OSNOVE UMETNE INTELIGENCE 2021/22

nomogrami

k najbližjih sosedov

regresija

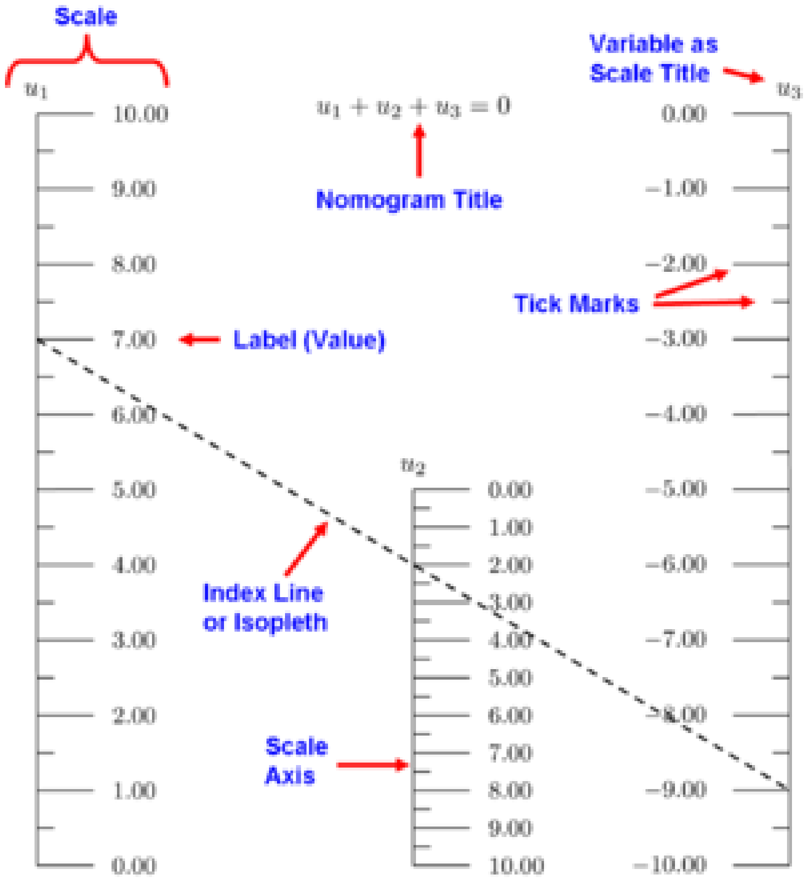
Pridobljeno znanje s prejšnjih predavanj

- **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)

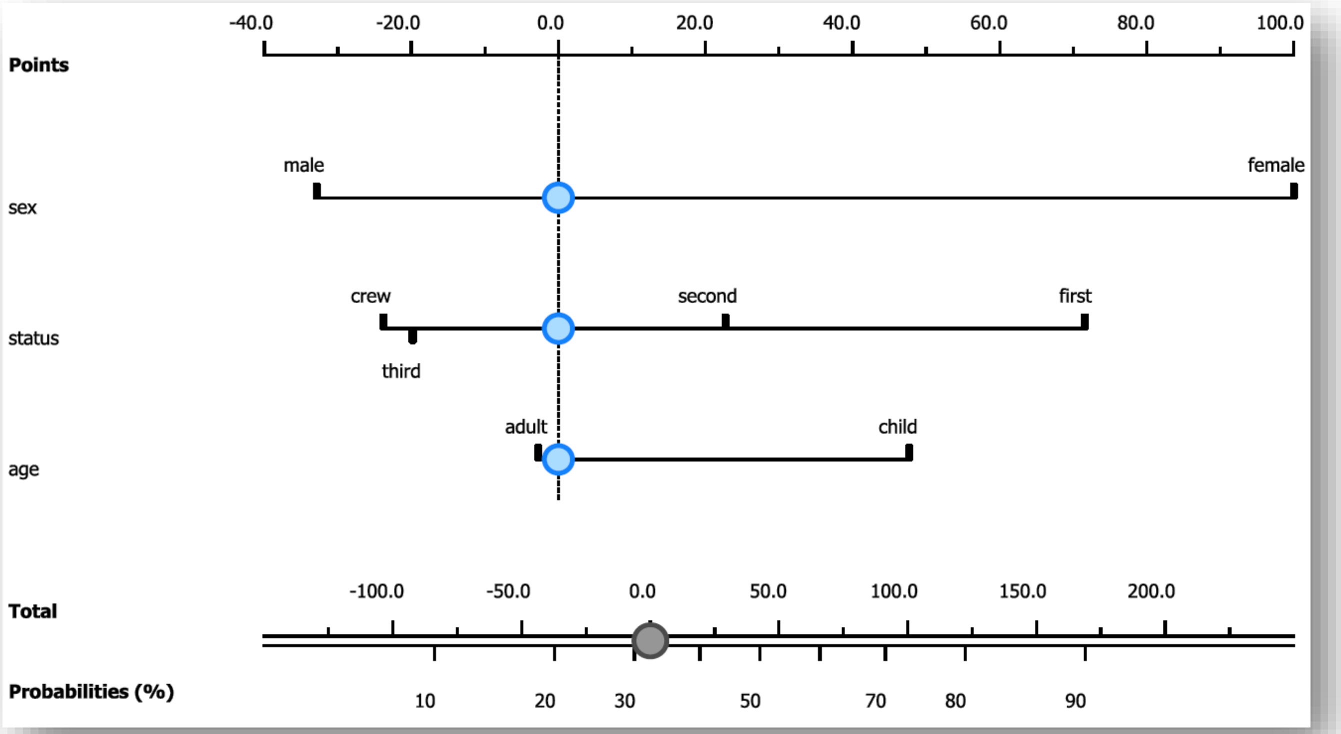
Nomogrami

- pristop za **vizualizacijo** naivnega Bayesovega modela

- "nomogram":
  - je *grafična upodobitev numeričnih odnosov med spremenljivkami*
  - omogoča uporabniku grafično pridobiti rezultat brez računanja
- uporaba:
  - matematika (iskanje vrednosti funkcij)
  - zdravniki v medicini (napovedovanje bolezni – npr. infarkta ali raka na podlagi vhodnih atributov)
- prikazuje:
  - pomembnost posameznih **vrednosti** vsakega atributa na ciljni razred
  - pomembnost posameznih **atributov** na ciljni razred
  - **vizualno razlago** napovedanih verjetnosti (brez kalkulatorja)



Primer - ideja



- vsaka **vrednost atributa doprinaša določeno število točk** k ciljnemu razredu
- točke vseh vrednosti atributov seštejemo v **skupno vsoto točk**, ki je povezana z verjetnostjo ciljnega razreda
- **razpon točk vsakega atributa** govori o pomembnosti atributa za napovedovanje ciljnega razreda (zgoraj urejeni od najbolj do najmanj pomembnega)

Izračun nomograma

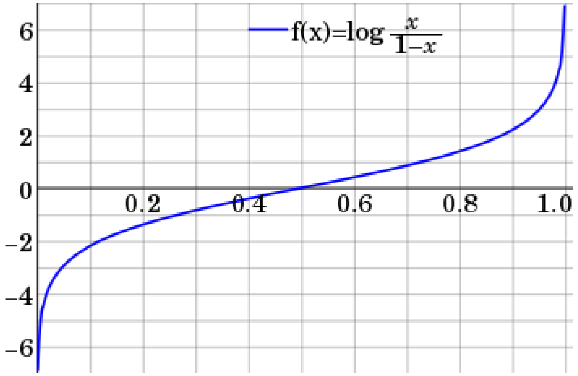
- verjetje razreda pri naivnem Bayesu:

$$h(C|X_1X_2 \ldots X_n) = P(C) \cdot \prod_{i=1}^{n} P(X_i|C)$$

- na zgornjem pravilu uporabimo logistično funkcijo (verjetnosti z intervala [0,1] preslikamo na interval $(-\infty, \infty)$, uporabimo logaritme)

$$\text{logit}\, P = \log \frac{P}{1-P}$$

logistična funkcija



- $\text{logit}\, h(C|X_1X_2 \ldots X_n) =$

$$\log \frac{P(C) \cdot \prod_{i=1}^{n} P(X_i|C)}{1 - P(C) \cdot \prod_{i=1}^{n} P(X_i|C)} = \log \frac{P(C) \cdot \prod_{i=1}^{n} P(X_i|C)}{P(\bar{C}) \cdot \prod_{i=1}^{n} P(X_i|\bar{C})} = log \frac{P(C)}{P(\bar{C})} + \log \frac{\prod_{i=1}^{n} P(X_i|C)}{\prod_{i=1}^{n} P(X_i|\bar{C})}$$

$$= logit\, P(C) + \sum_i \log \frac{P(X_i|C)}{P(X_i|\bar{C})}$$

samo ta člen je odvisen od vrednosti atributov X_i, to količino imenujemo razmerje verjetja (odds ratio)

Izračun nomograma

to količino imenujemo
razmerje verjetja
(odds ratio)

- $\operatorname{logit} P(C) + \sum_i \log \frac{P(X_i|C)}{P(X_i|\bar{C})} = \operatorname{logit} P(C) + \sum_i \log OR(X_i)$

- od zgornjih členov je edino razmerje verjetja (drugi člen) **odvisno od vrednosti atributov $\boldsymbol{X_i}$**, torej ga lahko uporabimo za "točkovanje" doprinosa atributa:

$$točke(C|X_i) = \log OR(X_i) = \log \frac{P(X_i|C)}{P(X_i|\bar{C})}$$

- **skupno število točk** za verjetnost celotnega primera (upoštevamo vse attribute):

$$točke(C|X_1 X_2 \ldots X_n) = \sum_i \log OR(X_i) = \sum_i \log \frac{P(X_i|C)}{P(X_i|\bar{C})}$$

- Kako izračunati **izraz znotraj logaritma $\boldsymbol{OR(X_i)}$**? Po Bayesovem pravilu (znova) velja:

$$\frac{P(X_i|C)}{P(X_i|\bar{C})} = \frac{\frac{P(C|X_i) \cdot P(X_i)}{P(C)}}{\frac{P(\bar{C}|X_i) \cdot P(X_i)}{P(\bar{C})}} = \frac{\frac{P(C|X_i)}{P(C)}}{\frac{P(\bar{C}|X_i)}{P(\bar{C})}} = \frac{\frac{P(C|X_i)}{P(\bar{C}|X_i)}}{\frac{P(C)}{P(\bar{C})}}$$

Primer

- učna množica titanic, 2201 učnih primerov (711 preživelih – razred YES, 1490 umrlih – razred NO)

atribut	vrednost	razred = YES	razred = NO
status	first	203	122
	second	118	167
	third	178	528
	crew	212	673
age	adult	654	1438
	child	57	52
sex	male	367	1364
	female	344	126
razred		711	1490

- Kako konstruiramo nomogram?
- Kako lahko vizualiziramo odločitev za **odraslega moškega**, ki je potoval v **drugem** razredu?

primer iz Možina, Martin, et al. "Nomograms for visualization of naive Bayesian classifier." European Conference on Principles of Data Mining and Knowledge Discovery. Springer, Berlin, Heidelberg, 2004.

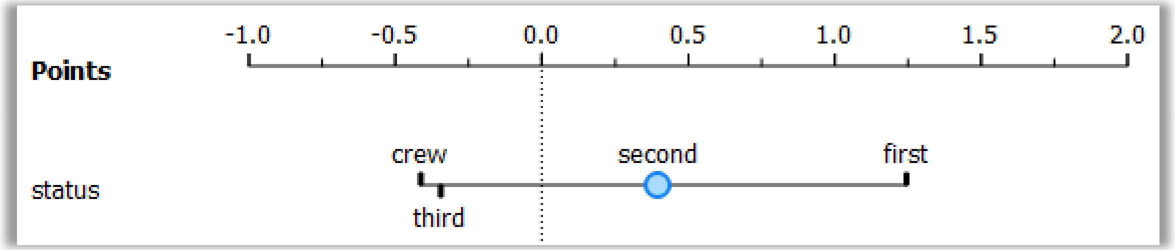
Konstrukcija nomograma

$$točke(yes|status = first) = \log\frac{\frac{P(yes|first)}{P(no|first)}}{\frac{P(yes)}{P(no)}} = \log\frac{\frac{203}{122}}{\frac{711}{1490}} = \log\frac{1{,}66}{0{,}48} = 1{,}25$$

$$točke(yes|status = second) = \log\frac{\frac{P(yes|second)}{P(no|second)}}{\frac{P(yes)}{P(no)}} = \log\frac{\frac{118}{167}}{\frac{711}{1490}} = \log\frac{0{,}71}{0{,}48} = 0{,}39$$

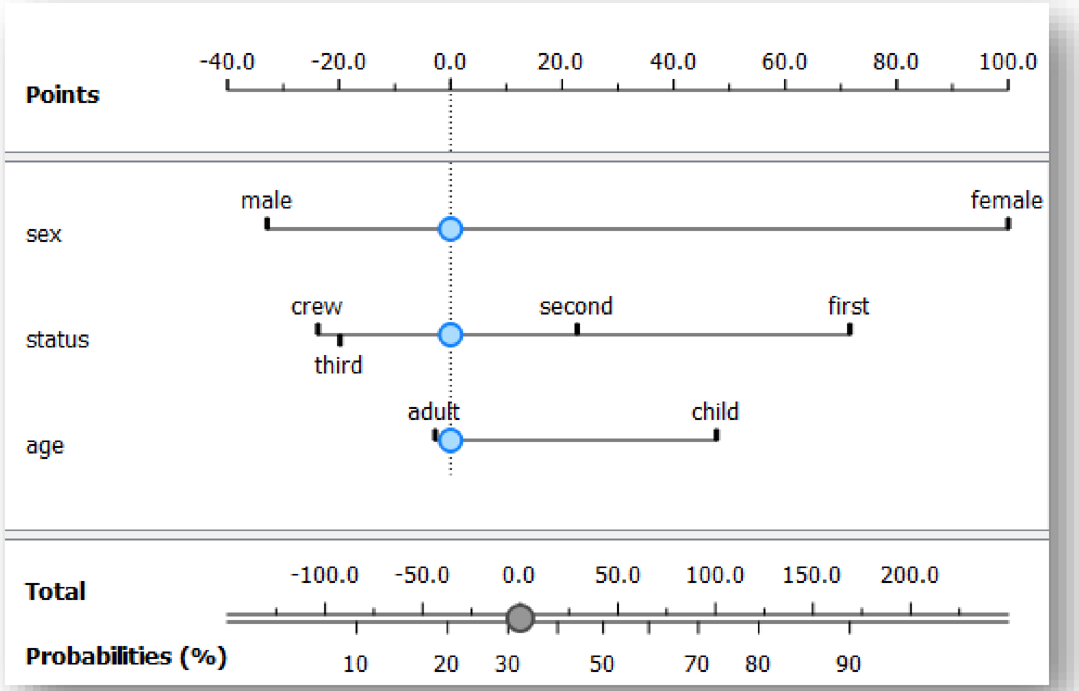
$$točke(yes|status = third) = \log\frac{\frac{P(yes|third)}{P(no|third)}}{\frac{P(yes)}{P(no)}} = \log\frac{\frac{178}{528}}{\frac{711}{1490}} = \log\frac{0{,}34}{0{,}48} = -0{,}35$$

$$točke(yes|status = crew) = \log\frac{\frac{P(yes|crew)}{P(no|crew)}}{\frac{P(yes)}{P(no)}} = \log\frac{\frac{212}{673}}{\frac{711}{1490}} = \log\frac{0{,}32}{0{,}48} = -0{,}42$$



Konstrukcija nomograma

- osi ostalih atributov **poravnamo glede na ničelno vrednost** prispevka atributa
- prikažemo lahko tudi **skupno skalo** za celotno napoved (vsoto točk)
- točke posameznih vrednosti atributov (log OR) lahko skaliramo v skalo točk, kjer s 100 točkami predstavimo prispevek največje vrednosti atributa



preslikane točke

skupna vsota točk

preslikava v verjetnosti

- skupne točke lahko preslikamo nazaj v verjetnosti

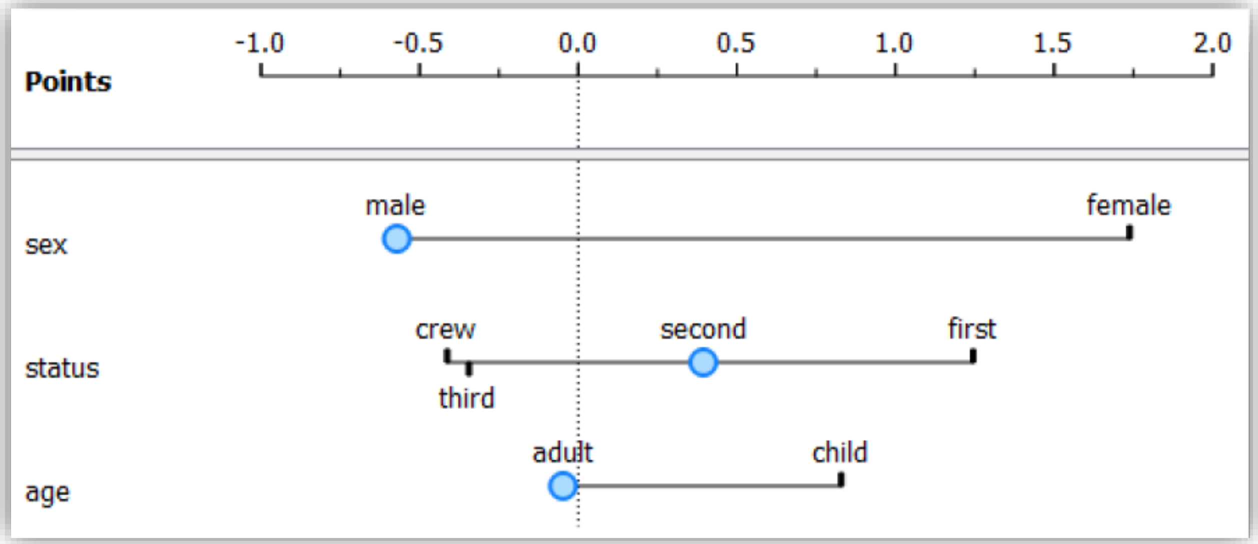
Neobvezna dodatna literatura: Možina, Martin, et al. "Nomograms for visualization of naive Bayesian classifier." European Conference on Principles of Data Mining and Knowledge Discovery. Springer, Berlin, Heidelberg, 2004.

Primer

- Kako lahko pojasnimo odločitev, da je **odrasli moški**, ki je potoval v **drugem** razredu, **preživel**?
- $$točke(yes|age = adult) = \log\frac{\frac{P(yes|adult)}{P(no|adult)}}{\frac{P(yes)}{P(no)}} = \log\frac{\frac{654}{1438}}{\frac{711}{1490}} = \log\frac{0{,}45}{0{,}48} = -0{,}05$$

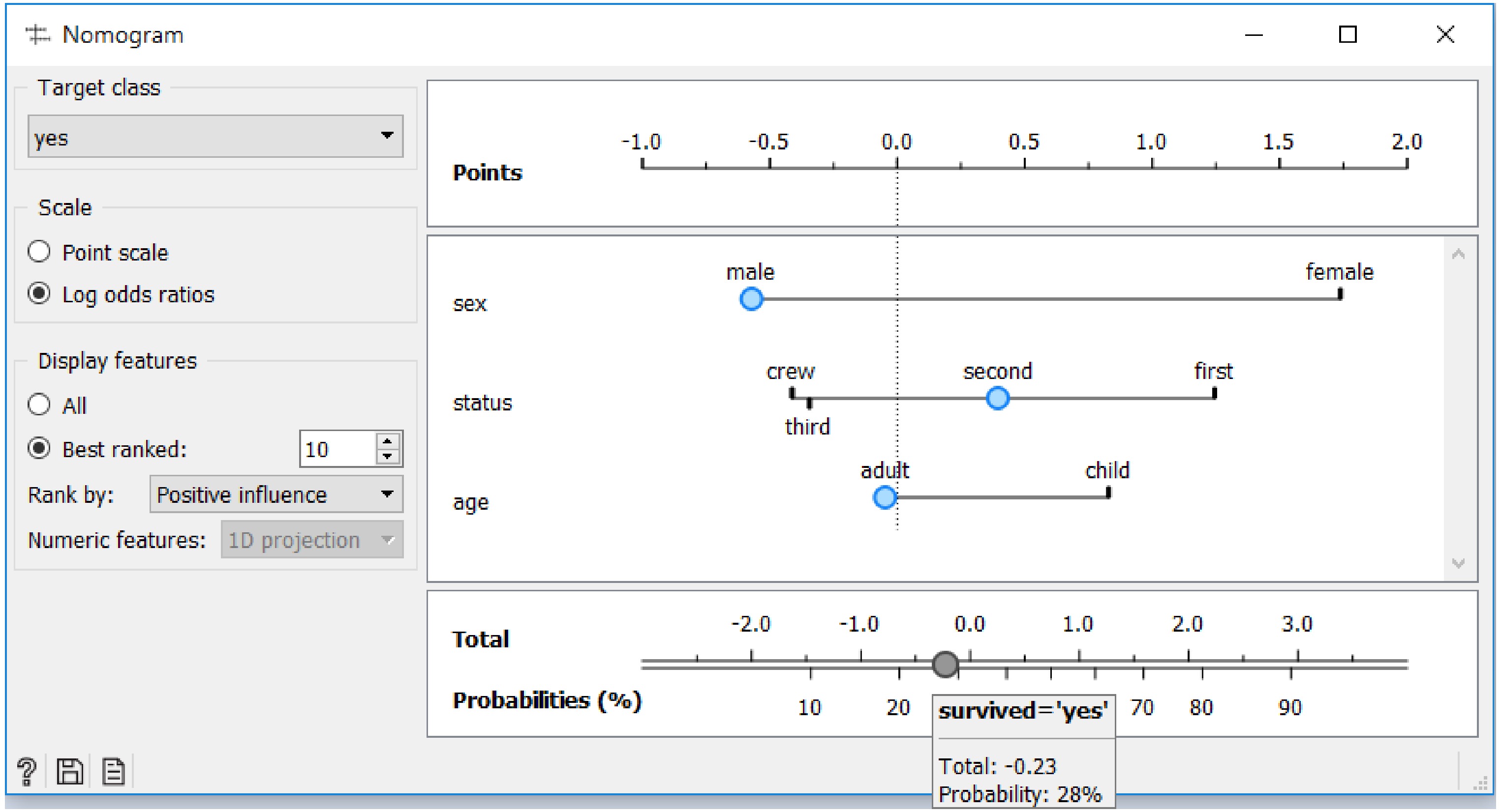
  $$točke(yes|sex = male) = \log\frac{\frac{P(yes|male)}{P(no|male)}}{\frac{P(yes)}{P(no)}} = \log\frac{\frac{367}{1364}}{\frac{711}{1490}} = \log\frac{0{,}27}{0{,}48} = -0{,}57$$

  $$točke(yes|status = second) = \log\frac{\frac{P(yes|second)}{P(no|second)}}{\frac{P(yes)}{P(no)}} = \log\frac{\frac{118}{167}}{\frac{711}{1490}} = \log\frac{0{,}71}{0{,}48} = 0{,}39$$



Primer

- Kako lahko pojasnimo odločitev, da je **odrasli moški**, ki je potoval v **drugem** razredu, **preživel**?
- $točke(yes|adult, male, second) = -0{,}05 - 0{,}57 + 0{,}39 = -0{,}23$



Izpitna naloga

- 3. izpitni rok, 2. 9. 2019

<u>3. NALOGA (10t):</u>

Podana je učna množica podatkov o pacientih, ki so zboleli za srčno boleznijo. Ta vsebuje dva atributa:

- **družina**: če je za sorodno boleznijo zbolel tudi še kak ožji družinski član (vrednosti: da/ne)
- **spol**: spol pacienta (vrednosti: moški/ženska).

  Vsak učni primer je označen z razredom **bolezen**, ki pove, ali je pacient zbolel (vrednosti: da/ne).

Povzetek števila učnih primerov s posameznimi vrednostmi atributov je podan z naslednjo tabelo:

		bolezen	
		DA	NE
družina	da	200	150
	ne	120	110
spol	moški	140	160
	ženska	180	100

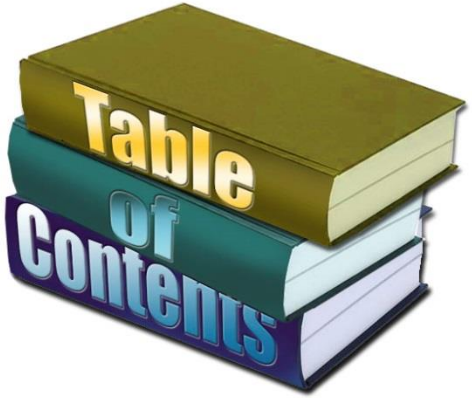
b.) Nariši nomogram za verjetnostno razlago modela za klasificiranje v razred bolezen=DA. Nomogram naj prikazuje:

- **skalo za prispevke atributa, ki naj prikazuje izračunano število točk (skaliranje ni potrebno),**
- **os za vsak atribut z jasno označenimi prispevki njegovih posameznih vrednosti.**

Pri izračunu uporabljaj naravni logaritem.

c) Pacient s kakšnimi lastnostmi ima največjo verjetnost, da ostane zdrav (pojasni odgovor)?

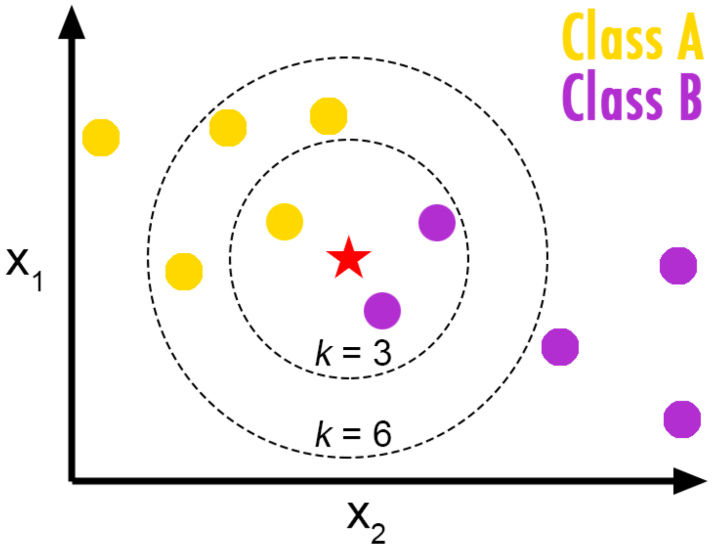
Pregled



- 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

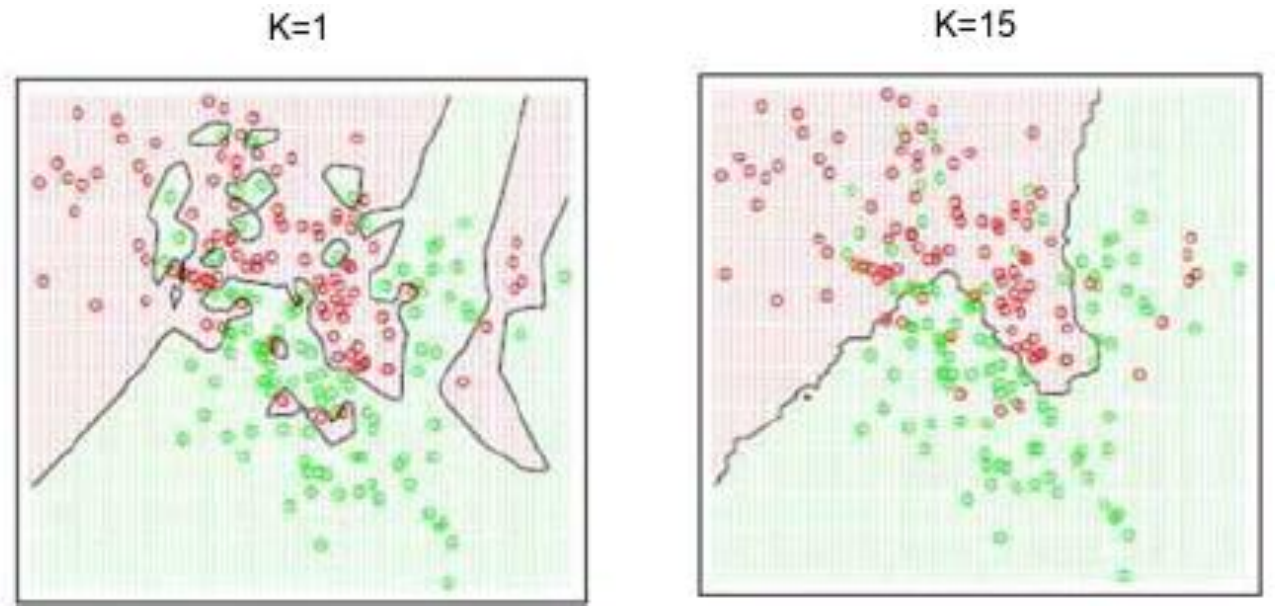
Metoda k najbližjih sosedov

- 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

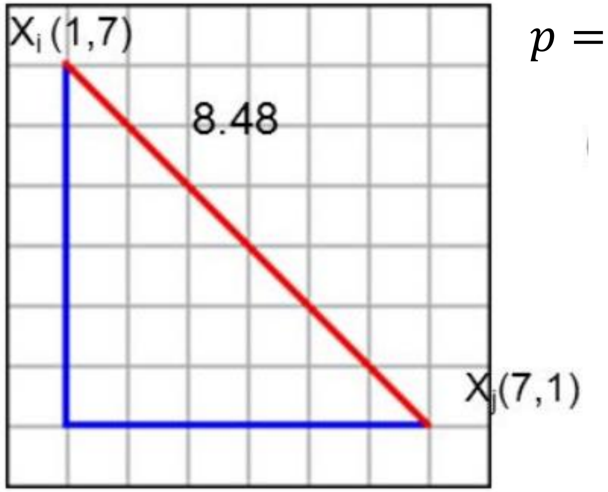
- pomembna je izbira ustreznega k:
  - premajhen k: pretirano prilagajanje
  - prevelik k: prešibko prilagajanje (pri $k = N$: napoved večinskega razreda)
  - v praksi običajno: $k = 5$



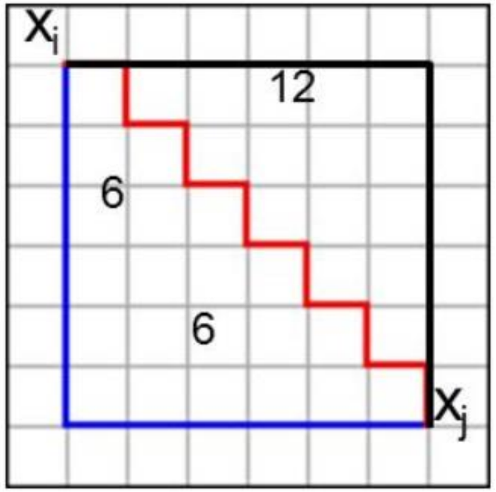
Hastie, Tibshirani, Friedman: Elements of Statistical Learning, 2009

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)



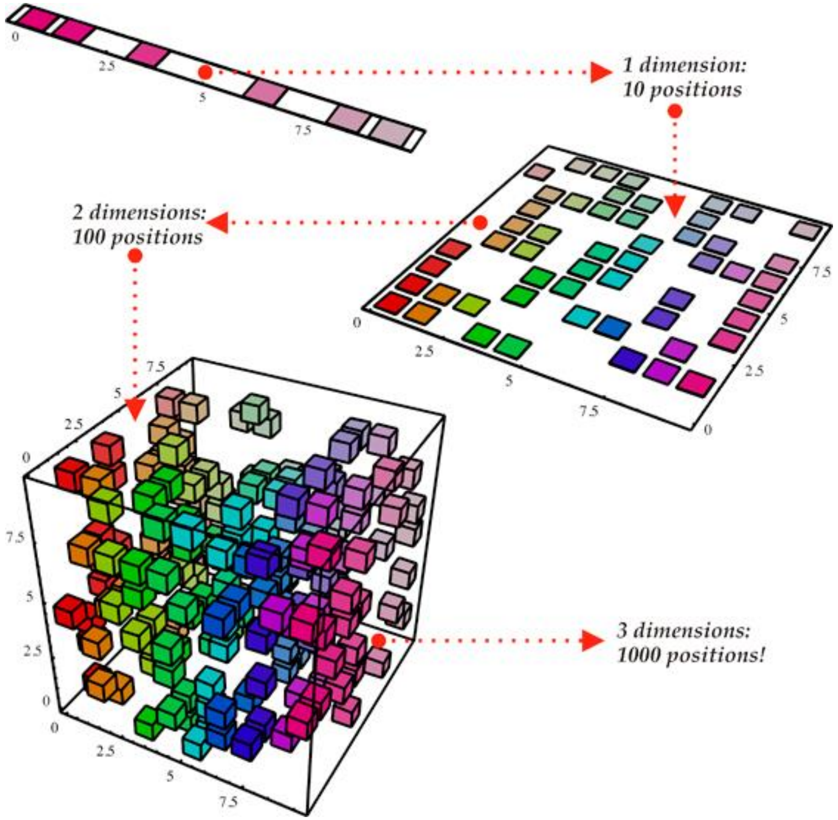
$p = 2$

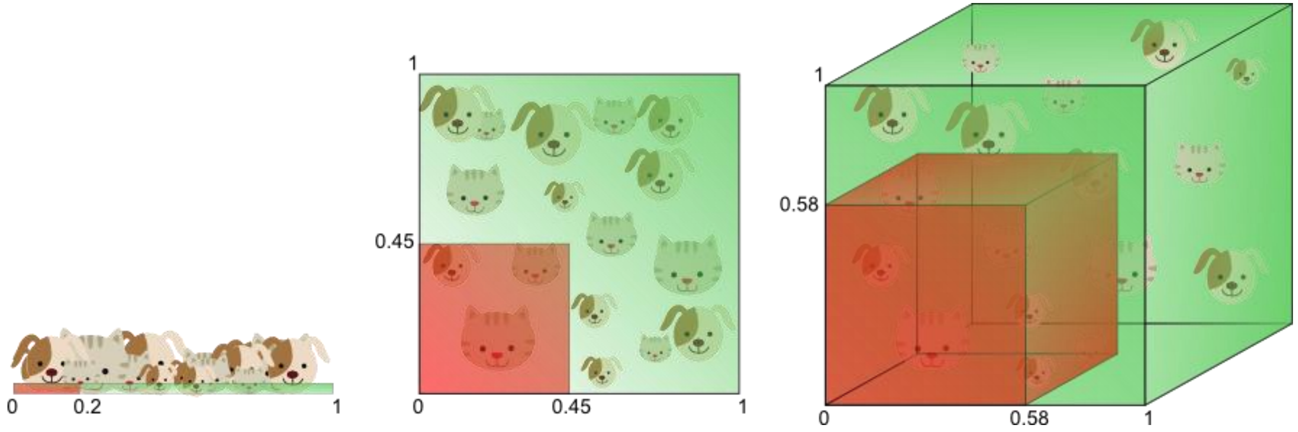


$p = 1$

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)

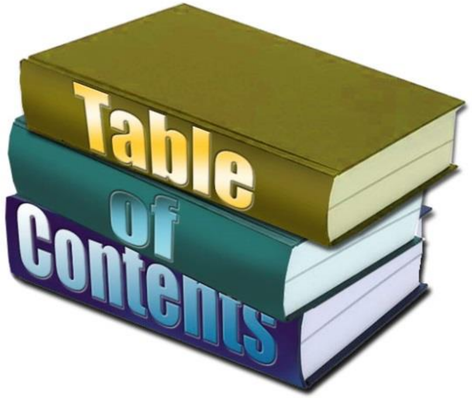
2. NALOGA (25t):

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

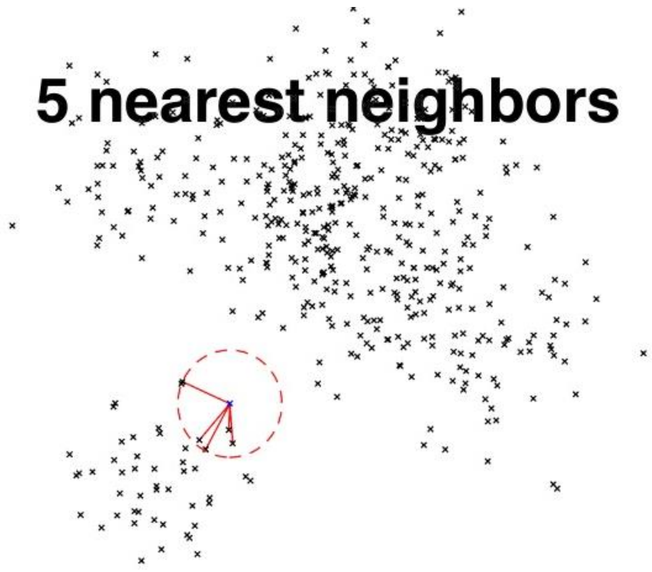


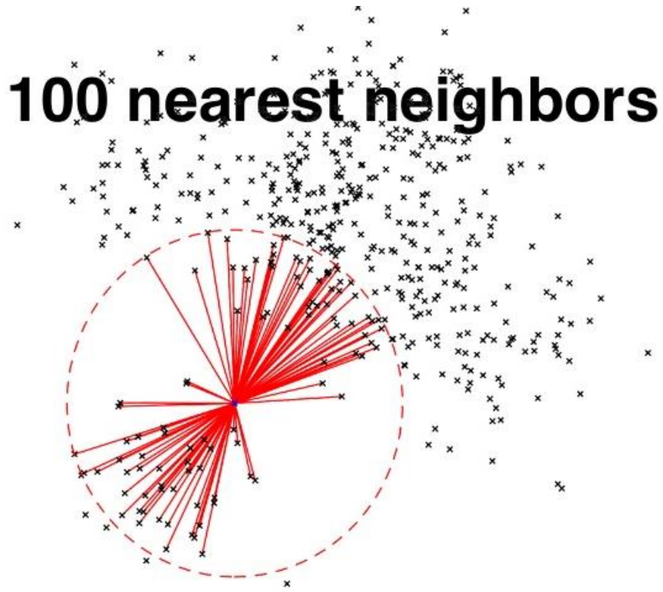
- strojno učenje
  - uvod v strojno učenje
  - učenje odločitvenih dreves
  - učenje dreves iz šumnih podatkov (rezanje dreves)
  - ocenjevanje učenja
  - diskretizacija atributov, obravnava manjkajočih vrednosti
  - naivni Bayesov klasifikator
  - nomogrami
  - k najbližjih sosedov
  - lokalna utežena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

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

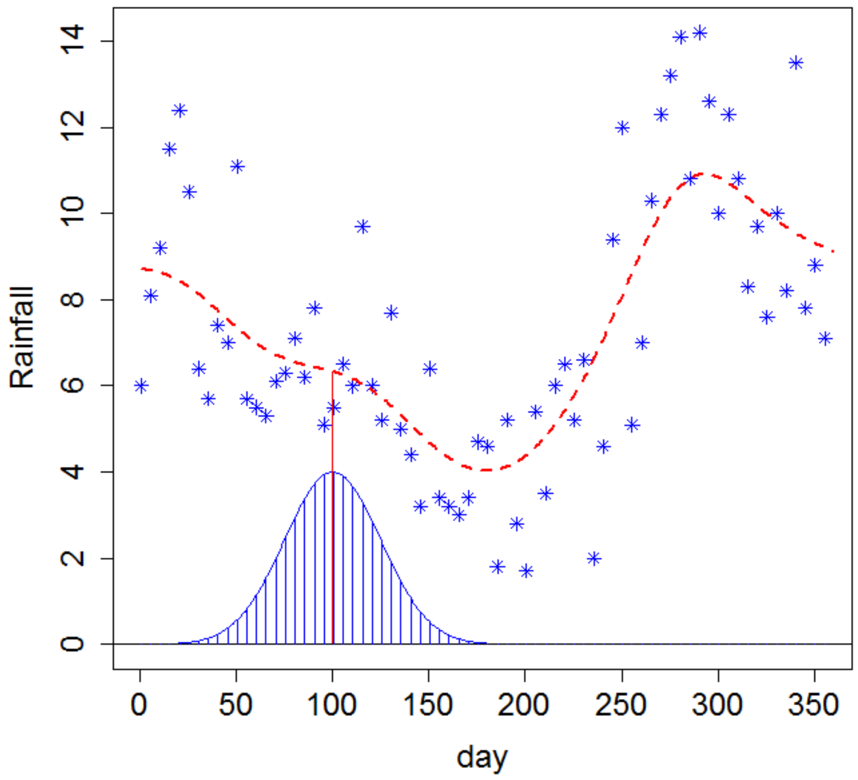
- **uteževanje z razdaljo** (lokalno utežena regresija)
  - $h(x_?) = \frac{\sum_{i=1}^{k} w_i \cdot f(x_i)}{\sum_{i=1}^{k} w_i}$
  - w_i je utež, ki je lahko enaka $w_i = \frac{1}{\left(d(x_?, x_i)\right)^2}$
  - pri uteževanju se lahko uporablja tudi **poljubna jedrna funkcija, npr. Gaussovo jedro:** $\frac{1}{\sqrt{2\pi}} e^{-\frac{\left(d(x_?, x_i)\right)^2}{2}}$

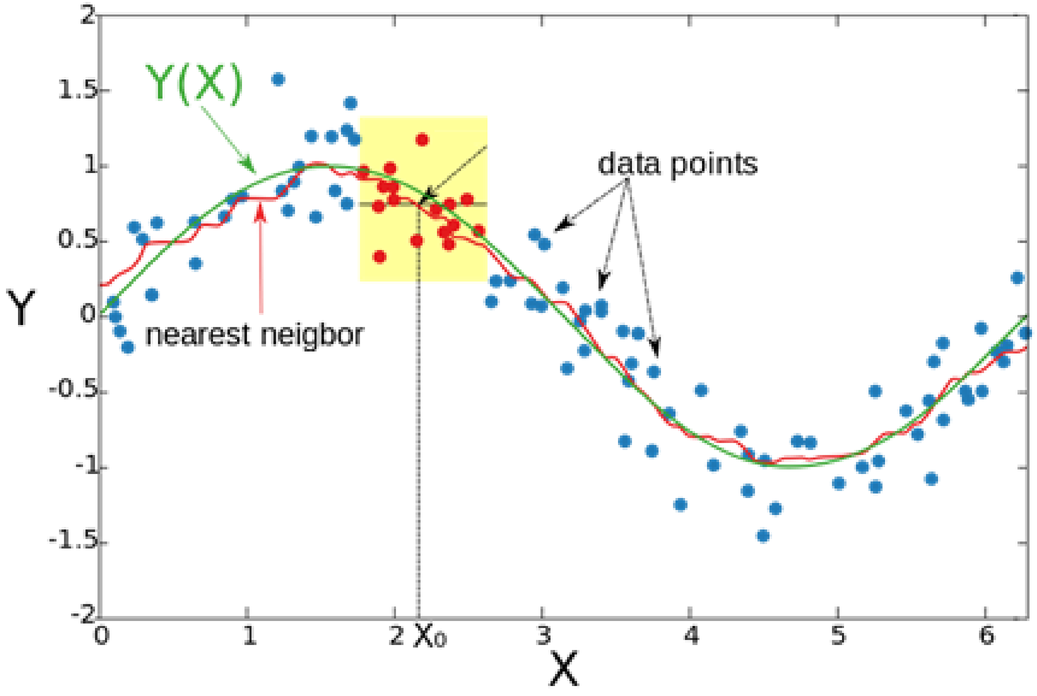




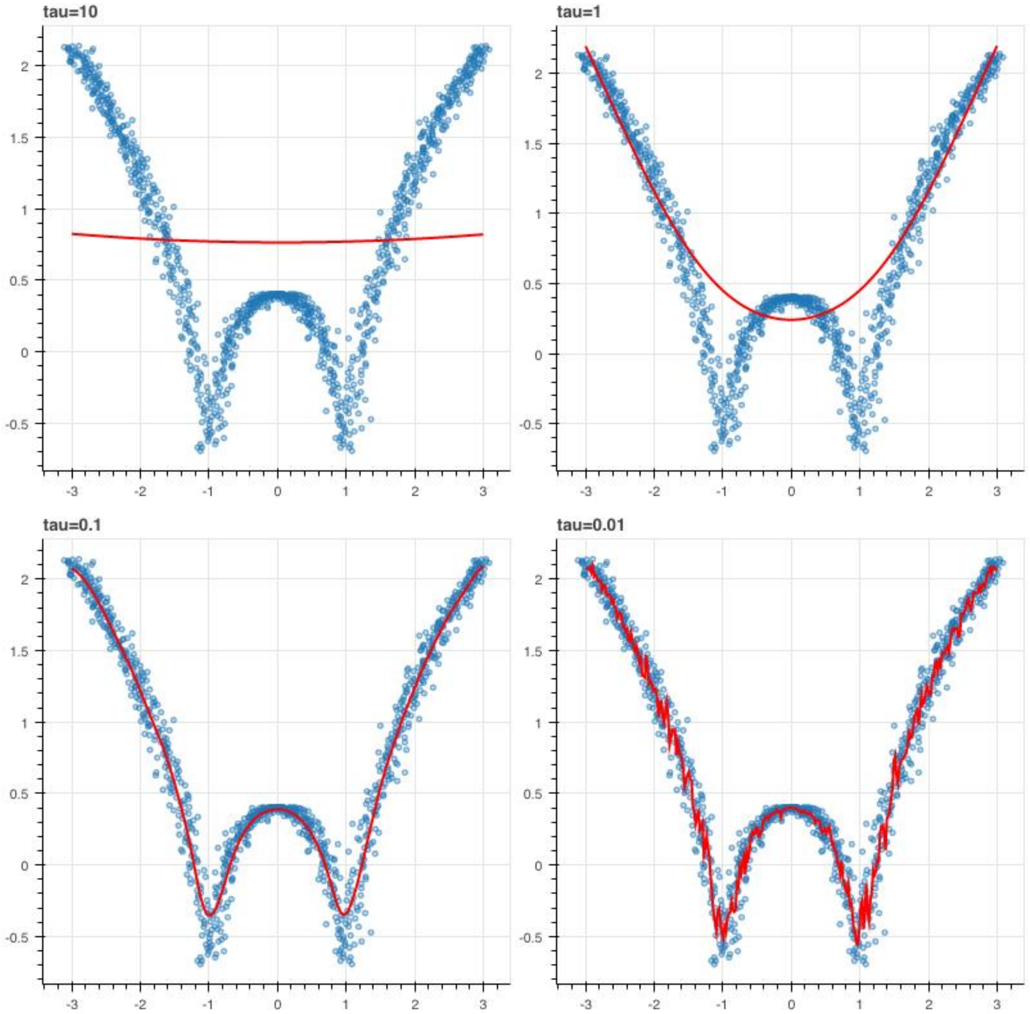
k najbližjih sosedov za regresijo

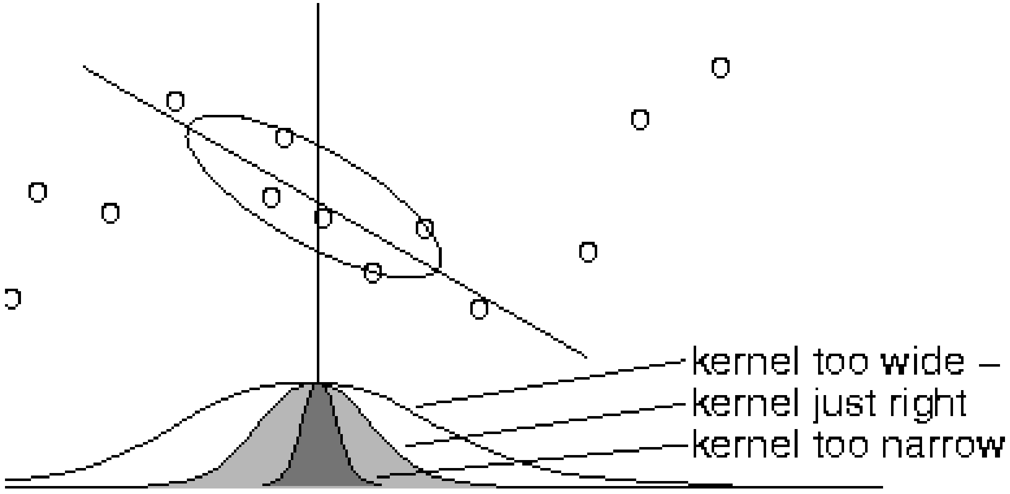
- primer v 2 dimenzijah (atribut X in odvisna spremenljivka Y)

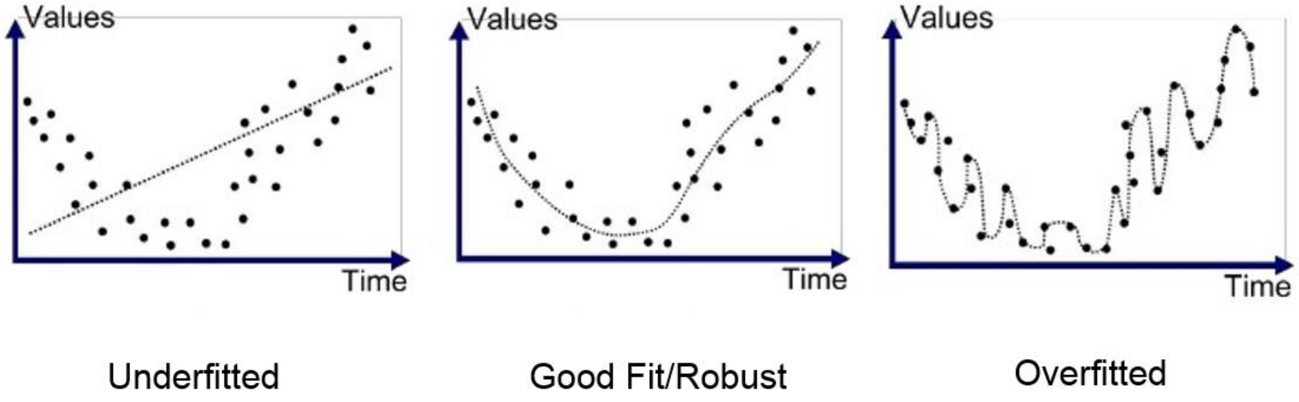




Pomen širine jedra za prileganje podatkom







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

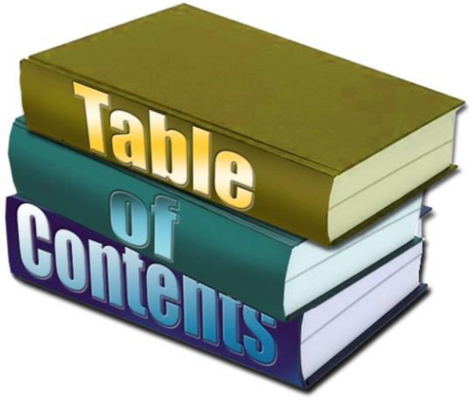
c.) Primeri 1-3 imajo vrednost odvisne spremenljivke enako 10, primera 4-5 pa vrednost odvisne spremenljivke enako 20. Kako bi naslednji napovedni modeli klasificirali primer z zaporedno številko 4 (predpostavi, da ga izločimo iz učne množice in obravnavamo kot testni ali nevideni primer):

- klasifikacijski model 3-NN:

- regresijski model 3-NN:

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

Pregled

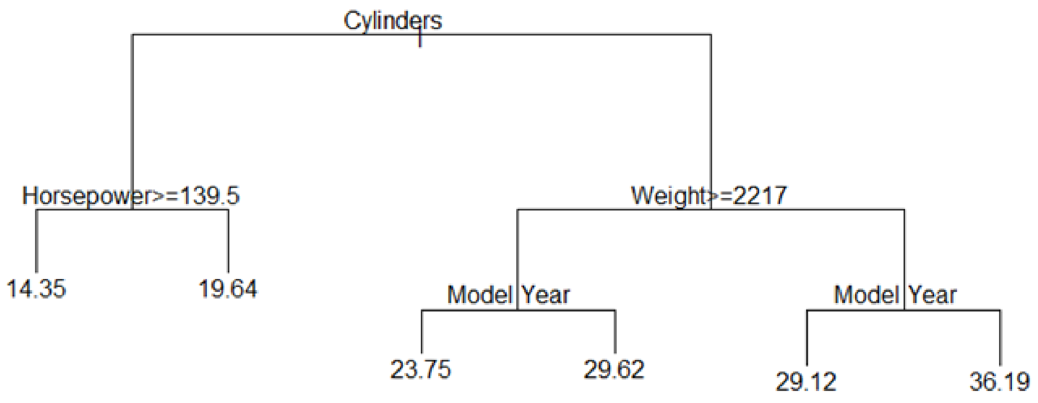


- strojno učenje
  - uvod v strojno učenje
  - učenje odločitvenih dreves
  - učenje dreves iz šumnih podatkov (rezanje dreves)
  - ocenjevanje učenja
  - diskretizacija atributov, obravnava manjkajočih vrednosti
  - naivni Bayesov klasifikator
  - nomogrami
  - k najbližjih sosedov
  - lokalna utežena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

Regresijska drevesa

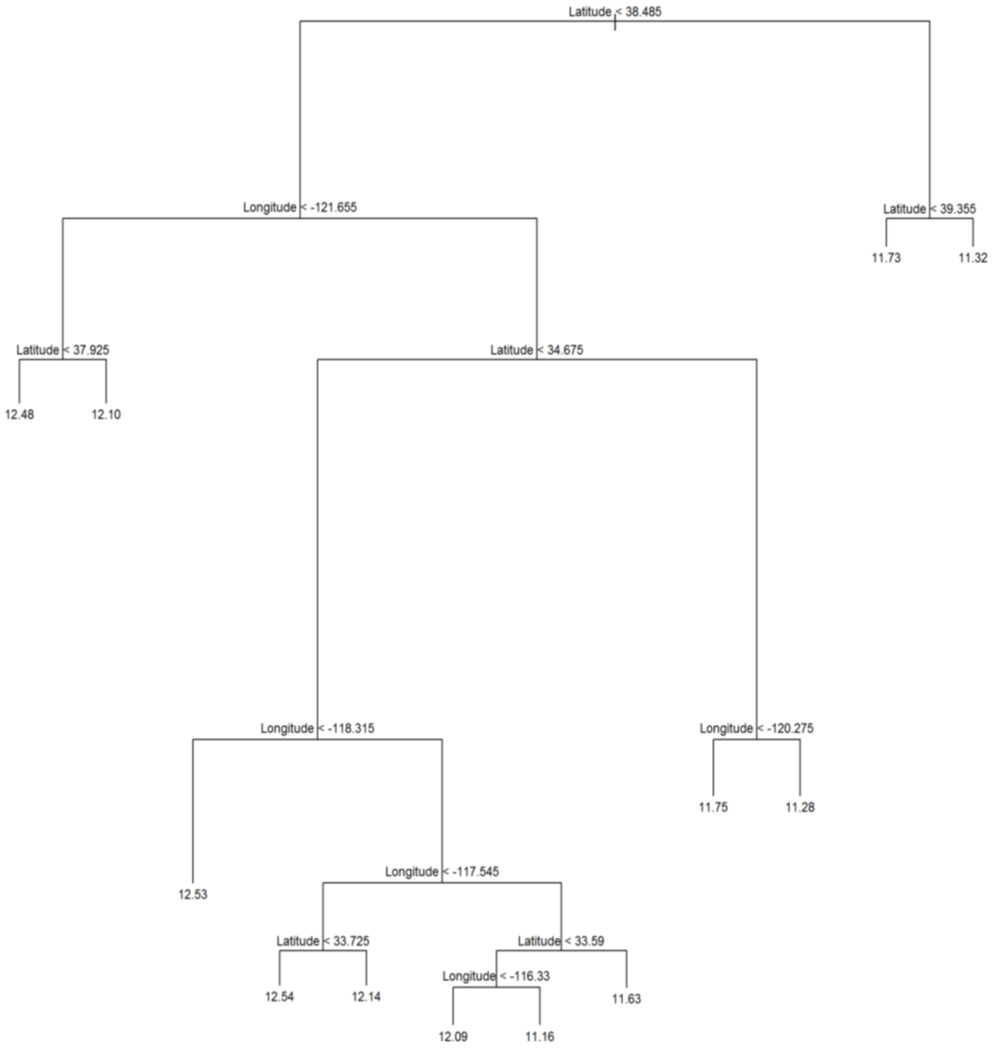
- **zvezna ciljna spremenljivka** – regresijski problem
- regresijska drevesa so podobna odločitvenim drevesom, le za regresijske probleme
- sistemi: CART (Breiman et al. 1984), RETIS (Karalič 1992), M5 (Quinlan 1993), WEKA (Witten and Frank, 2000)

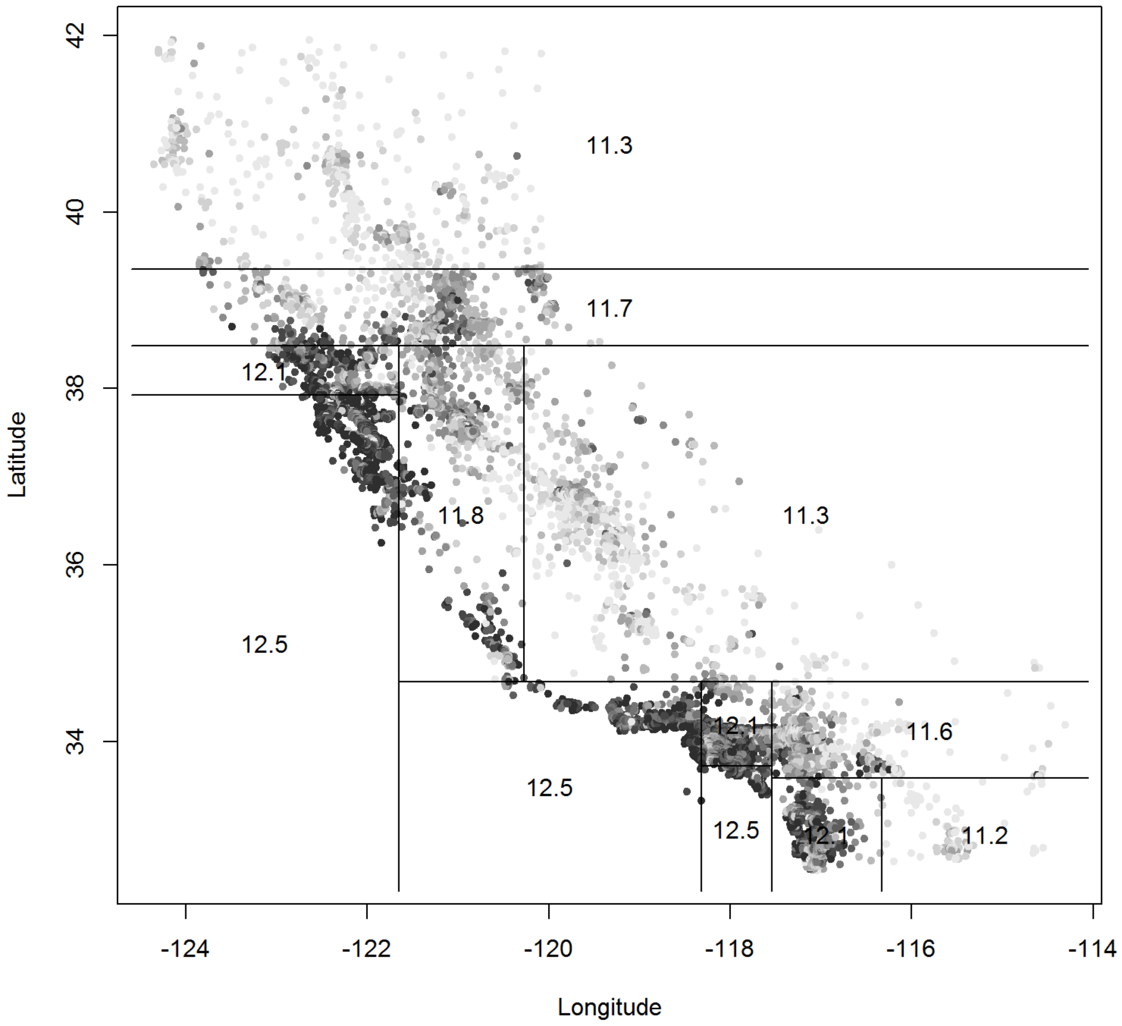
- listi v regresijskem drevesu predstavljajo bodisi:
  - **povprečno vrednost** označb ("razreda") primerov v listu
  - **preprost napovedni model** (npr. linearna regresija) za nove primere



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

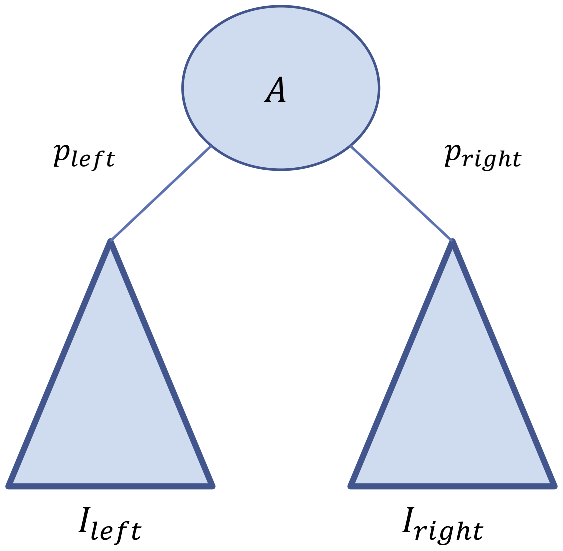
- drugačna mera za merjenje nedoločenosti/nečistoče: srednja kvadratna napaka v vozlišču v:

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

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

- pričakovana rezidualna nečistost

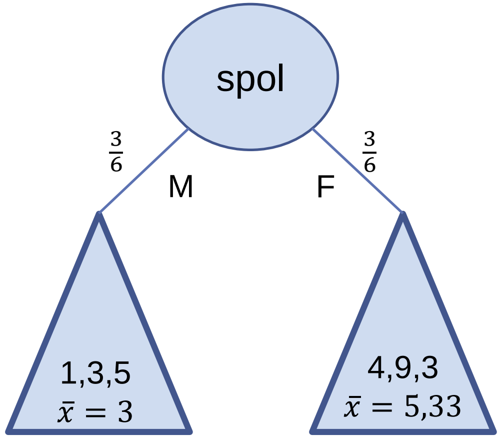
$$I_{res}(A) = p_{left} \cdot I_{left} + p_{right} \cdot I_{right}$$



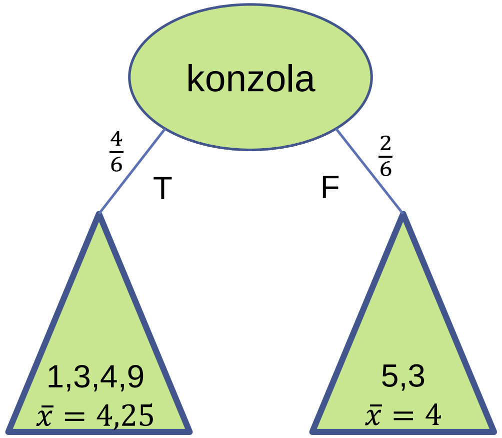
Primer

- 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



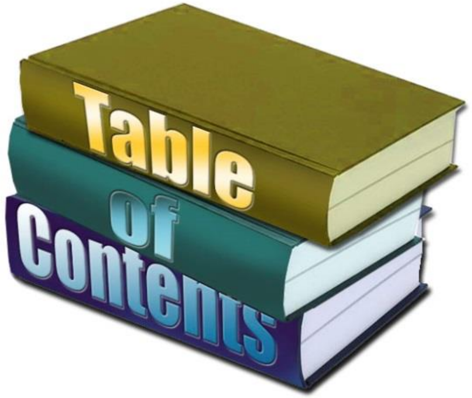
$$I_{res}(A) = p_{left} \cdot I_{left} + p_{right} \cdot I_{right}$$
$$MSE(v) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2$$



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

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

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

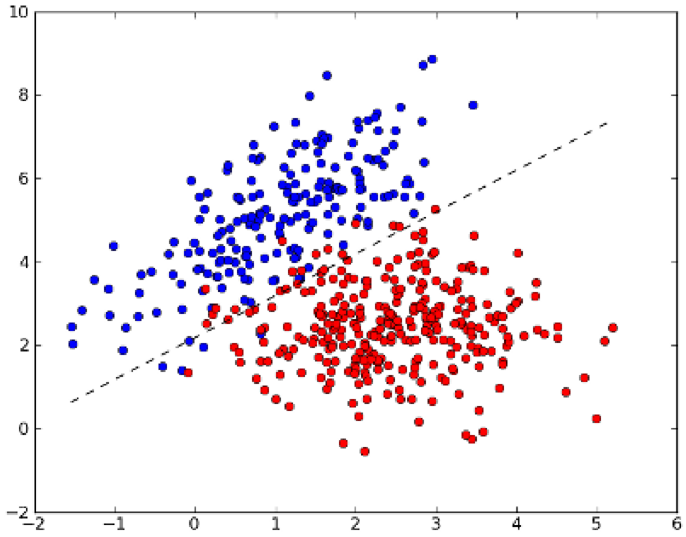
Linearni modeli

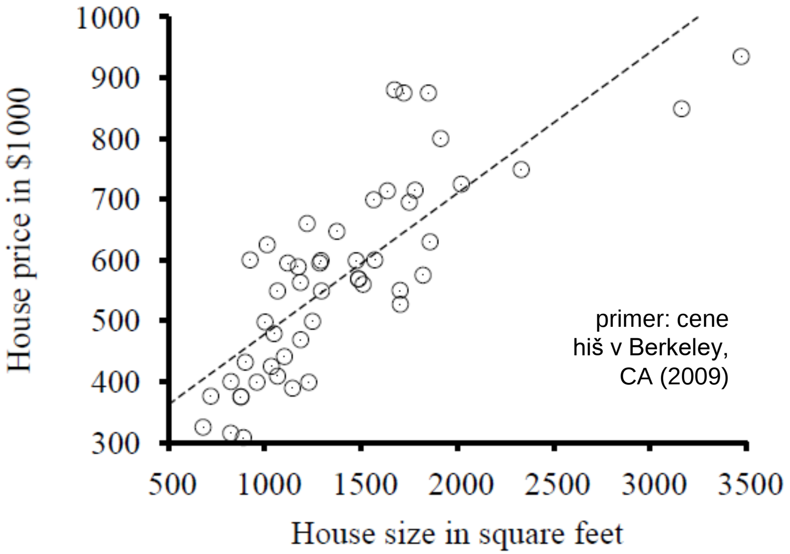
- uporaba pri **klasifikaciji** (kot separator razredov) in **regresiji** (kot prileganje skozi podane točke)
- linearni model z **eno odvisno** spremenljivko (angl. *univariate linear model*):

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

  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






Nenadzorovano učenje,
preiskovanje