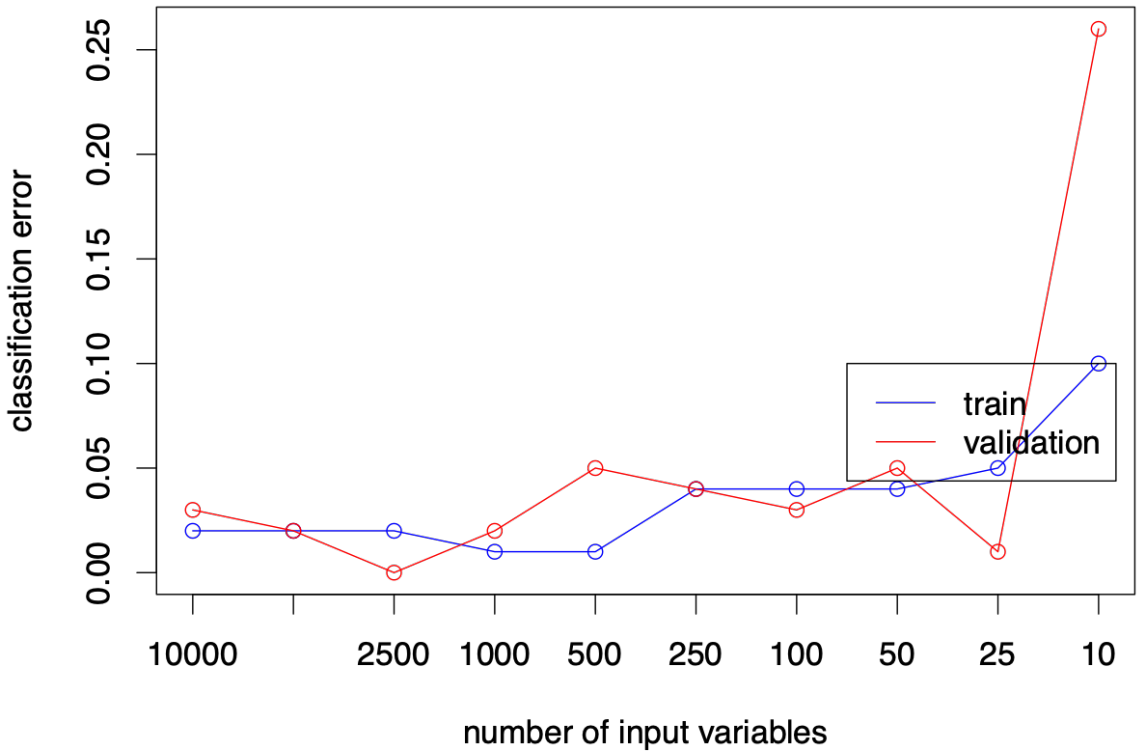


# Izbira spremenljivk na osnovi $t$-testa: *Arcene*

V učni množici izberemo $q$ najbolj relevantnih spremenljivk, $q < p$

- Iz tako skrčene učne množice $S_q$ se naučimo drevo $T_q$
- Na testni množici izmerimo napovedno točnost drevesa $T_q$



# $X_i$ in $Y$ numerični: korelacijski koeficienti

Mere linearne povezanosti med $X_i$ in $Y$

## Pearson

$$\rho^2(U, V) = \frac{(E[UV] - E[U]E[V])^2}{Var[U]Var[V]}$$

$U$ in $V$ ustrezata $X_i$ in $Y$.

## Sperman

Enako kot zgoraj, $U$ in $V$ sta rangiranji $\pi(D_i)$ and $\pi(Y)$.

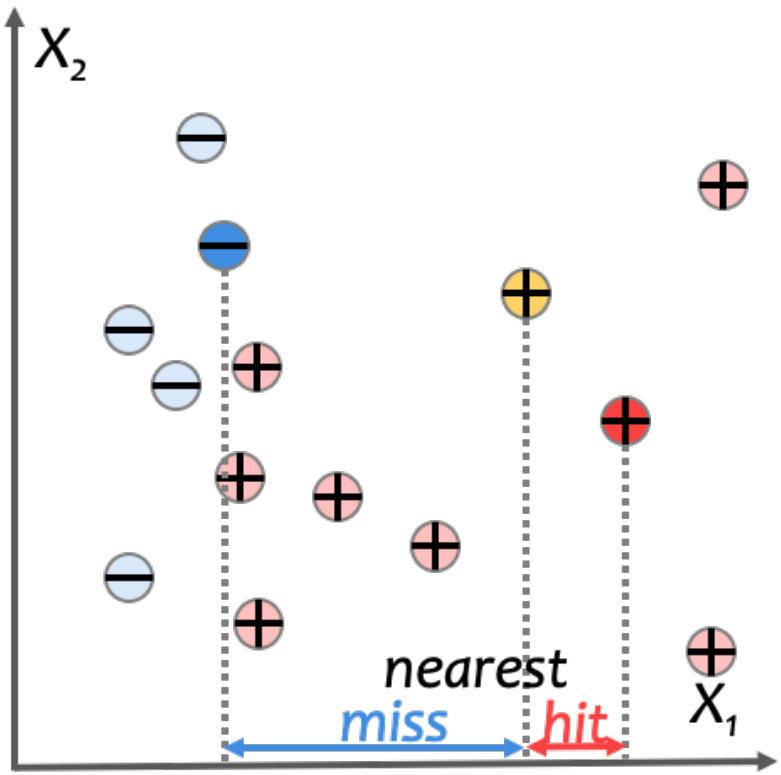
## Kendall

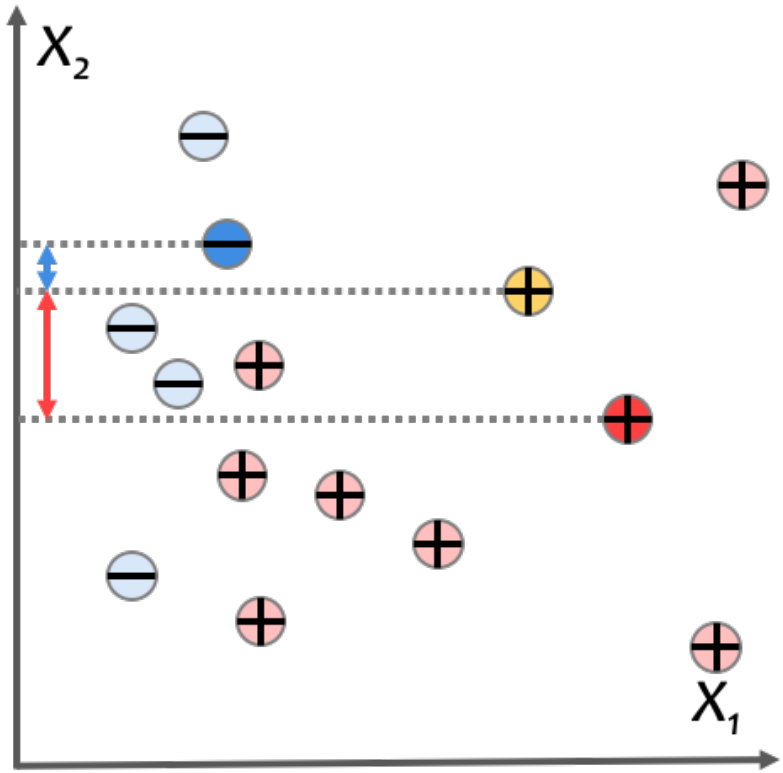
$$\rho(U, V) = \frac{2}{n(n-1)} \sum_{i<j} \text{sgn}(u_i - u_j)\,\text{sgn}(v_i - v_j)$$

Opazujemo $\rho^2$, ker nas zanima moč (in ne smer) povezave.

# *Relief*: ilustracija ideje

Na osnovi modela najbližjih sosedov



# *Relief*: razlaga ideje

Za naključno izbran primer $e$ iz $S$ opazujemo najbližji

- Zadetek $e_h$: najbližji sosed $e$, ki pripada istemu razredu
- Zgrešek $e_m$: najbližji sosed $e$, ki ne pripada istemu razredu

Dva primera

1. $|x_h - x| \ll |x_m - x|$ je dober znak za relevanco $X_i$
2. $|x_h - x| \gg |x_m - x|$ je slab znak za relevanco $X_i$

$x_h$, $x_m$ in $x$ so vrednosti $X_i$ za primere $e$, $e_h$ in $e_v$.

# *Relief*: algoritem

1. Nastavi $r_S(X_i) = 0$, za vse $i = 1, 2, \ldots p$
2. Ponovi $m$ krat (vrednost $m$ določi uporabnik)
   - Izberi naključen primer $e \in S$
   - Poišči zadetek $e_H \in S$: najbližji sosed $e$, $y = y_H$
   - Poišči zgrešek $e_H \in S$: najbližji sosed $e$, $y \neq y_M$
3. Posodobi za vse $i$

$$r_S(X_i) = r_S(X_i) - \frac{\delta(x_h, x) - \delta(x_m, x)}{m}$$

# *Relef*: algoritem, nadaljevanje

## Oznake

- $x_h$, $x_m$ in $x$ so vrednosti $X_i$ za primere $e$, $e_h$ in $e_v$.
- $y_h$, $y_m$ in $y$ so vrednosti $Y$ za primere $e$, $e_h$ in $e_v$.

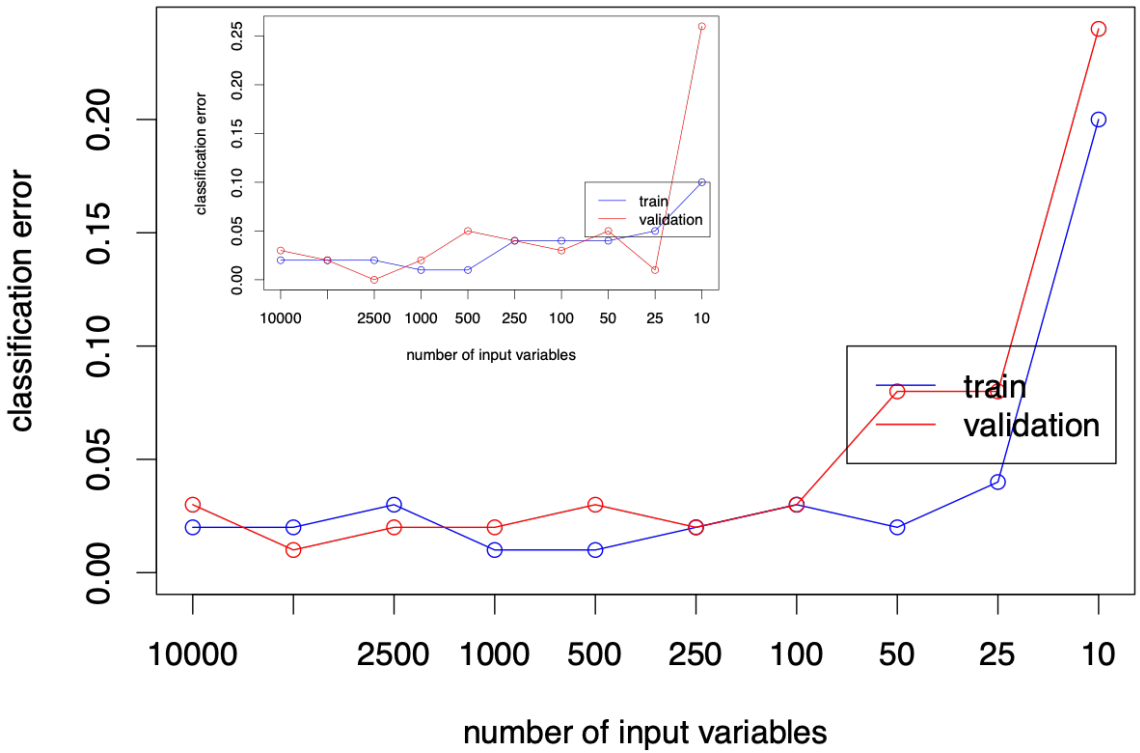
## Funkcija razdalje $\delta$

- Za numerične $X_i$: $\delta(u, v) = |u - v|/(\max(X_i) - \min(X_i))$
- Za diskretne $X_i$: $\delta(u, v) = I(u \neq v)$

# Izbira spremenljivk na osnovi *Relief*: *Arcene*

V učni množici izberemo $q$ najbolj relevantnih spremenljivk, $q < p$

- Iz tako skrčene učne množice $S_q$ se naučimo drevo $T_q$
- Na testni množici izmerimo napovedno točnost drevesa $T_q$



# *Relief* za več kot dva razreda, $|D_Y| > 2$

## Iščemo več zgreškov

- Po enega za vsako vrednost $y_m \in D_Y : y_m \neq y$
- Utežena vsota za vse zgreške $\sum_{y_m \in D_Y : y_m \neq y} p(y_m)/(1 - p(y))\delta(x_m, x)$, ki zamenja $\delta(x_m, x)$ v zadnji vrstici algoritma
- $p(y_m)$ je pogostost vrednosti $y_m$ v podatkovni množici $S$

## Nadaljnjo posploševanje algoritma

- Iščemo $k$ najbližjih zadetkov in zgreškov (namesto po enega)
- Za njih računamo povprečen $\delta$
- Vrednost parametra $k$ je uporabniško določena
- Poveča robustnost metode na šumne podatke

# *RRelief* za regresijo, $Y$ numerična

## Verjetnostna interpretacija *Relief*

$$r_S(X_i) = P(raz\ X_i|zadetki) - P(raz\ X_i|zgreski)$$

1. Verjetnost, da ima zadetek $e_h$ različno vrednost $X_i$ od primera $e$
2. Verjetnost, da ima zgrešek $e_m$ različno vrednost $X_i$ od primera $e$

## Ocenjevanje teh verjetnosti iz najbližjih sosedov od $e$

- $P(raz\ X_i|zadetki) = (1 - P(raz\ Y|raz\ X_i)) \cdot P(raz\ X_i)/P(raz\ Y)$
- $P(raz\ X_i|zgreski) = P(raz\ Y|raz\ X_i) \cdot P(raz\ X_i)/P(raz\ Y)$

## Ocena $P(raz\ Y)$ za regresijo

Iz porazdelitve vrednosti $y_n - y$ za $k$ najbližjih sosedov $e_n$ od $e$

# Razumljivi modeli in relevantnost

Iz učne množice se naučimo **razumljiv** model $m$

Iz modela $m$ **preberemo** relevantnosti napovednih spremenljivk.

## Linearni modeli

- $r_S(X_i)$ je proporcionalna absolutni vrednosti ustrezne uteži $|w_i|$
- Lahko upoštevamo še intervale zaupanja, če smo jih izračunali

## Odločitvena drevesa, možni oceni

1. Število vozlišč v drevesu, kjer test sloni na $X_i$
2. Povprečno zmanjševanje nečistoče ($IR$) v vozliščih s testom $X_i$

# Naključni gozdovi in relevantnost: *IR*

## Relevantnost napovedne spremenljivke $X$ v odločitvenem drevesu $t$

$IR(t, X)$ je povprečna vrednost zmanjševanja nečistosti $IR$ v notranjih vozliščih drevesa, ki testirajo vrednost neodvisne spremenljivke $X$.

## Relevantnost spremenljivke $X$ v ansamblu $M$

$$IR(M, X) = \frac{1}{r} \sum_{i=1}^{r} IR(m_i, X)$$

- Povprečje relevantnosti $X$ v vseh sestavinah ansambla
- Normalizacija: največja vrednost pomena enaka 1 (ali 100%)

# Naključni gozdovi in relevantnost: zmanjševanje točnosti

Izračunamo razliko $D = Err(M, S_P) - Err(M, S)$ med

- Napako modela $M$ na množici $S$
- Napako modela $M$ na množici $S_P$, kjer so vrednosti $X$ naključna permutacija pravih vrednosti $X$ v $S$
- POZOR: Modela nam ni treba znova graditi, za ocenjevanje $Err(M, S_P)$ uporabimo trik *OOB*

Vrednost razlike $D$ je povečanje napake (in zmanjševanje točnosti)

- Visoka, če je $X$ za napoved relevantna spremenljivka
- Nizka, če je $X$ nepomembna za napovedovanje

# Zmanjševanje točnosti za poljuben model

**Rahlo drugačen izračun $D = Err(M_P, S) - Err(M, S)$**

- Napaka modela $M$ na množici $S$
- Napaka modela $M_P$ naučenega na množici $S_P$, kjer so vrednosti $X$ naključna permutacija pravih vrednosti $X$ v $S$
- POZOR: Moramo zgraditi dva modela $M$ in $M_P$

Tak trik "permutacije" lahko uporabimo v kombinaciji s poljubnim modelom oz. algoritmom za strojno učenje.

# Metodi filtriranja

## Filtriranje s pragom (spodnjo mejo) relevantnosti $r_T$

- Izberi le spremenljivke $X_i : r_S(X_i) \geq r_T$
- Priporočena vrednost za *Relief*: $r_T = 1/\sqrt{\alpha p}$, kjer je $\alpha$ verjetnost, da med relevantne izberemo irelevantno spremenljivko $X_i$

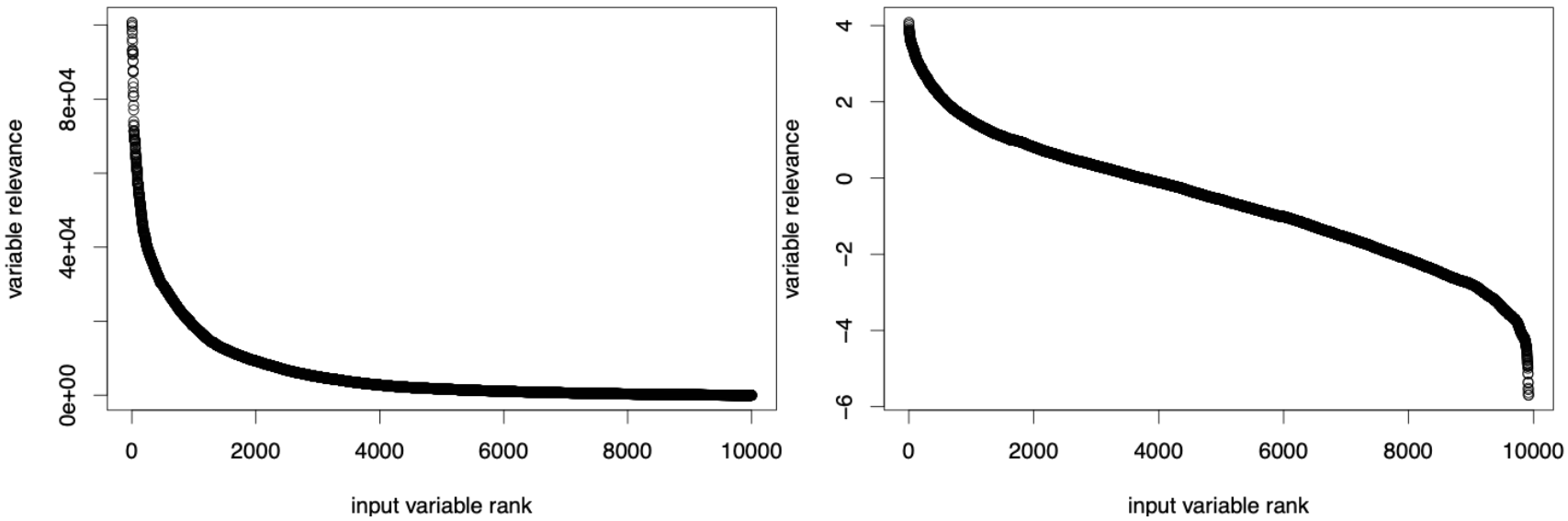
## Filtriranje $q$ najbolj relevantnih spremenljivk

## Pri obeh si običajno pomagamo z rangirnim grafom

- x-os: spremenljivke urejen po padajoči vrednosti $r_S(X_i)$
- y-os: relevantnost spremenljivk

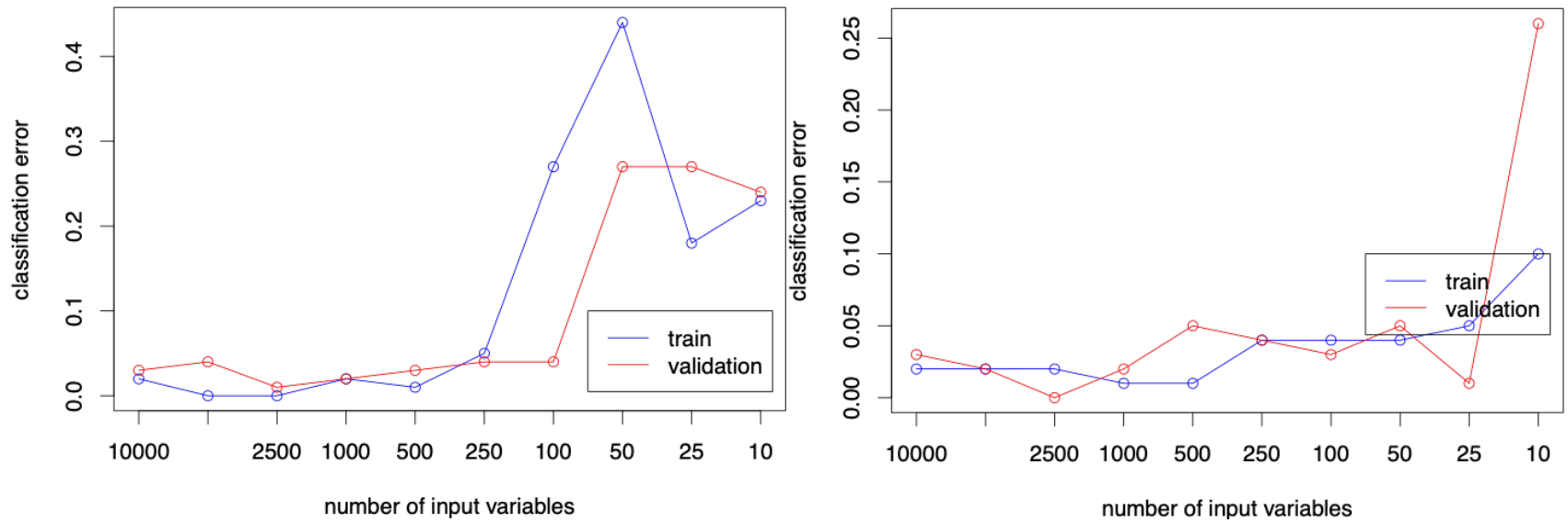
# Rangirni grafi: *Arcene*

Hitro izračunljivi

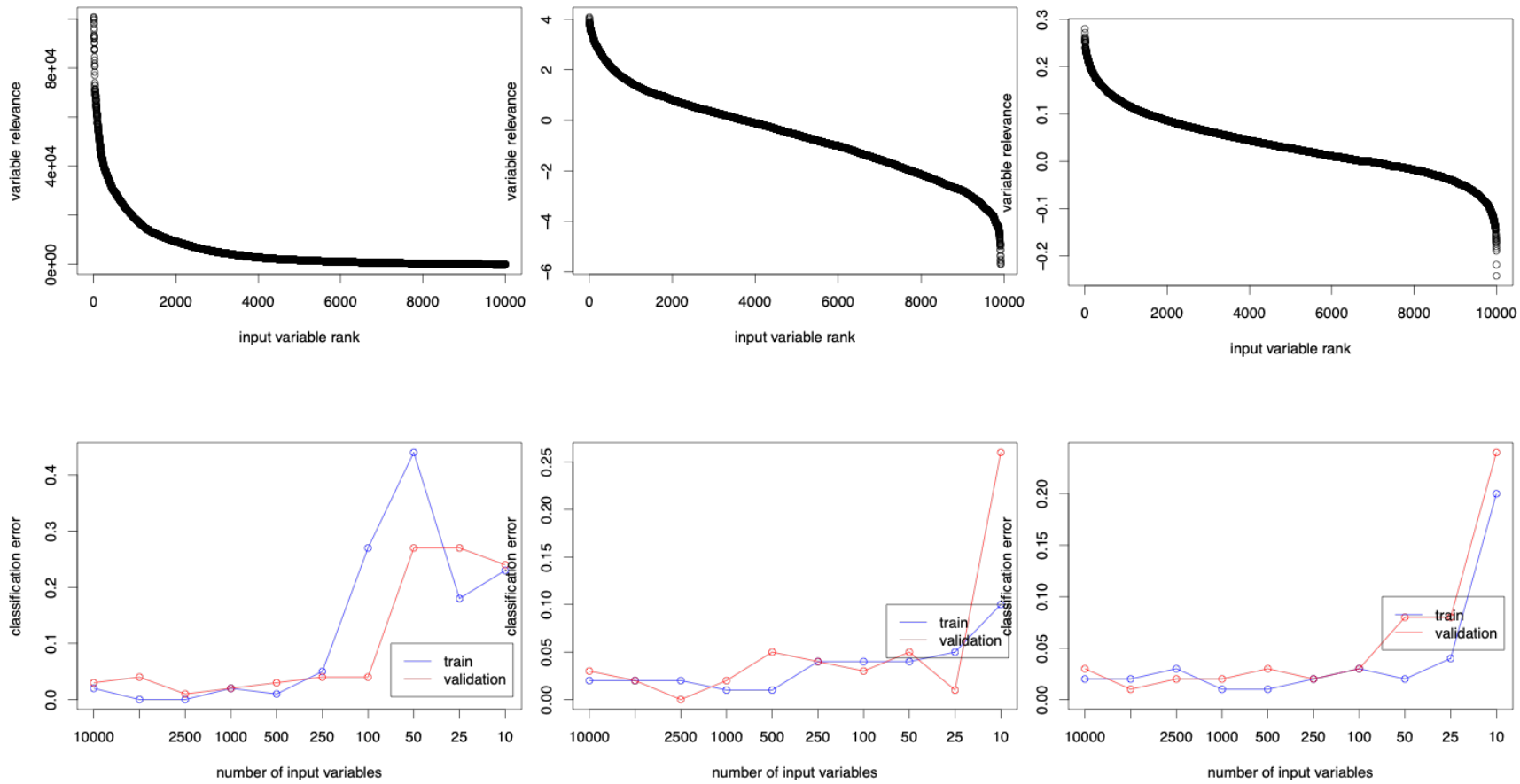


# Opazovanje napovedne napake: *Arcene*

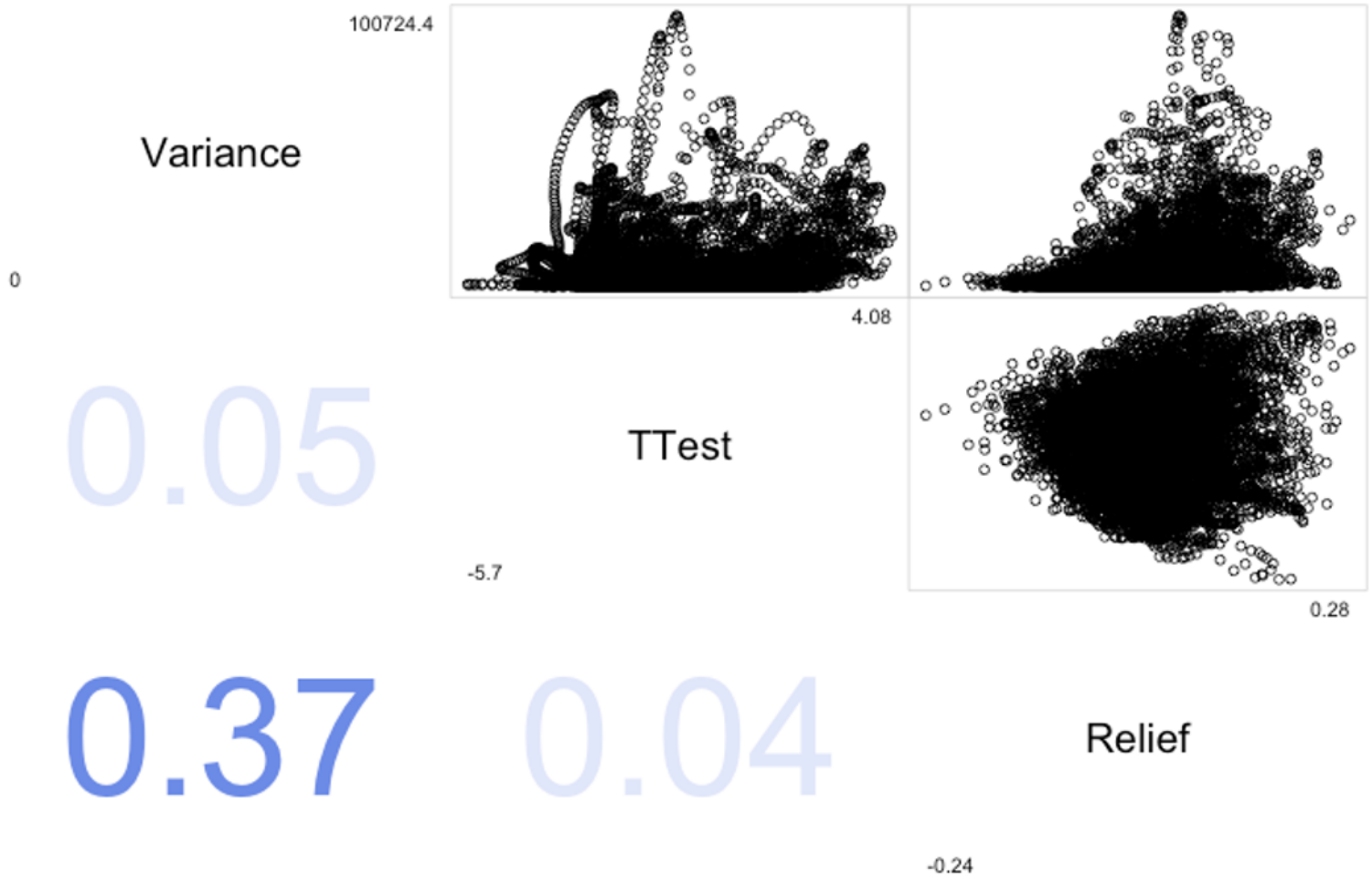
Računsko zahtevni, a bolj informativni



# Primerjava mer relevantnosti: *Arcene*



# Korelacije med merami relevantnosti: *Arcene*



# Znani algoritmi in implementacije

## Relief (Kira in Rendell 1992)

Osnovni Relief za dvojiško razvrščanje.

## RRelief (Robnik-Šikonja in Kononenko 2003)

Različne nadgradnje osnovnega algoritma, tudi za regresijo.

## CORELearn

Implementacija Relief in RRelief v R.

# Naloga zmanjševanja razsežnosti

## Vhodi

- Množica spremenljivk $X = \{X_1, X_2, \ldots X_p\}$ z zalogami vrednosti $D_i$
- Podatkovna množica $S \subseteq D_1 \times D_2 \times \ldots \times D_p$
- Pozor: ni ciljne spremenljivke $Y$, nenadzorovani scenarij

## Izhod

- Množica spremenljivk $Z = \{Z_1, Z_2, \ldots Z_q\}$ z zalogami vrednosti $D_i^Z$
- Transformirani prostor nižje razsežnosti, $q < p$, pogosto $q \ll p$
- Običajno linearna transformacija oblike $Z = XW$

# Zmanjševanje razsežnosti in izbira spremenljivk

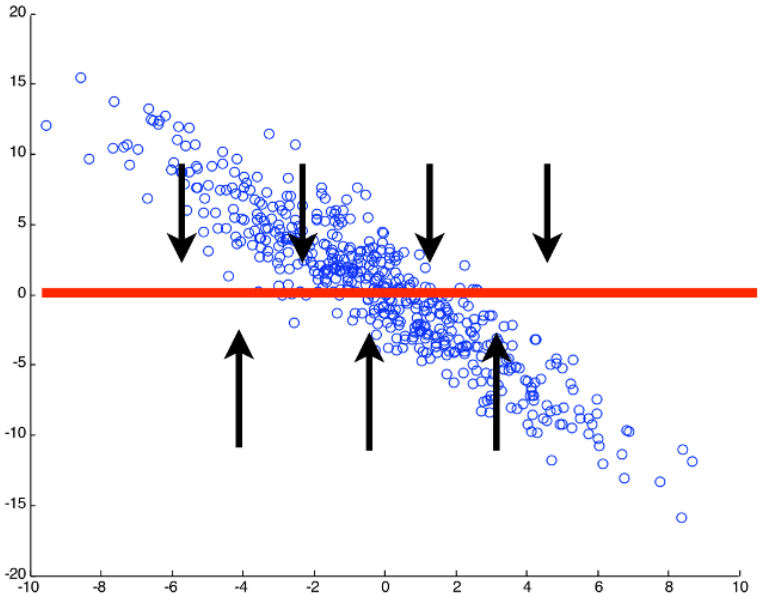
Izbira spremenljivk je tudi linearna transformacija oblike $Z = XW$

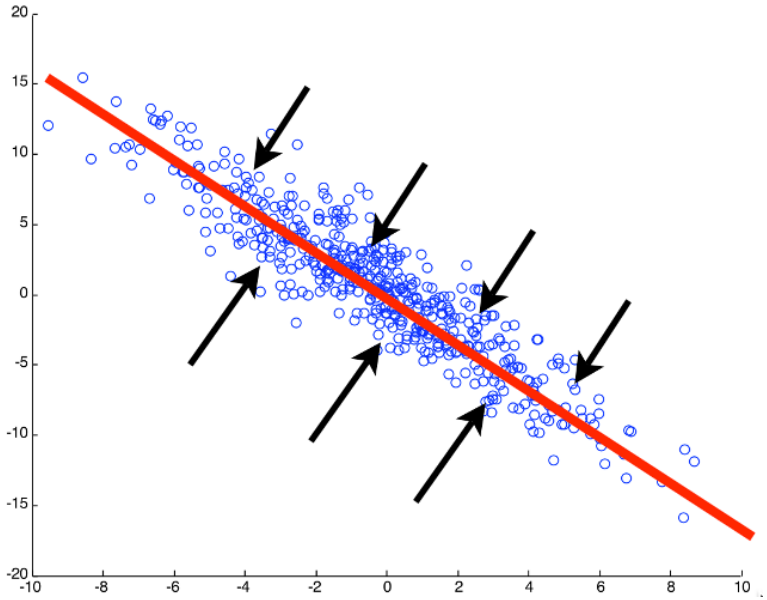
$$W = \begin{bmatrix} I_{q \times q} \\ 0_{(p-q) \times q} \end{bmatrix}$$

- Matrika $W$ je matrika dimenzij $p \times q$
- Predpostavka: spremenljivke $X$ urejene po padajoči relevantnosti

# Izbira spremenljivk in tvorba novih spremenljivk

- Levo: izbira spremenljivk
- Desno: tvorba novih spremenljivk



# Naloga

## Iščemo $W$ za transformacijo $Z = XW$ z omejitvama

1. Matrika $W$ je ortogonalna, velja torej $W^T W = WW^T = I$
2. Dimenzije $Z$ pojasnijo čim več variance osnovnih podatkov

## Predpostavimo

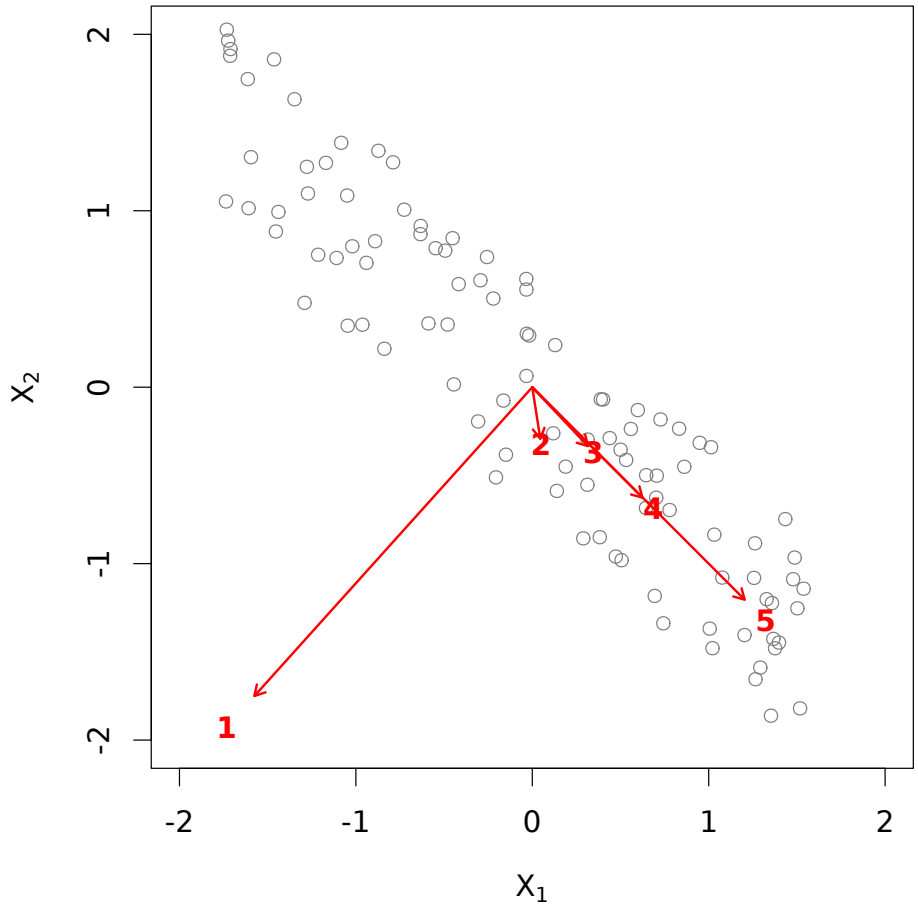
Vse spremenljivke $X$ imajo povprečje 0, kovariančna matrika je torej $X^T X$

# Osnovna ideja: od variance do lastnih vektorjev

- Kovariančna matrika $C = X^T X$
- Opazujmo iteracijo $w \leftarrow Cw$
- Za poljuben začetni vektor (1)

- Iteracija zapelje vektor v smer največje variance

- Torej za smer $w$ največje variance velja $Cw = \lambda w$



# Lastni vektorji in lastne vrednosti

Spomnimo se (LA): za lastni vektor $w$ matrike $C$ velja

$$Cw = \lambda w$$

$w$ je lastni vektor matrike $C$ pri lastni vrednosti $\lambda$

Kovariančna matrika $C$ je simetrična, torej $C^T = C$

Izberimo dva lastna vektorja $u$ in $v$ pri lastnih vrednostih $\lambda$ in $\mu \neq \lambda$

$$\begin{aligned}(\lambda - \mu)\langle u, v\rangle &= \lambda u^T v - \mu u^T v \\ &= (\lambda u)^T v - u^T(\mu v) = (Cu)^T v - u^T(Cv) \\ &= u^T C^T v - u^T C v = 0\end{aligned}$$

Njihov skalarni produkt $\langle u, v\rangle = 0$, lastna vektorja sta torej ortogonalna.

# Algoritem za izračun $W$

**Za podan $S$ izračunamo kovariančno matriko $C$**

Če imajo vse spremenljivke iz $X$ povprečje 0, je to $S^T S$

**Za $C$ izračunamo lastne vektorje $w_i$ in lastne vrednosti $\lambda_i$**

$$W = [w_1 w_2 \cdots w_p], \ \Lambda = diag(\lambda_1, \lambda_2, \cdots \lambda_p)$$

- Ker je $C$ simetrična, velja $W^T W = I$
- Velja tudi $C = W \Lambda W^T$

# Matrika $W$ in glavne komponente

**Glavne komponente so stolpci matrike $W$**

Torej lastni vektorji matrike $C$.

**$\lambda_i$ pove koliko variance pojasni komponenta $w_i$**

Lastne vrednosti matrike $C$ pretvorimo v deleže celotne variance.

# Matrika $W$ in zmanjševanje razsežnosti

$$Z = XW_q$$

**Izberemo prvih $q$ glavnih komponent (lastnih vektorjev)**

- $W_q$ je matrika dimenzij $p \times q$
- Vsebuje je prvih $q$ stolpcev matrike $W$

Kako izberemo vrednost $q$?

# Hevristike za izbiro $q$

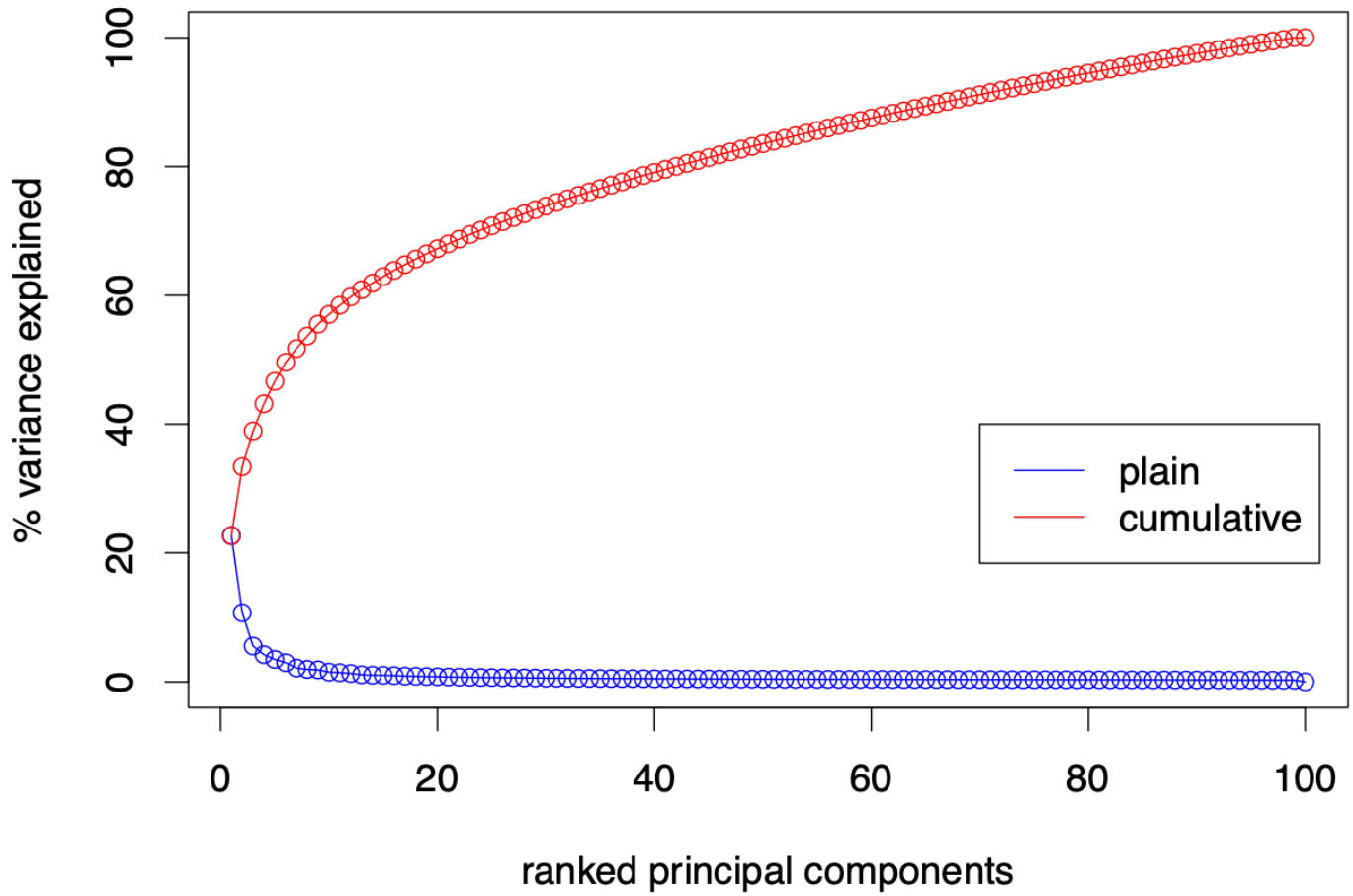
Najprej nariši graf *scree* (naslednja prosojnica)

Izberi $q$, kjer je

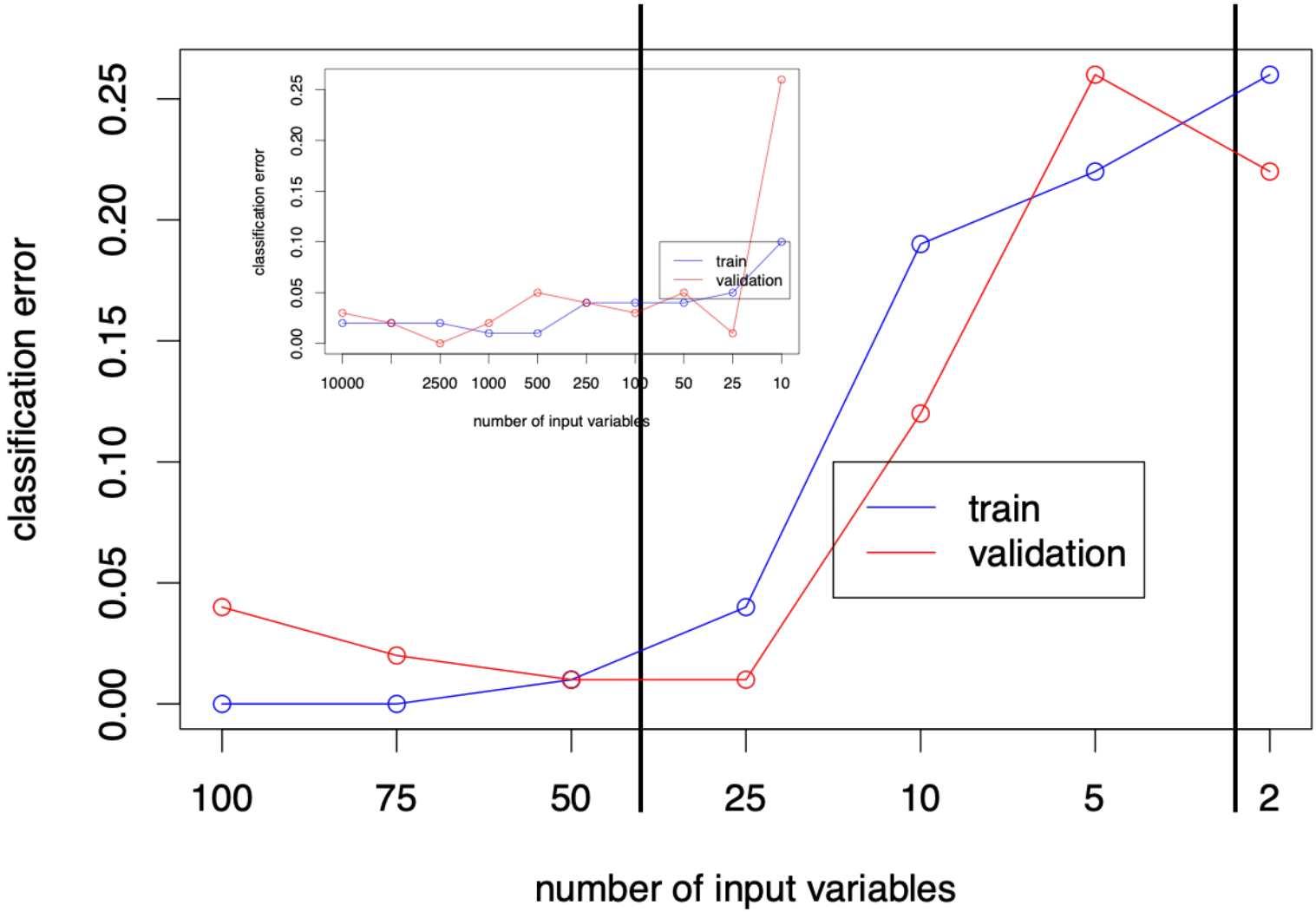
1. Kumulativna pojasnjena varianca večja od podanega praga (običajno je vrednost praga 80% ali 90%)
2. Prispevek k pojasnjeni varianci še vedno večji od podanega praga (običajno je vrednost praga 5% ali 10%)
3. Prvo vidno koleno v krivulji *scree*

# Primer grafa *scree*: *Arcene*

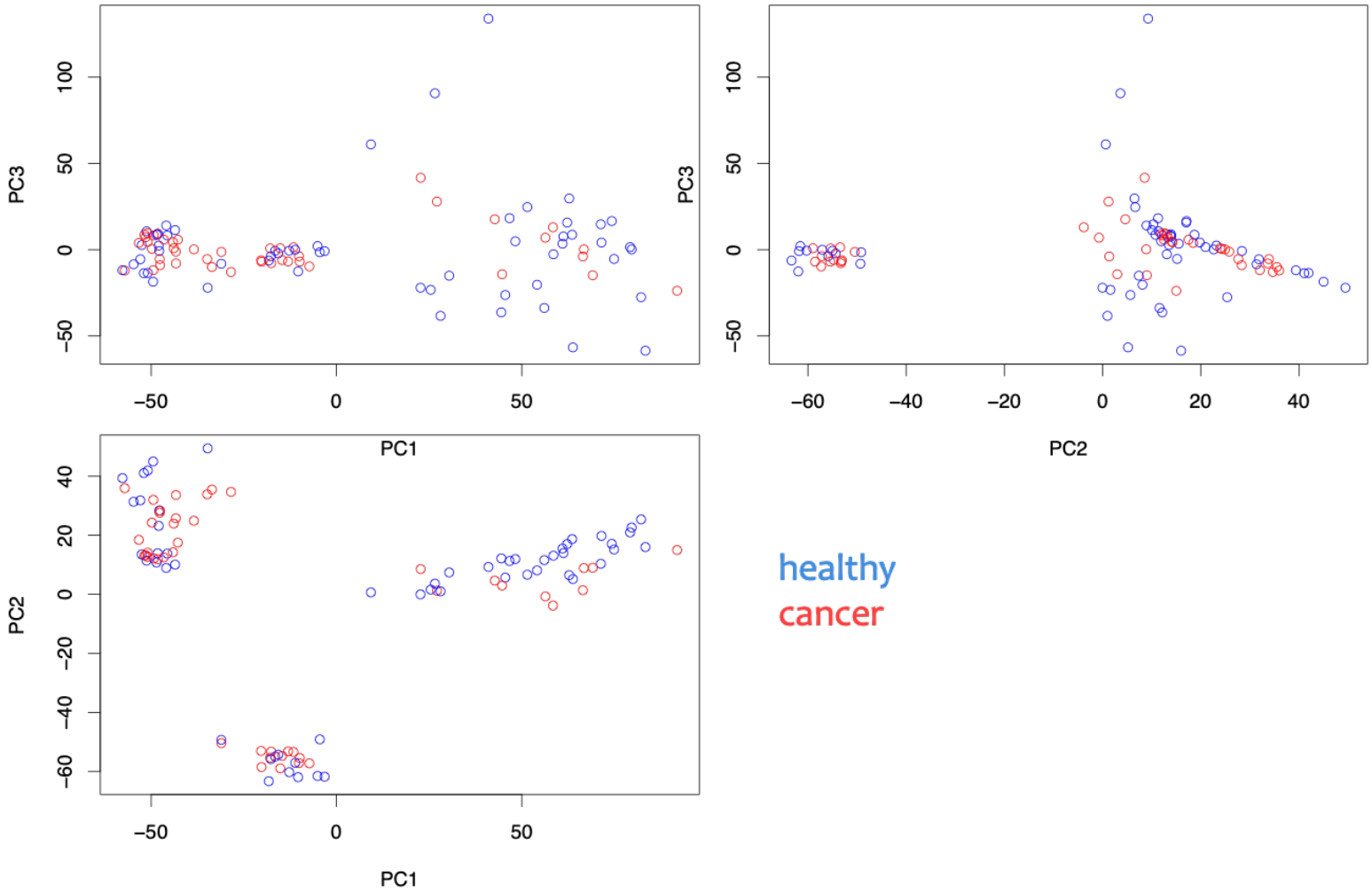
Prva hevristika $q = 42$, druga in tretja $q = 3$



# Izbira glavnih komponent: *Arcene*



# Glavne komponente in vizualizacija: *Arcene*



# Implementacija

**prcomp**

Vgrajena funkcija v R implementira metodo glavnih komponent.