OSNOVE UMETNE INTELIGENCE

2022/23

k najbližjih sosedov

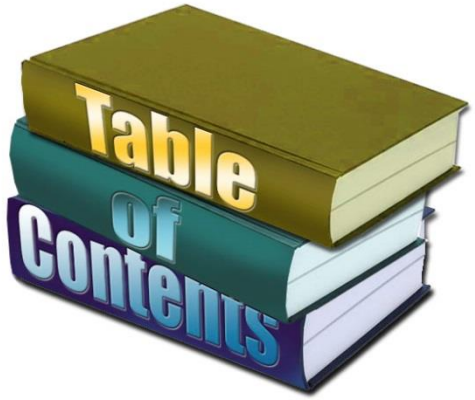
regresija

nenadzorovano učenje

Pridobljeno znanje s prejšnjih predavanj

- **strojno učenje**
  - različne vrste atributov: diskretni (nominalni, ordinalni) in zvezni
  - **diskretizacija** zveznih atributov (intervali z enako frekvenco, intervali enake širine, maksimizacija informacijskega prispevka)
  - obravnava **manjkajočih** atributov (učenje z manjkajočimi vrednostmi, nadomeščanje, verjetnostno napovedovanje)
  - **naivni Bayesov klasifikator**
    - diagnostično sklepanje, vzročno sklepanje
    - preslikava v strojno učenje (evidenca in hipoteza → atributi in razred)
    - "naivna" poenostavitev pogojnih verjetnosti
    - klasifikacija v najbolj verjeten razred
    - primeri (sadeži, naloga z izpita)
  - **nomogrami** kot orodje za vizualizacijo naivnega Bayesovega klasifikatorja:
    - razumevanje vizualizacije: predstavitev prispevkov vrednosti posameznih atributov
    - točke, izračunane z razmerjem verjetja
    - os za vsak atribut, os za vsoto vseh točk

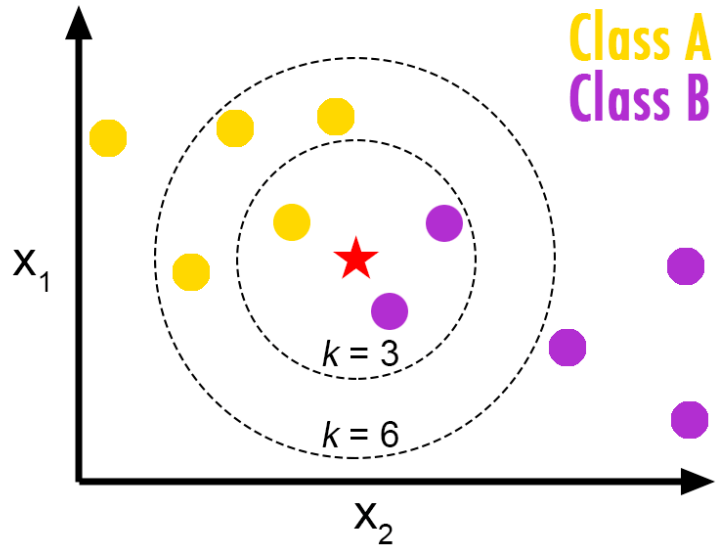
Pregled



- strojno učenje
  - uvod v strojno učenje
  - učenje odločitvenih dreves
  - učenje dreves iz šumnih podatkov (rezanje dreves)
  - ocenjevanje učenja
  - diskretizacija atributov, obravnava manjkajočih vrednosti
  - naivni Bayesov klasifikator
  - nomogrami
  - k najbližjih sosedov
  - lokalna utežena regresija
  - regresijska drevesa
  - nenadzorovano učenje

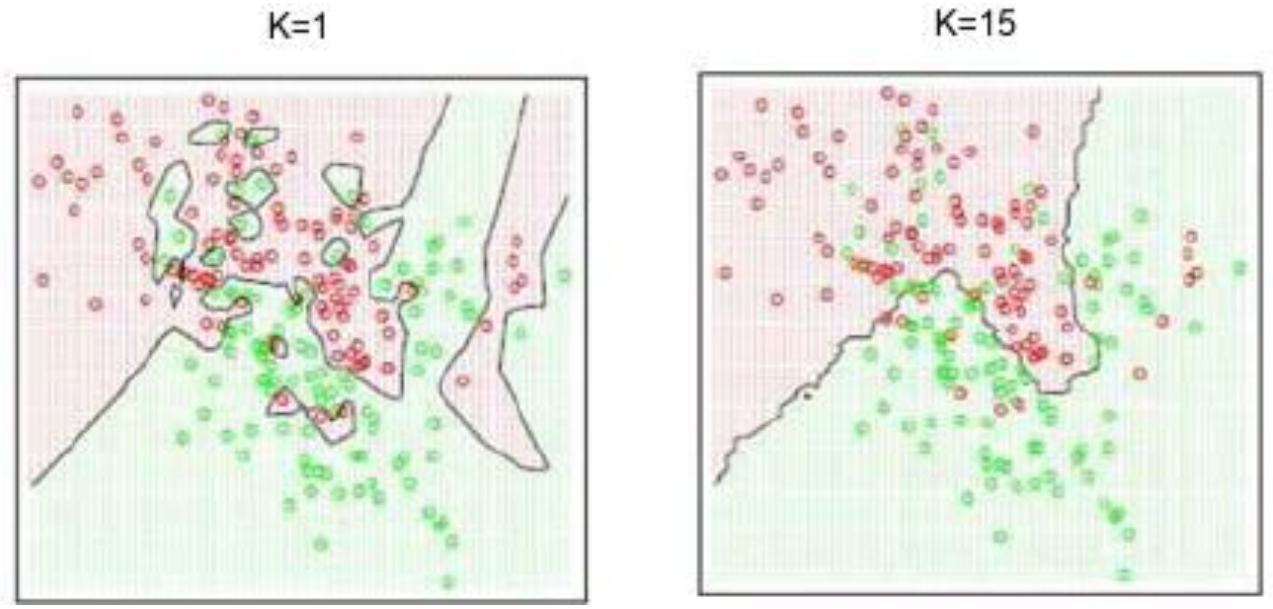
Metoda k najbližjih sosedov

- angl. *k nearest neighbors*
- lastnosti:
  - **neparametrična** metoda (ne ocenjuje parametrov izbranega modela)
  - učenje na podlagi **posameznih primerov** (angl. *instance-based learning*)
  - **leno učenje** (angl. *lazy learning*): z učenjem odlaša vse do povpraševanja o novem primeru
- ideja: ob vprašanju po vrednosti odvisne spremenljivke za novi primer:
  - poišči **k primerov**, ki so **najbližji** glede na podano **mero razdalje**
  - napovej
    - **pri klasifikaciji**: npr. večinski razred med sosedi
    - **pri regresiji**: npr. povprečno vrednost/mediano označb sosedov
- v izogib neodločenemu glasovanju za večinski razred pri klasifikaciji običajno izberemo, da je *k* liho število



Metoda k najbližjih sosedov

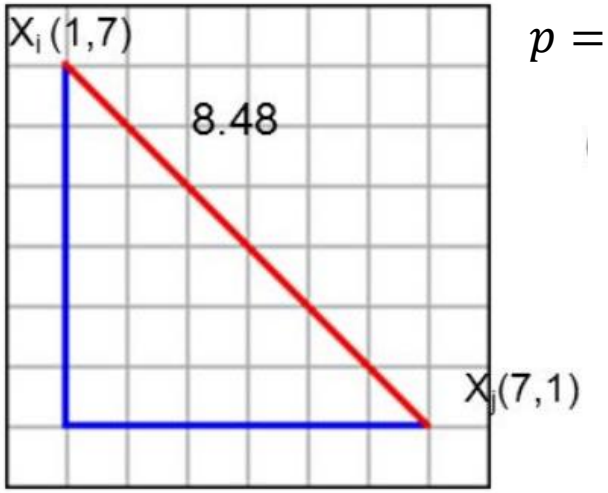
- pomembna je izbira ustreznega k:
  - premajhen k: pretirano prilagajanje
  - prevelik k: prešibko prilagajanje (pri $k = N$: napoved večinskega razreda)
  - v praksi običajno: $k = 5$



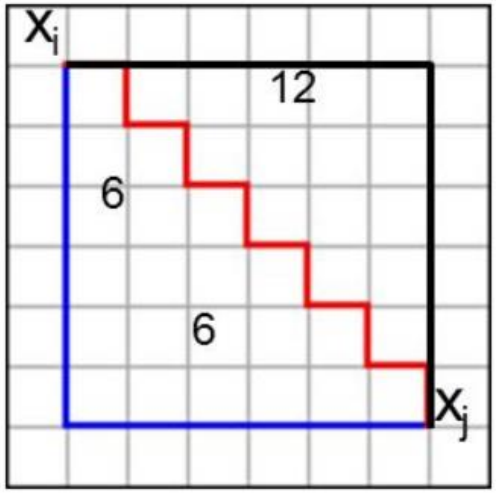
Hastie, Tibshirani, Friedman: Elements of Statistical Learning, 2009

Metoda k najbližjih sosedov

- razdaljo običajno merimo z razdaljo Minkowskega: $L^p(\boldsymbol{x}_i, \boldsymbol{x}_j) = \left(\sum_k |x_{i,k} - x_{j,k}|^p\right)^{\frac{1}{p}}$
  - za $p = 2$ je to evklidska razdalja: $L^2(\boldsymbol{x}_i, \boldsymbol{x}_j) = \sqrt{\sum_k (x_{i,k} - x_{j,k})^2}$
  - za $p = 1$ je to manhattanska razdalja: $L^1(\boldsymbol{x}_i, \boldsymbol{x}_j) = \sum_k |x_{i,k} - x_{j,k}|$
- različni pristopi:
  - za zvezne atribute: razlika med vrednostima
  - za diskretne atribute: Hammingova razdalja (število neujemajočih diskretnih atributov pri obeh primerih)



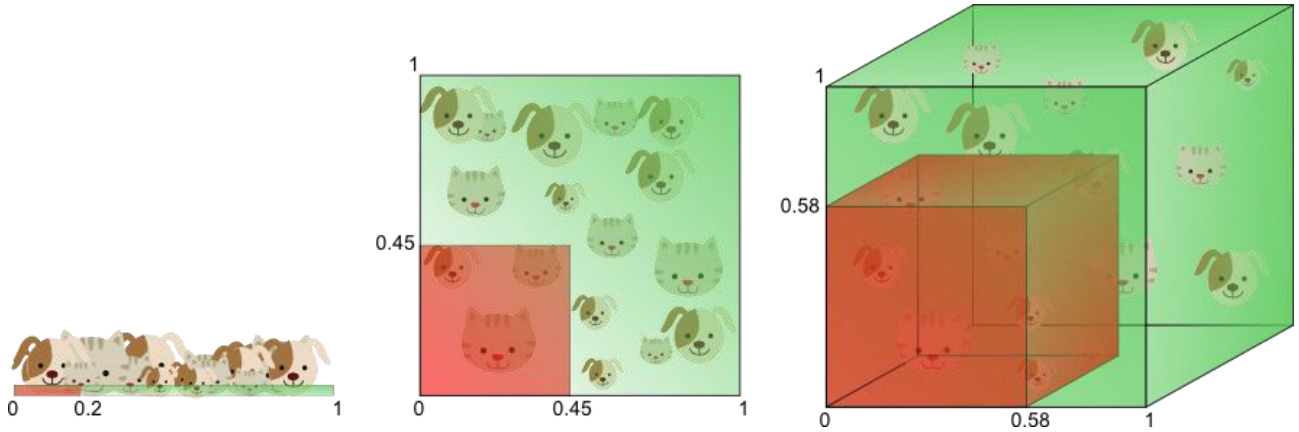
$p = 2$

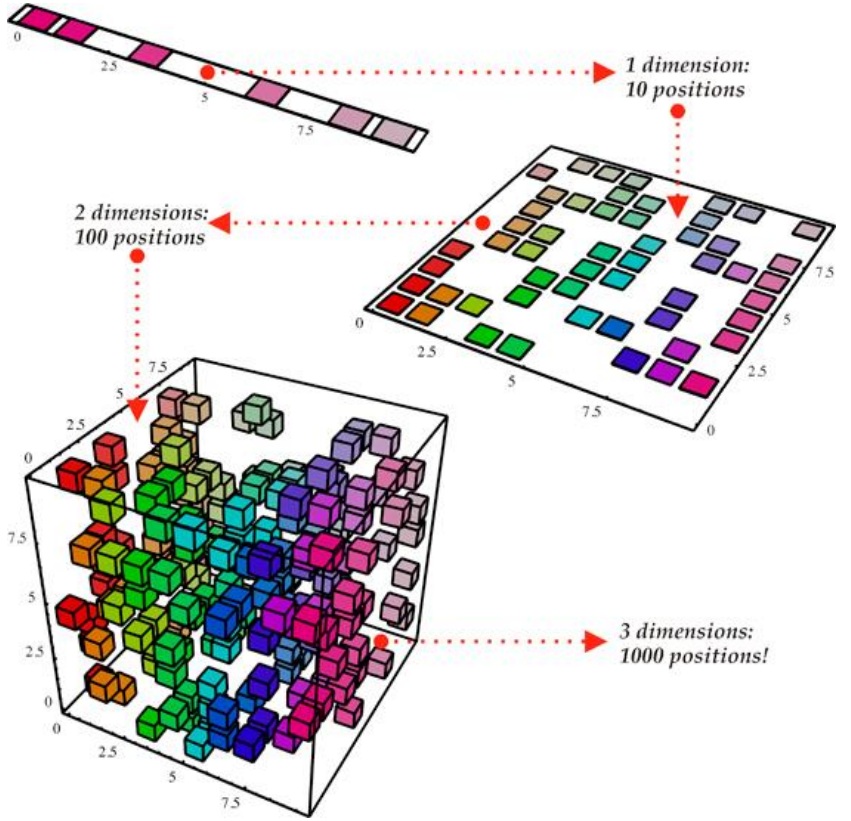


$p = 1$

Opombe

- vpliv intervala vrednosti na izračunano razdaljo vpliva na najdene najbližje sosede → **potrebna normalizacija**
- pri velikem številu dimenzij lahko postanejo primeri zelo oddaljeni – **prekletstvo dimenzionalnosti** (angl. *the curse of dimensionality*)
- implementacije iskanja najbližjih sosedov: $O(N)$, $O(logN)$, $O(1)$





← pokritje 20% problemskega prostora s povečevanjem števila dimenzij

Izpitna naloga

- 2. izpitni rok, 15. 2. 2018 (prilagojena naloga)

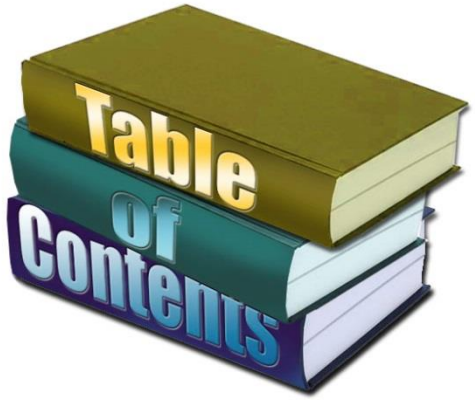
2. NALOGA (25t):

Podana je učna množica primerov, ki je prikazana v tabeli (*vreme* in *pritisk* sta atributa, *glavobol* pa je razred). Naloge:

c) V kateri razred bi klasifikator k-NN (pri k=3 in uporabi Hammingove razdalje) klasificiral učni primer z vrednostmi atributov *vreme=deževno, pritisk=srednji*?

vreme	pritisk	glavobol
sončno	nizek	ne
sončno	nizek	ne
sončno	srednji	da
sončno	visok	ne
sončno	nizek	ne
sončno	nizek	da
deževno	srednji	ne
deževno	srednji	da
deževno	visok	da

Pregled



- strojno učenje
  - uvod v strojno učenje
  - učenje odločitvenih dreves
  - učenje dreves iz šumnih podatkov (rezanje dreves)
  - ocenjevanje učenja
  - diskretizacija atributov, obravnava manjkajočih vrednosti
  - naivni Bayesov klasifikator
  - nomogrami
  - k najbližjih sosedov
  - lokalna utežena regresija
  - regresijska drevesa
  - nenadzorovano učenje

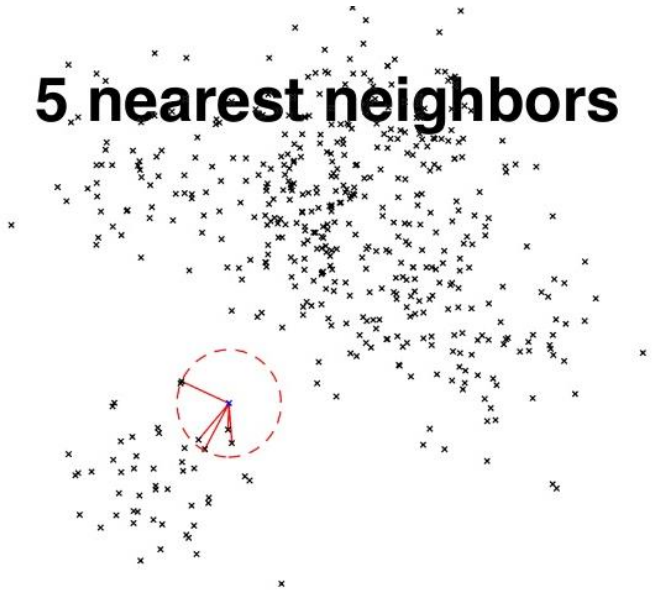
k najbližjih sosedov za regresijo

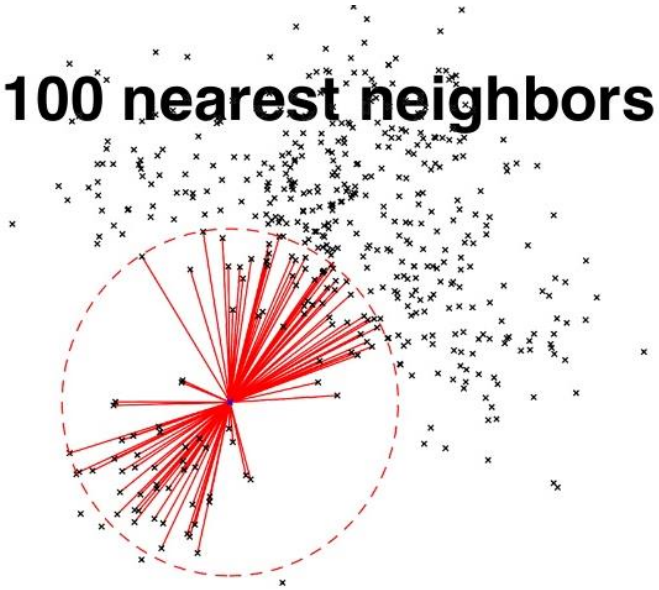
- poišči **k primerov**, ki so **najbližji** glede na podano **mero razdalje**
- možnosti za izračun napovedi:
  - povprečna vrednost/mediana označb sosedov
  - utežena vsota

- **uteževanje z razdaljo** (lokalno utežena regresija)
  - $h(x_?) = \frac{\sum_{i=1}^{k} w_i \cdot f(x_i)}{\sum_{i=1}^{k} w_i}$
  - w_i je utež, ki je lahko enaka $w_i = \frac{1}{\left(d(x_?, x_i)\right)^2}$

  - pri uteževanju se lahko uporablja tudi **poljubna jedrna funkcija, npr. Gaussovo jedro:**

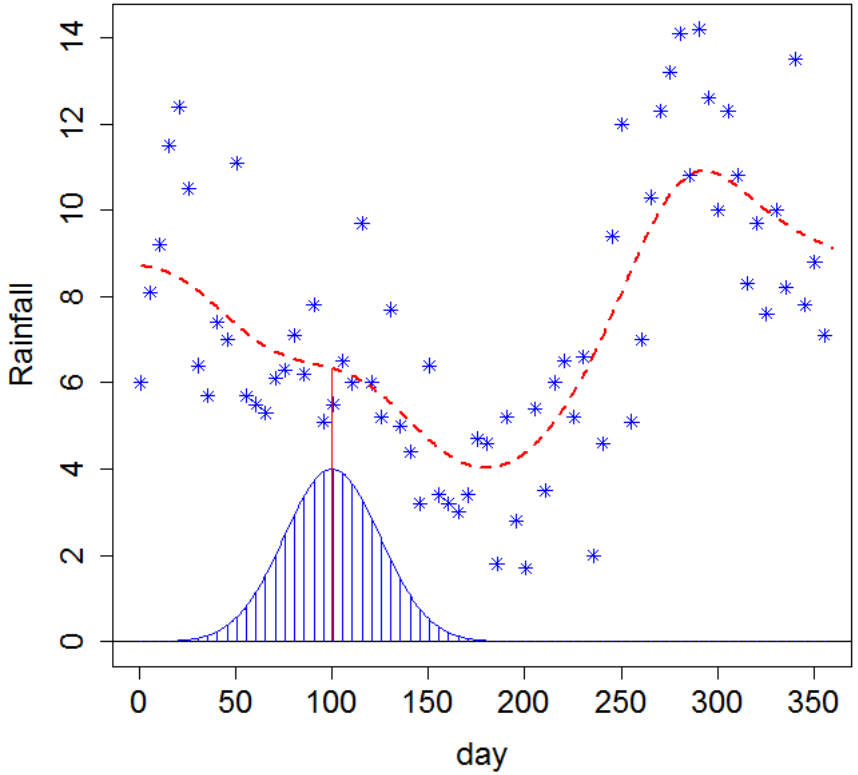
$$\frac{1}{\sqrt{2\pi}} e^{-\frac{\left(d(x_?, x_i)\right)^2}{2}}$$

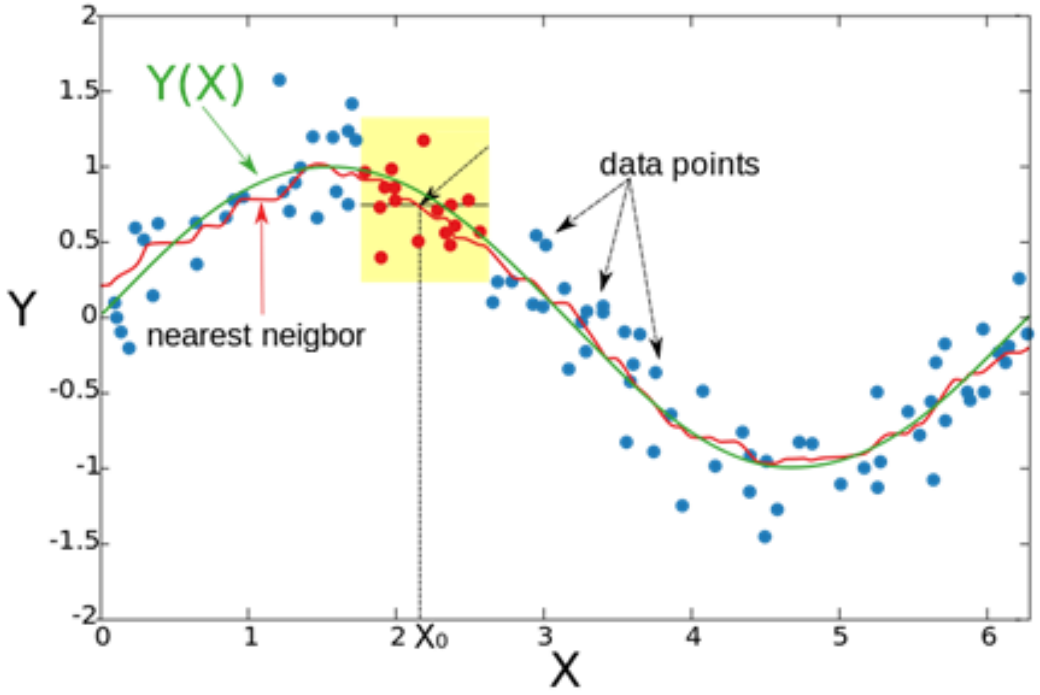




k najbližjih sosedov za regresijo

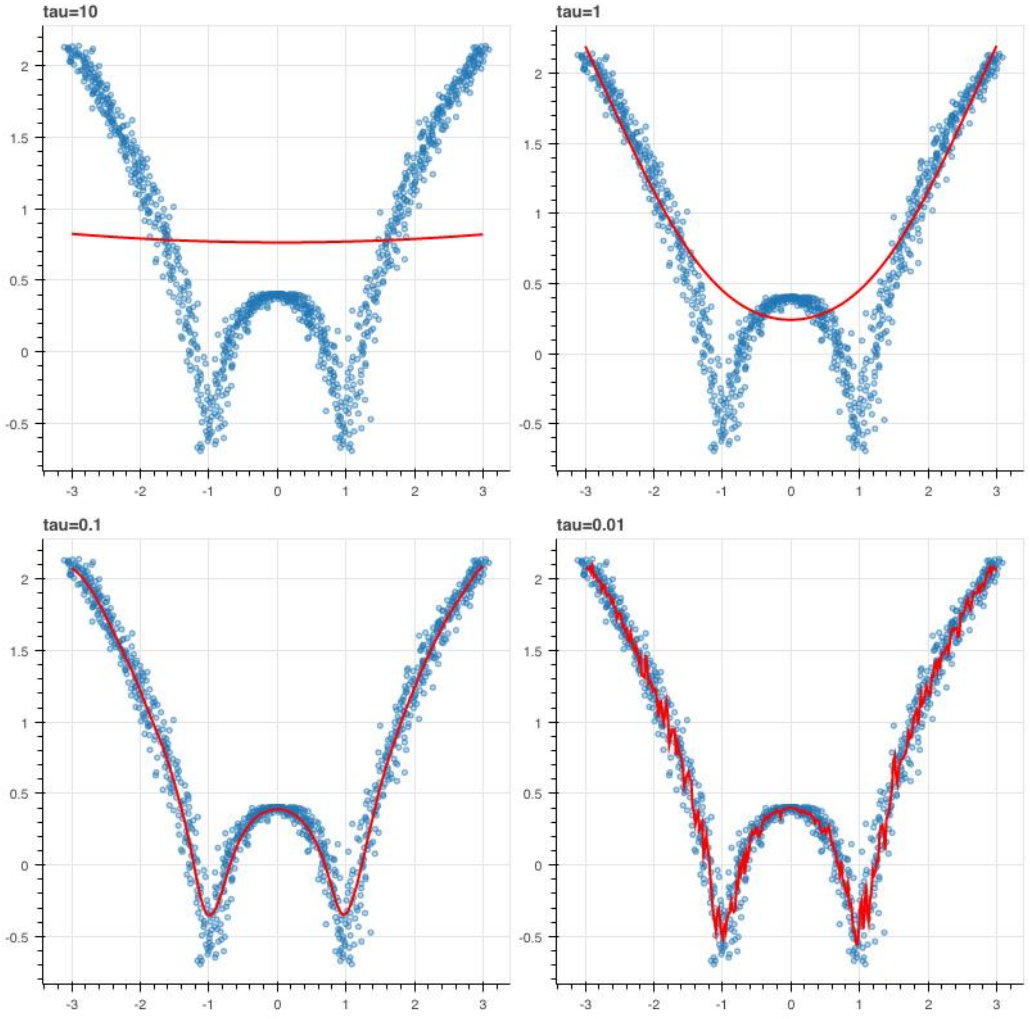
- primer v 2 dimenzijah (atribut X in odvisna spremenljivka Y)

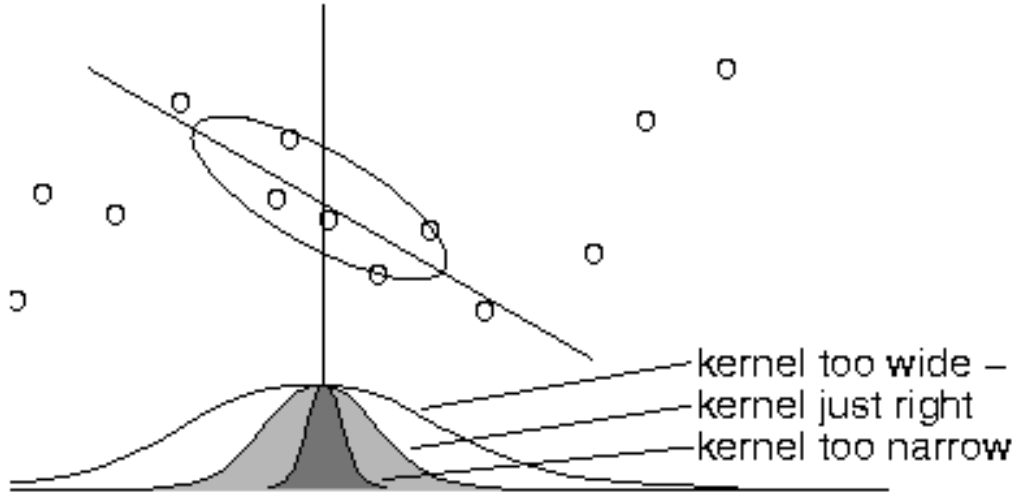


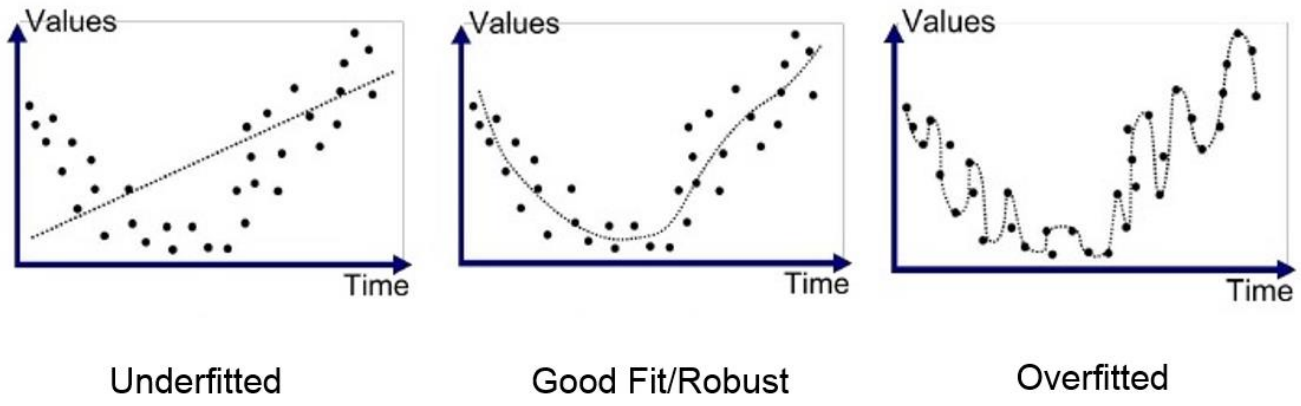


http://blogs.cornell.edu/modelmeanings/2014/10/17/local-interpretation/

Pomen širine jedra za prileganje podatkom







https://medium.com/greyatom/what-is-underfitting-and-overfitting-in-machine-learning-and-how-to-deal-with-it-6803a989c76

Izpitna naloga

- 2. izpitni rok, 12. 2. 2020 (prilagojena naloga)

1. NALOGA (10t):

Učna množica vsebuje 5 učnih primerov, katerih medsebojne razdalje (izračunane z neko mero razdalje) so podane v tabeli na desni strani.

	1	2	3	4	5
1	0	18	14	14	16
2	18	0	4	20	26
3	14	4	0	20	22
4	14	20	20	0	26
5	16	26	22	26	0

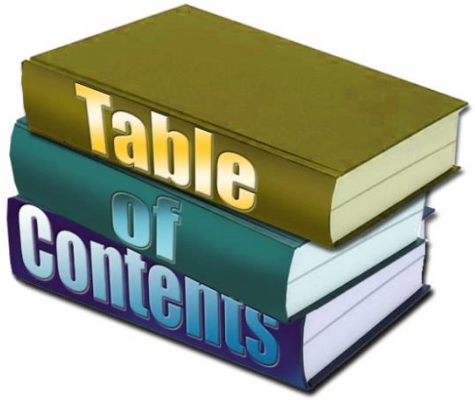
c.) Primeri 1-3 imajo vrednost odvisne spremenljivke enako 10, primera 4-5 pa vrednost odvisne spremenljivke enako 20. Kako bi naslednji napovedni modeli klasificirali primer z zaporedno številko 4 (predpostavi, da ga izločimo iz učne množice in obravnavamo kot testni ali nevideni primer):

- klasifikacijski model 3-NN:

- regresijski model 3-NN:

- lokalno uteženа regresija s funkcijo za uteževanje primerov $w = 1$:

Pregled

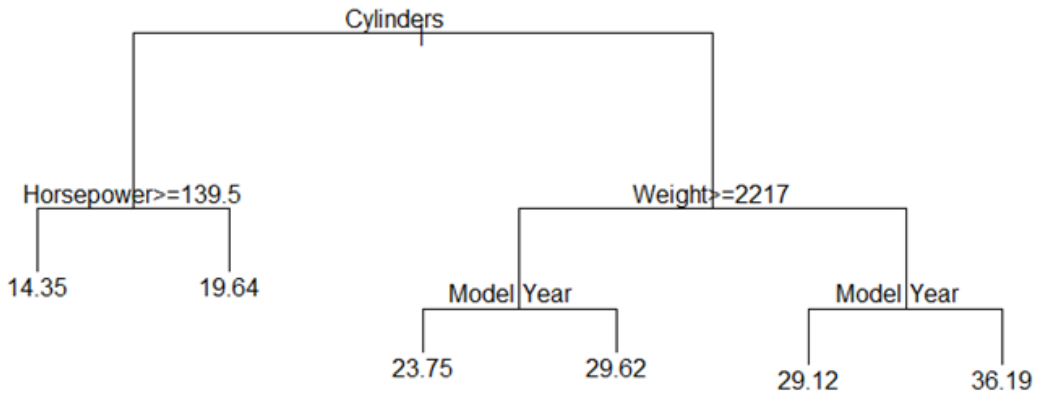


- strojno učenje
  - uvod v strojno učenje
  - učenje odločitvenih dreves
  - učenje dreves iz šumnih podatkov (rezanje dreves)
  - ocenjevanje učenja
  - diskretizacija atributov, obravnava manjkajočih vrednosti
  - naivni Bayesov klasifikator
  - nomogrami
  - k najbližjih sosedov
  - lokalna utežena regresija
  - regresijska drevesa
  - nenadzorovano učenje

Regresijska drevesa

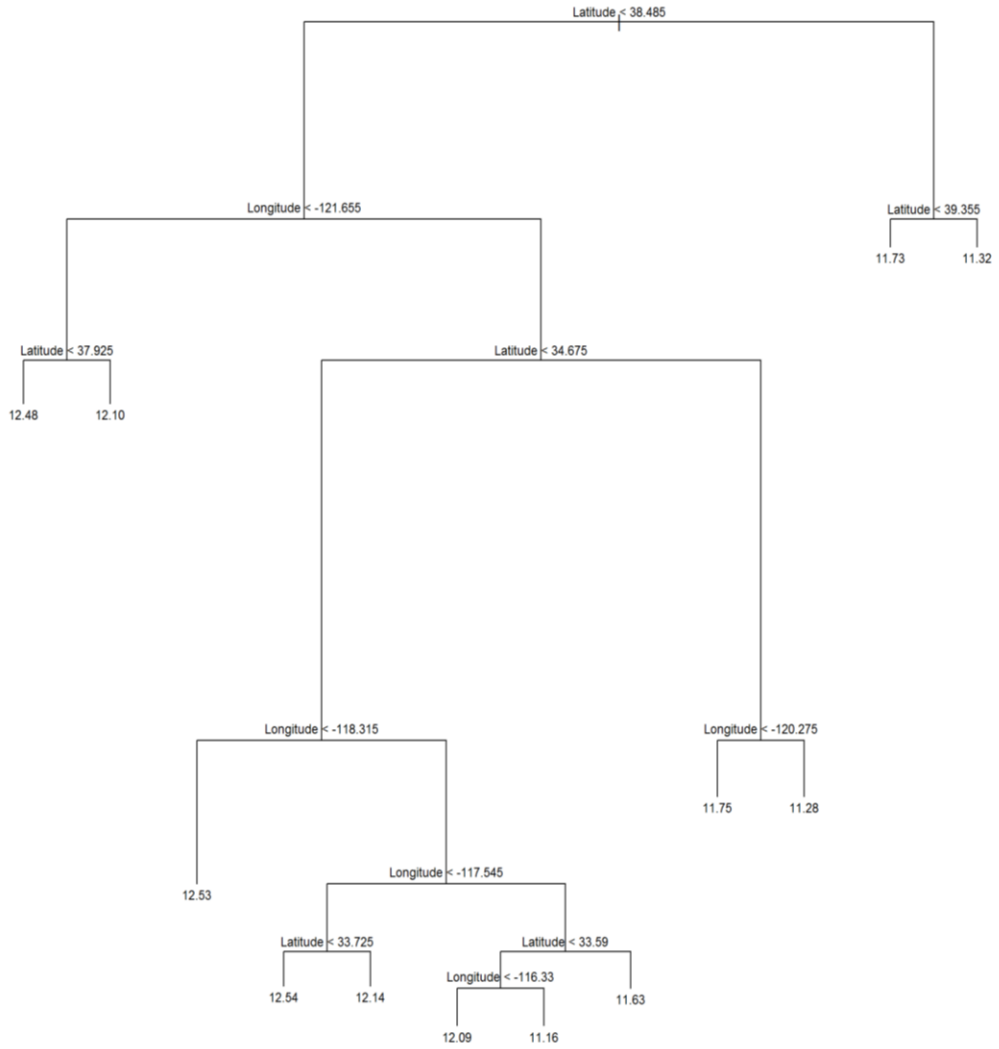
- **zvezna ciljna spremenljivka** – regresijski problem
- regresijska drevesa so podobna odločitvenim drevesom, le za regresijske probleme
- sistemi: CART (Breiman et al. 1984), RETIS (Karalič 1992), M5 (Quinlan 1993), WEKA (Witten and Frank, 2000)

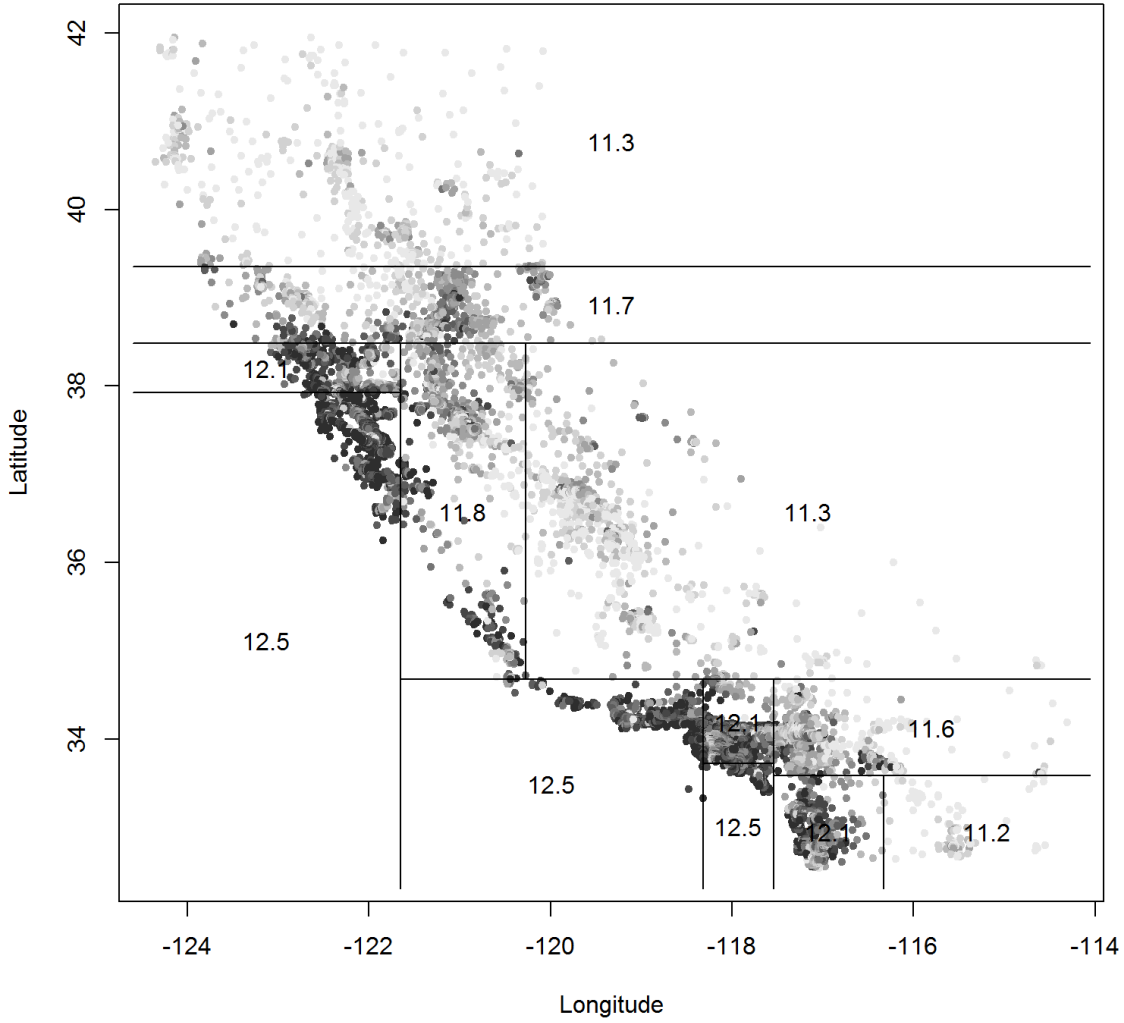
- listi v regresijskem drevesu predstavljajo bodisi:
  - **povprečno vrednost** označb ("razreda") primerov v listu
  - **preprost napovedni model** (npr. linearna regresija) za nove primere



Bora Beran, On Anything Data, domena MPG

Regresijska drevesa





João Neto, Classification & Regression Trees, http://www.di.fc.ul.pt/~jpn/r/tree/tree.html

Gradnja regresijskih dreves

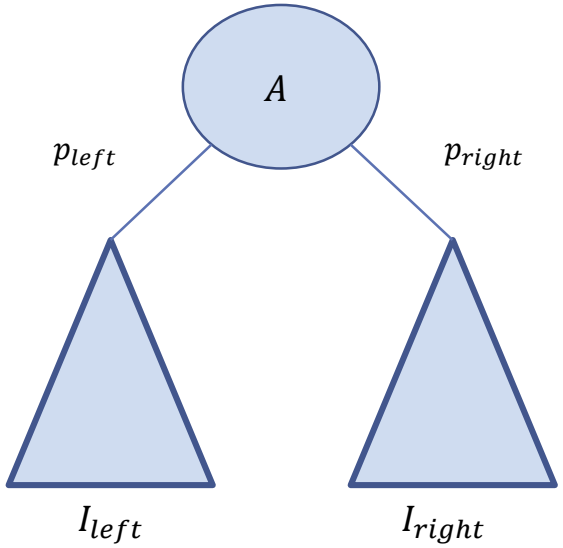
- drugačna mera za merjenje nedoločenosti/nečistoče: srednja kvadratna napaka v vozlišču v:

$$MSE(v) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2$$

- cilj: minimiziramo rezidualno nedoločenost po delitvi primerov glede na vrednosti atributa A

- pričakovana rezidualna nečistost

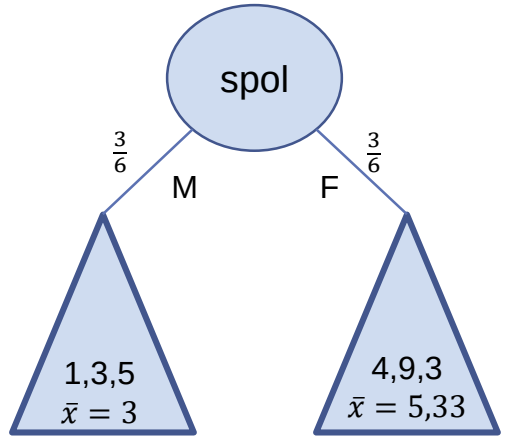
$$I_{res}(A) = p_{left} \cdot I_{left} + p_{right} \cdot I_{right}$$



Primer

- napovedovanje števila točk pri igri

spol	konzola	točke
M	T	1
M	T	3
M	F	5
F	T	4
F	T	9
F	F	3

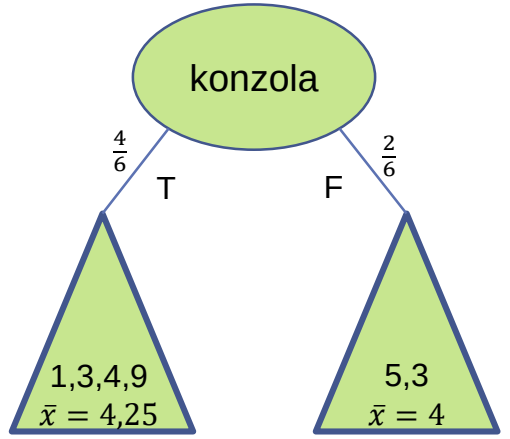


$$I_{res}(A) = p_{left} \cdot I_{left} + p_{right} \cdot I_{right}$$

$$MSE(v) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2$$

$$I_{res}(spol) = \frac{3}{6}\left[\frac{(1-3)^2+(3-3)^2+(5-3)^2}{3}\right]+\frac{3}{6}\left[\frac{(4-5{,}33)^2+(9-5{,}33)^2+(3-5{,}33)^2}{3}\right]=4{,}77$$

$$I_{res}(konzola) = \frac{4}{6}\left[\frac{(1-4{,}25)^2+(3-4{,}25)^2+(4-4{,}25)^2+(9-4{,}25)^2}{4}\right]+\frac{2}{6}\left[\frac{(5-4)^2+(3-4)^2}{2}\right]=6{,}125$$



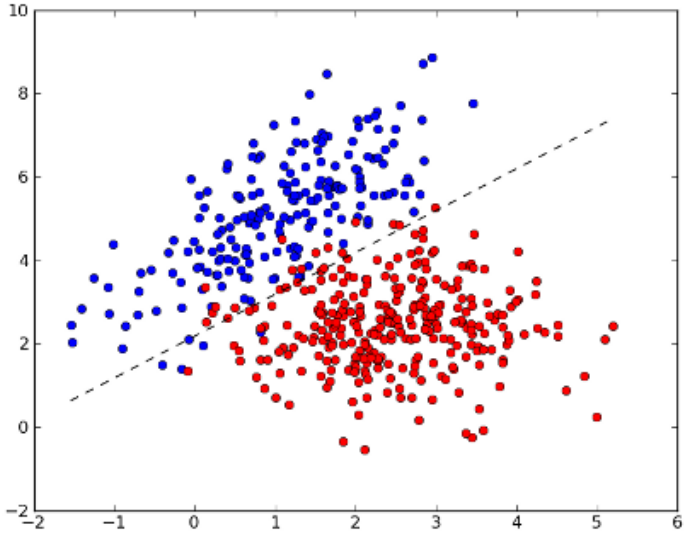
Linearni modeli

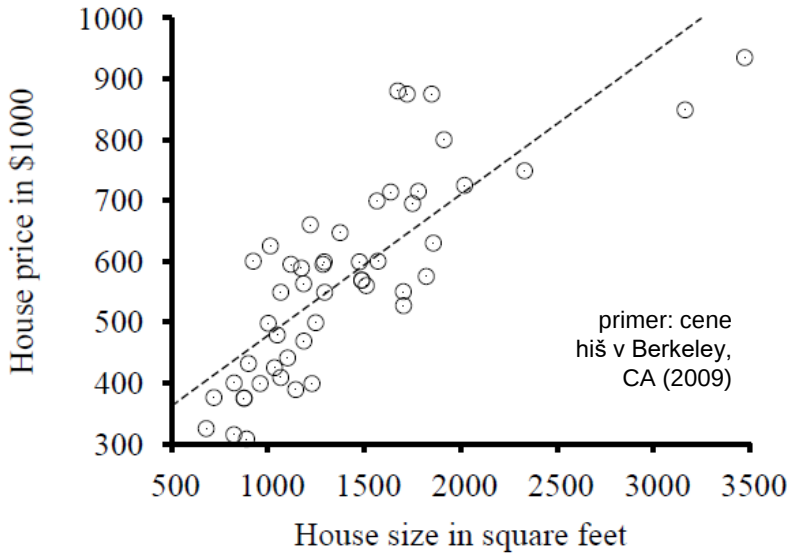
- uporaba pri **klasifikaciji** (kot separator razredov) in **regresiji** (kot prileganje skozi podane točke)
- linearni model z **eno odvisno** spremenljivko (angl. *univariate linear model*):

$$h(x) = w_1 x + w_0$$

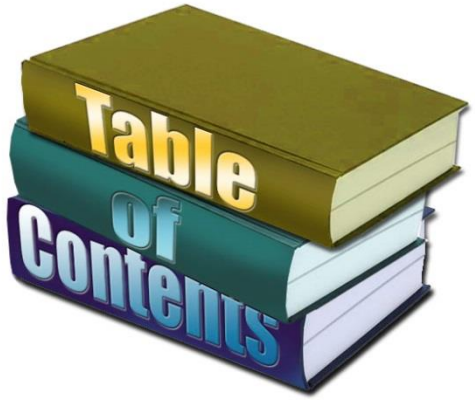
  w_0 in w_1 sta **uteži** (angl. *weights*) spremenljivk (koeficienta)

- **linearna regresija**: postopek iskanja funkcije $h(x)$ (oziroma uteži w_0 in w_1), ki se najbolje prilega učnim podatkom





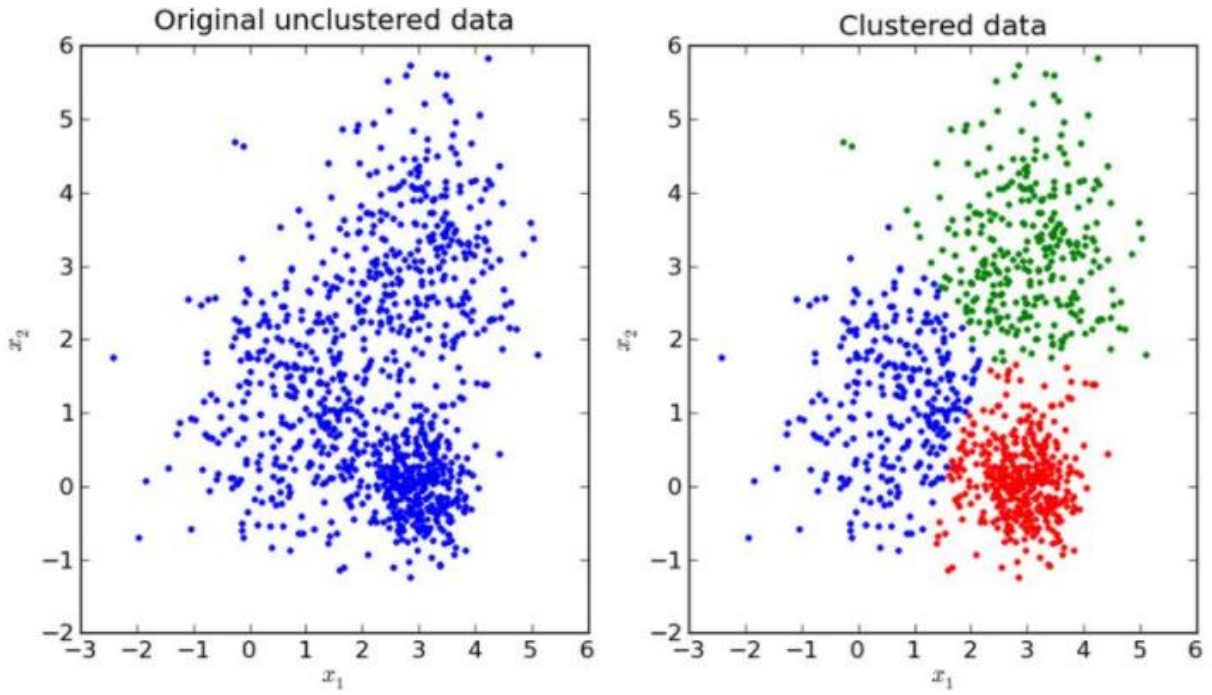
Pregled



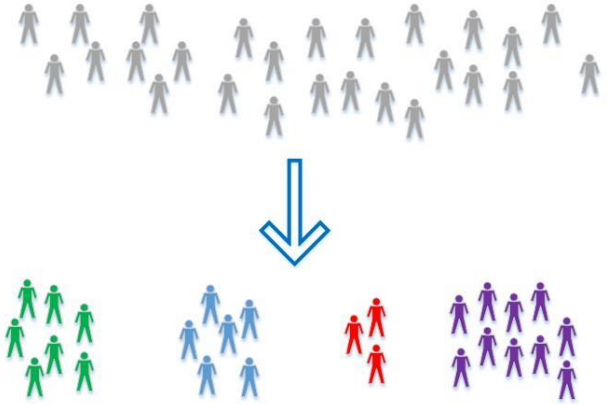
- strojno učenje
  - uvod v strojno učenje
  - učenje odločitvenih dreves
  - učenje dreves iz šumnih podatkov (rezanje dreves)
  - ocenjevanje učenja
  - diskretizacija atributov, obravnava manjkajočih vrednosti
  - naivni Bayesov klasifikator
  - nomogrami
  - k najbližjih sosedov
  - lokalna utežena regresija
  - regresijska drevesa
  - nenadzorovano učenje

Nenadzorovano učenje

- drugačni scenarij in cilji učenja kot pri nadzorovanem učenju:
  - **nimamo ciljne (odvisne) spremenljivke**, zato nas ne zanima napoved primera
  - podani so samo atributi primerov
- cilj: odkrivanje zakonitosti glede porazdelitve učnih primerov. Vprašanja:
  - ali lahko primere razdelimo v **smiselne skupine**?
  - ali obstaja priročen način za **vizualizacijo** podatkov?



Nenadzorovano učenje

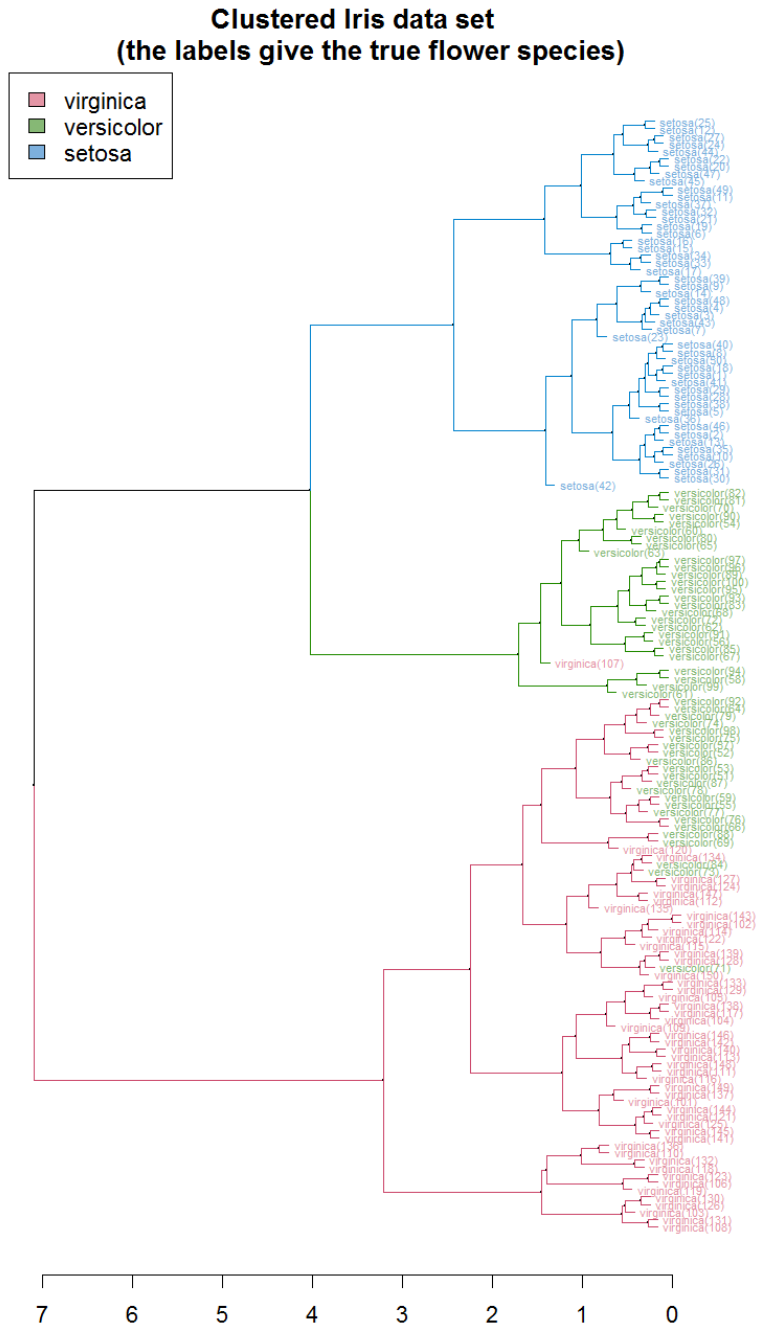


- **lastnosti**:
  - nenadzorovano učenje je **bolj subjektivno** kot nadzorovano učenje, ker nima enoznačnega formalnega cilja kot je "napovedovanje vrednosti odvisne spremenljivke" pri nadzorovanem učenju
  - velikokrat lažje **(cenejše) pridobimo neoznačene podatke** (podatke brez odvisne spremenljivke): drage meritve, ekspertno mnenje, globalna ocena (npr. filma)?

- **primeri uporabe**:
  - odkrivanje skupin rakavih bolnikov, grupiranih po različnih rezultatov meritev izraženosti genov,
  - odkrivanje skupin kupcev, grupiranih po njihovi zgodovini brskanja in nakupovanja
  - odkrivanje skupin filmov, grupiranih glede na ocene, podane s strani gledalcev

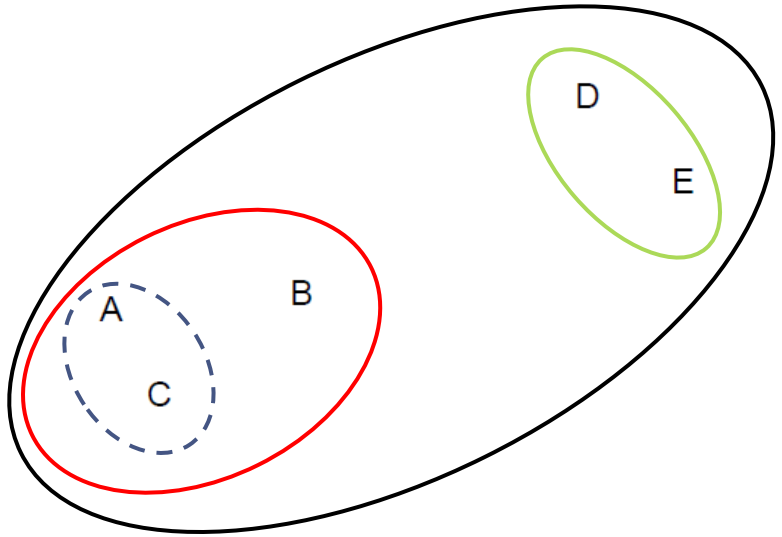
Gručenje

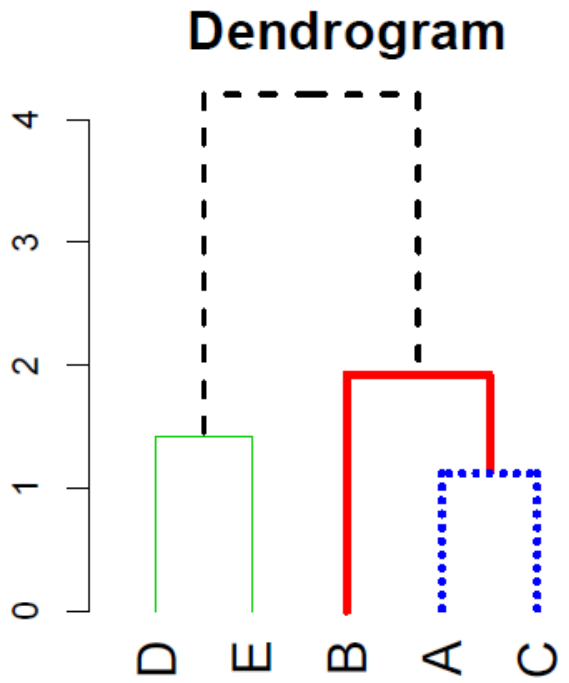
- **gručenje (angl. clustering)** je najbolj uporabljana metoda nenadzorovanega učenja
- **cilj**: iskanje homogenih podskupin v učnih podatkih
- metode:
  - **hierarhično gručenje**: iščemo vnaprej neznano število gruč. Rezultat gručenja je vizualna reprezentacija skupin, imenovana dendrogram, ki nam nudi vpogled v oblikovanje različnega števila gruč
  - **metoda k-means**: optimizacijski algoritem, ki poskuša iterativno primere gručiti v vnaprej podano število *k* gruč



Hierarhično gručenje

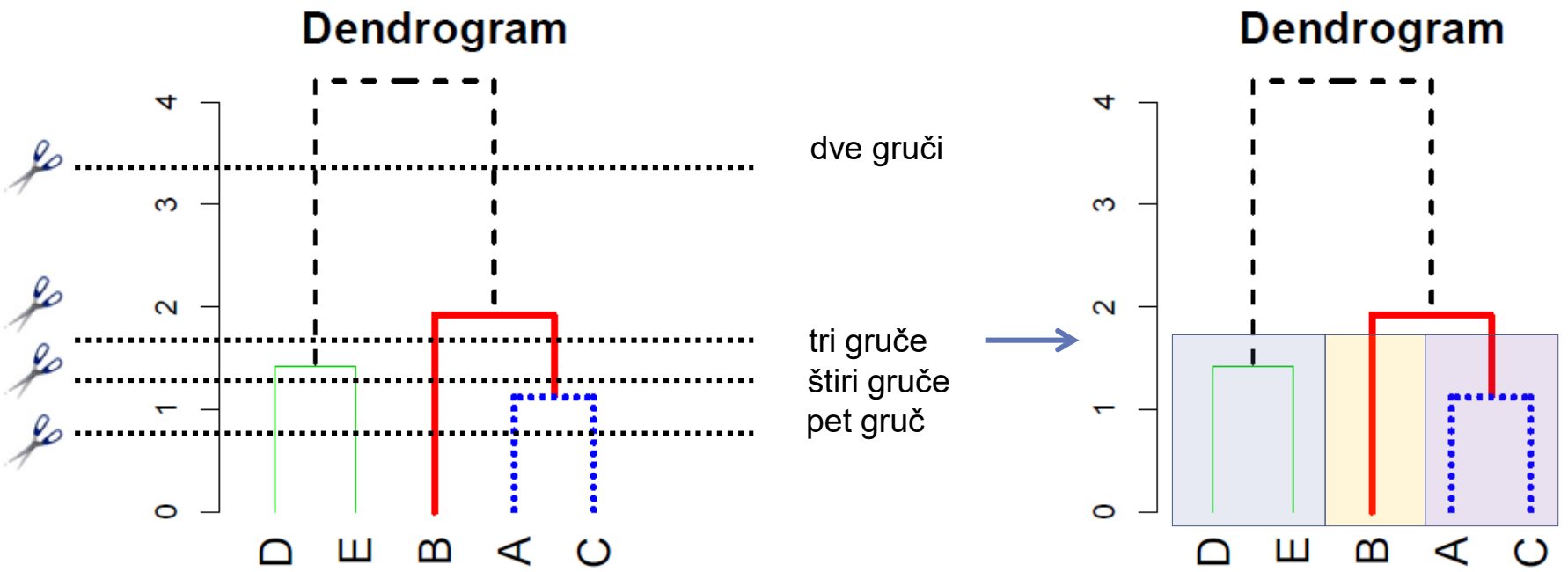
- dva pristopa:
  - **združevalni** (angl. *agglomerative*): gradnja dendrograma začenši od listov proti korenu s postopkom **združevanja** glede na razdaljo
  - **delilni** (angl. *divisive*): gradnja dendrograma od korena proti listom, na vsakem koraku **delimo** gručo na podgruče
- primer združevalnega pristopa:
  1. začni z vsako točko v svoji gruči
  2. najdi dve najbližji gruči in ju združi
  3. ponavljaj, dokler ne združiš vseh gruč





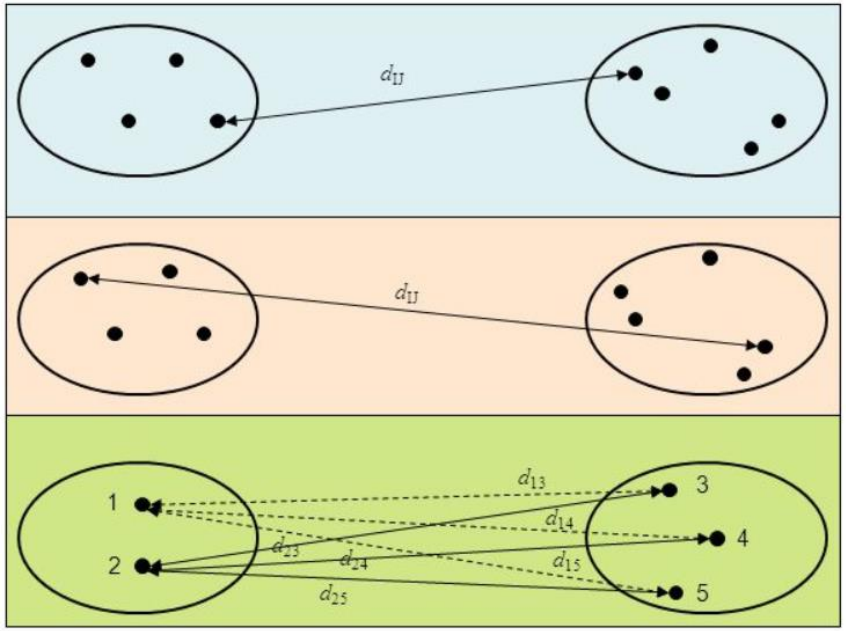
Interpretacija dendrograma

- rezanje dendrograma določi mejo, pri kateri prenehamo z združevanjem gruč
- z rezanjem dendrograma na različnih višinah torej določamo število ciljnih gruč



Merjenje razdalj

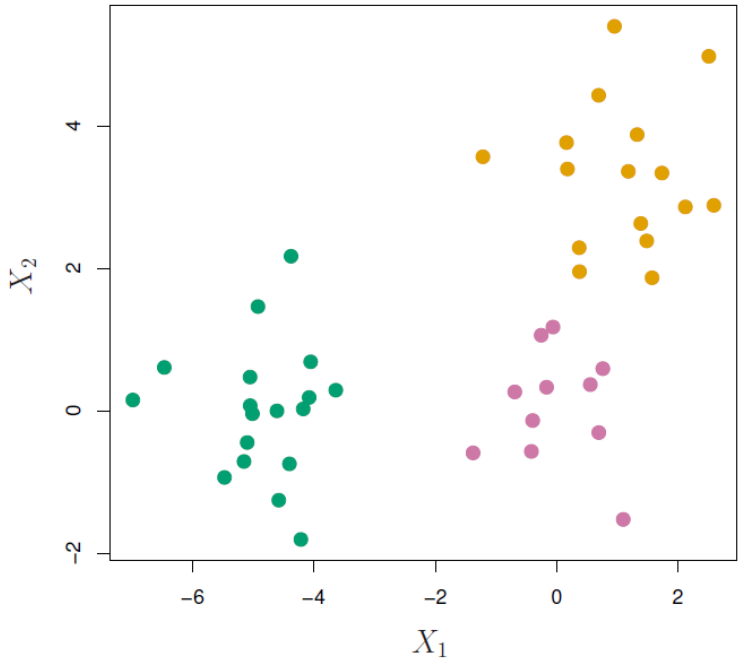
- med učnimi primeri uporabljamo že znane mere za merjenje razdalj (evklidska razdalja, manhattanska razdalja, korelacija med vrednostmi atributov ...)
- posebno obravnavo moramo posvetiti merjenju razdalj:
  - med **posameznim učnim primerom** in **gručo**
  - med **dvema gručama**
- kot razdaljo v teh primerih lahko upoštevamo:

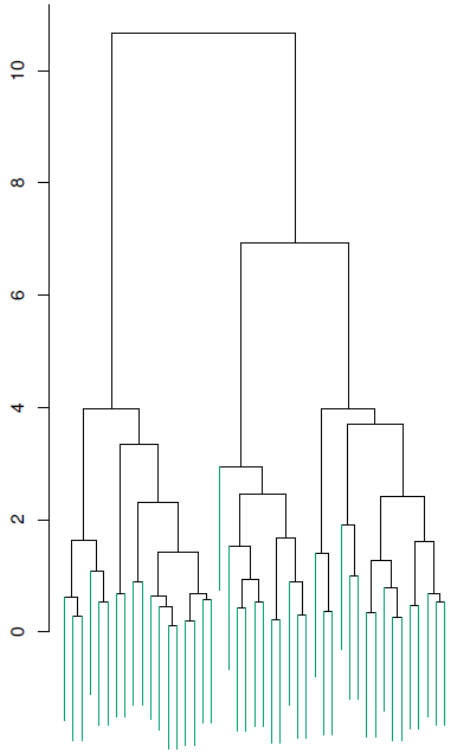


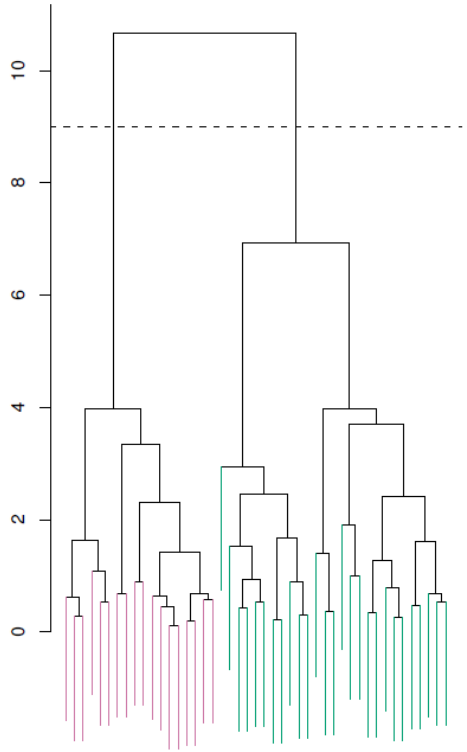
- razdaljo med **najbližjima** primeroma (enojna povezanost, angl. *single linkage*)
  $d(C_1, C_2) = \min_{i,j}\{d_{ij} | i \in C_1, j \in C_2\}$
- razdaljo med **najbolj oddaljenima** primeroma (popolna povezanost, angl. *complete linkage*)
  $d(C_1, C_2) = \max_{i,j}\{d_{ij} | i \in C_1, j \in C_2\}$
- **povprečno razdaljo** med vsemi primeri (povprečna povezanost, angl. *average linkage*)
  $d(C_1, C_2) = \sum_{i \in C_1, j \in C_2} \frac{d_{ij}}{|C_1||C_2|}$

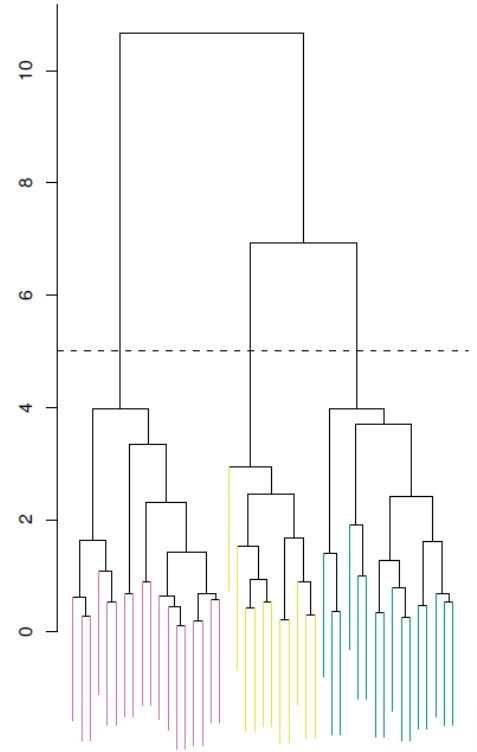
Primer

- 45 primerov, 2 atributa (oznaka razreda je skrita pred algoritmom za gručenje)
- uporabljena evklidska razdalja in merjenje razdalj s polno povezanostjo (complete linkage)
- dendrogram prikazuje rezanja na različnih višinah







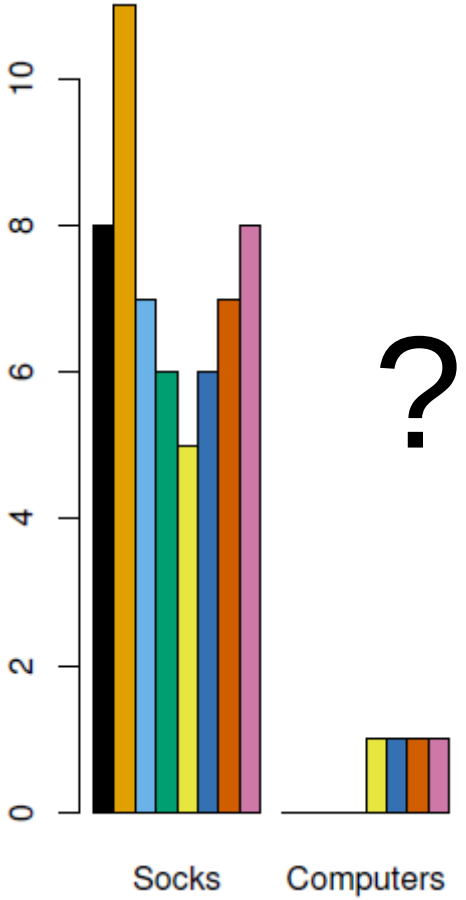


Opombe

- **normalizacija atributov** (glej desno sliko →)

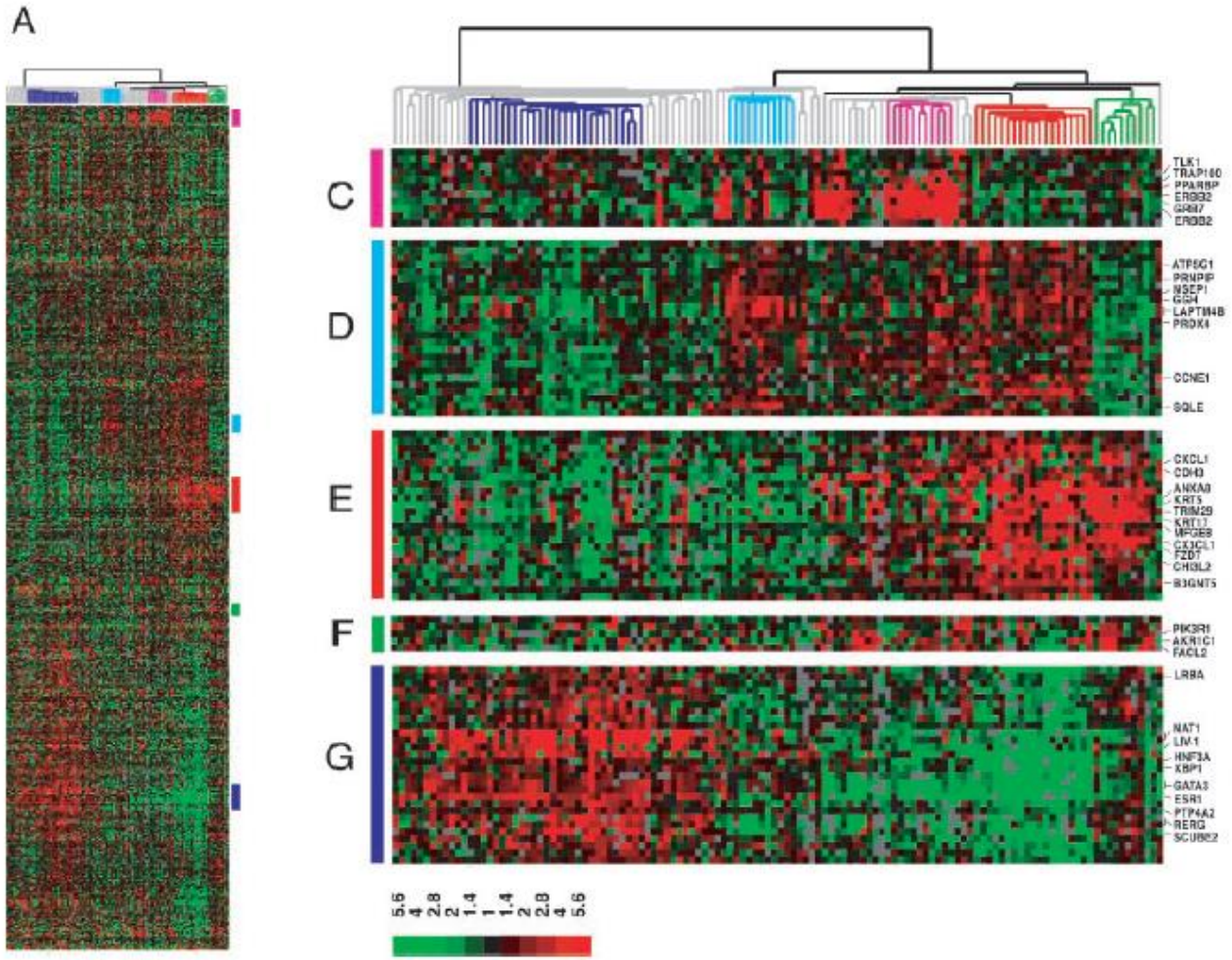
- **časovna zahtevnost**:
  - združevalni pristop: $O(n^2 log n)$:
    n^2 časa za izračun matrike razdalj, $log n$ za urejanje razdalj
  - delilni pristop: $O(2^n)$:
    za iskanje optimalne delitve na dve podgruči

- **parametri**
  - katero mero razdalje izbrati?
  - kateri pristop merjenja razdalj med gručami izbrati?
  - kolikšno naj bo ciljno število gruč?



Primer uporabe

- analiza različnih vrst tumorja na prsih glede na izraženost genov
  Sørlie, Therese, et al. "Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications." Proceedings of the National Academy of Sciences 98.19 (2001): 10869-10874.
- mera razdalje: korelacija
- razdalja med gručami: povprečna razdalja med primeri
- atributi: izraženost 500 genov
- rezultati:
  - identifikacija sorodnih skupin pacientov
  - identifikacija izraženih genov v skupinah pacientov



Izpitna naloga

- 2. izpitni rok, 13. 2. 2019 (prilagojena naloga)

4. NALOGA (10t):

Podanih je pet točk z vrednostmi atributov X in Y, ki predstavljata koordinati na grafu.

točka	X	Y
A	1	1
B	3	1
C	1	3
D	3	2
E	4	3

a) **(5t) Izvedi algoritem hiearhičnega razvrščanja naštetih točk in nariši dendrogram. Uporabi Manhattansko razdaljo in pristop popolne povezanosti (angl. complete linkage) merjenja razdalj med gručami.**

b) **(3t) Dendrogram iz prejšnje naloge poreži tako, da dobimo dve gruči.**


Nenadzorovano učenje,
preiskovanje