

# Naivni Bayesov klasifikator

Jure Žabkar

[jure.zabkar@fri.uni-lj.si](mailto:jure.zabkar@fri.uni-lj.si)

13. 4. 2021



# Vsebina

- Ocenjevanje verjetnosti (m-ocena)
- Naivni Bayesov klasifikator
- Nomogrami
- Ocenjevanje učenja: CV, leave-one-out, train/test set
- Primeri praktičnih nalog

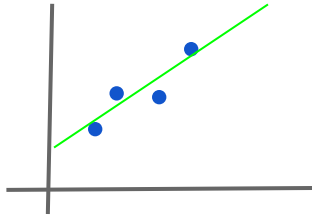
# Strojno učenje

**Nadzorovano**

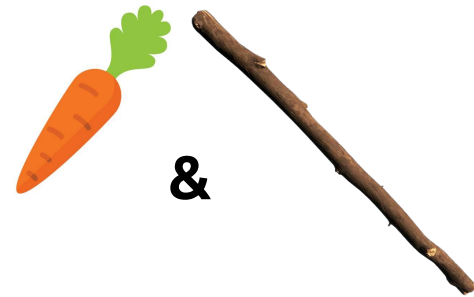
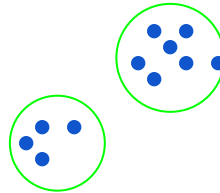
**Nenadzorovano**

**Spodbujevano  
učenje**

Regresija, **Klasifikacija**



Gručenje, povezovalna pravila



# Atributna predstavitev podatkov

survived	status	age	sex
no	third	adult	male
no	crew	adult	male
no	first	adult	male
no	third	adult	male
yes	crew	adult	male
yes	third	adult	male
no	second	adult	male
no	crew	adult	male
no	first	adult	male
no	first	adult	male
yes	crew	adult	male

# Ocenjevanje verjetnosti

Točnost  $T$  = verjetnost pravilne klasifikacije.

Napaka =  $1 - T$

$N$  ... število vseh primerov,  $n$  ... število uspešnih poskusov

- relativna frekvenca:  $p = n/N$

- m-ocena:  $p = (n + p_a * m) / (N + m)$

ekspert zaupa v  $p_a \Rightarrow$  velik  $m$ , sicer majhen  $m$  (tipično  $m=2$ )

- Laplace:  $p = (n+1)/(N+k)$

# Naivni Bayesov klasifikator



- Verjetnosti klasifikator,
- Naivnost: v predpostavki, da so atributi med seboj pogojno neodvisni,
- Algoritem se v praksi pogosto izkaže kot dober,
- Priljubljen zaradi svoje preprostosti in hitrosti,
- Vizualizacija z nomogrami

# Naivni Bayesov klasifikator



$$P(y|x) = \frac{P(y)P(x|y)}{P(x)},$$

kjer je:

- $P(y|x)$  posteriorna verjetnost razreda  $y$  pri danih vrednostih atributa  $x$ ,
- $P(y)$  apriorna verjetnost razreda  $y$ ,
- $P(x|y)$  verjetje (*angl. likelihood*) oz. pogojna verjetnost  $x$  pri danem  $y$  in
- $P(x)$  apriorna verjetnost vrednosti atributa.

# Naivni Bayesov klasifikator



V primeru več atributov:

$$P(y|x_1, \dots, x_n) = \frac{P(y)P(x_1, \dots, x_n|y)}{P(x_1, \dots, x_n)}.$$

Imenovalec v zgornji formuli ni odvisen od razreda in ga za potrebe klasifikatorja zanemarimo, števec pa razpišemo z verižno uporabo pravila:

$$P(A \wedge B|C) = P(A|C) P(B|C \wedge A).$$



# Naivni Bayesov klasifikator



Pogojno verjetnost v števcu razpišemo takole:

$$\begin{aligned} P(x_1, \dots, x_n | y) &= P(x_1 | y) P(x_2, \dots, x_n | y \wedge x_1) \\ &= P(x_1 | y) P(x_2 | y \wedge x_1) P(x_3, \dots, x_n | y \wedge x_1 \wedge x_2) \\ &= P(x_1 | y) P(x_2 | y \wedge x_1) P(x_3 | y \wedge x_1 \wedge x_2) P(x_4, \dots, x_n | y \wedge x_1 \wedge x_2) \\ &= \vdots \\ &= P(x_1 | y) P(x_2 | y \wedge x_1) \dots P(x_n | y \wedge x_1 \wedge \dots \wedge x_{n-1}) \end{aligned}$$

# Naivni Bayesov klasifikator



Privzemimo, da je vsak atribut  $x_i$  pogojno neodvisen od  $x_j$ , za  $i \neq j$ , torej  $P(x_i|y \wedge x_j) = P(x_i|y)$ . Zdaj lahko poenostavimo zadnjo vrstico zgornje izpeljave:

$$P(x_1, \dots, x_n|y) = P(x_1|y) P(x_2|y) \dots P(x_n|y)$$

# Naivni Bayesov klasifikator



Verjetnost razreda pri danih vrednostih atributov je ob predpostavki pogojne neodvisnosti

$$P(y|x_1, \dots, x_n) = \frac{P(y) \prod_{i=1}^n P(x_i|y)}{P(x_1, \dots, x_n)}.$$

Algoritem nov primer klasificira v razred z največjo verjetnostjo:

$$\operatorname{argmax}_{y \in Y} P(y) \prod_{i=1}^n P(x_i|y),$$

kjer je  $Y$  množica vrednosti razreda  $y$ .

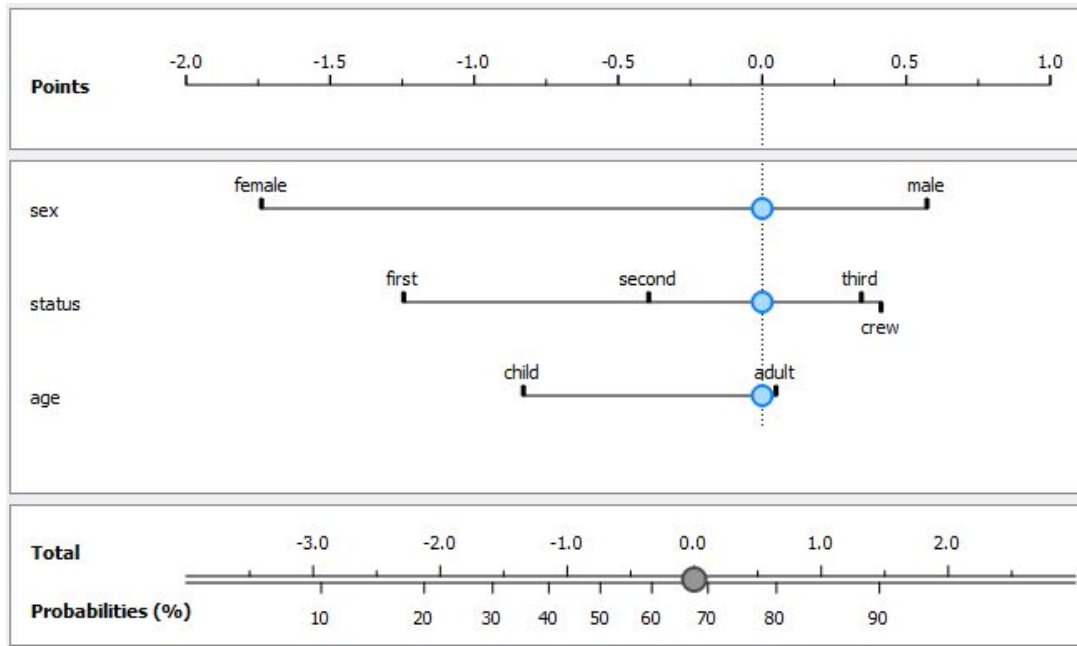
# Naivni Bayesov klasifikator



- Klasifikator: klasificiramo v razred, ki je najbolj verjeten
- Učenje: ocenimo verjetnosti  $P(y)$  in  $P(x_i|y)$  za vse razrede  $y$  in attribute  $x_i$
- Algoritem se v praksi pogosto izkaže kot dober, če so atributi med seboj dovolj pogojno neodvisni

# Nomogrammi

- Vizualizacija modela NB
- Prikazuje vpliv posameznih atributov na verjetnost ciljnega razreda
- Pomembnost atributa za ciljni razred



# Vrste atributov

- Diskretni
  - Nominalni, npr. (dež: da / ne), (spol: M / Ž)
  - Ordinalni, npr. (tlak: nizek / srednji / visok)
- Zvezni (numerični)
  - Poljubna številska vrednost

# Obravnava zveznih atributov v klasifikaciji

Običajno naredimo diskretizacijo v dva ali več diskretnih intervalov.

Načini diskretizacije:

- Intervali z enako frekvenco primerov (equal frequency)
- Enako široki intervali (equal width)
- Intervali, ki maksimizirajo informacijski prispevek
- Ročno (domenski ekspert postavi meje intervalov)

# Diskretizacija z maksimizacijo inf. prispevka

- Mesto vejitve je številka in možnosti za njeno vrednost je neskončno!
- Vejimo na sredini med dvema vrednostma
- $n-1$  možnosti ( $n$  je število učnih primerov); poskusimo vse!

9 da, 5 ne

Entropija pred vejitvijo:

$$-\frac{9}{14} \log_2\left(\frac{9}{14}\right) - \frac{5}{14} \log_2\left(\frac{5}{14}\right) = 0.94 \text{ bit}$$

**Inf. Gain =  
0.001 bit**

temp.

golf

24	25	28	29	30	31	32	35	40	41	43	45
da	ne	da	da	da	ne	ne da	da da	ne	da	da	ne

4 da, 2 ne

5 da, 3 ne

Entropija po vejitvi:

$$\frac{5}{14} \left( -\frac{4}{5} \log_2\left(\frac{4}{5}\right) - \frac{1}{5} \log_2\left(\frac{1}{5}\right) \right) + \frac{9}{14} \left( -\frac{5}{9} \log_2\left(\frac{5}{9}\right) - \frac{4}{9} \log_2\left(\frac{4}{9}\right) \right) = 0.939 \text{ bit}$$



# Ocenjevanje učenja

- Učna / testna množica (train / test set)
- Prečno preverjanje (cross-validation)
- Izpusti enega (leave-one-out)

# Prečno preverjanje (cross-validation)

