

13.12. - cvičenia (F1-248)

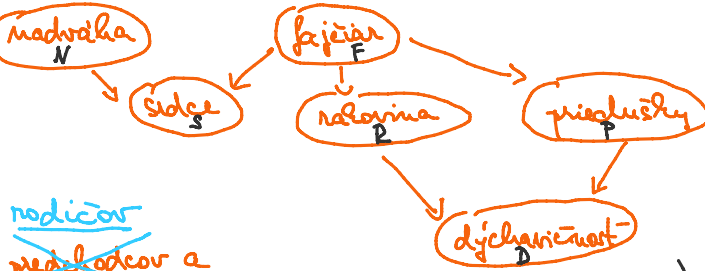
do 8.12. - výber projektu: Announcement do G. Classroom

Minule: B. sieť

1. Ako modelujúci pp.

$$(A \perp \tilde{A}) \mid pa(A)$$

$\tilde{A}$: všetky ~~predchodcov~~ ~~nasledovníkov~~ ~~predchodcov~~ ~~nasledovníkov~~ A



$$\Rightarrow (D \perp N, S, F) \mid R, P$$

$$\Rightarrow P(n, f, s, r, p, d) = P(n) \cdot P(f) \cdot P(s \mid n, f) \cdot P(r \mid f) \cdot P(p \mid f, r) \cdot P(d \mid r, p)$$

$$P(N=n), n \in \{0, 1\}$$

$$\text{všeob.: } P(x_1, \dots, x_p) = \prod_{j=1}^p P(x_j \mid pa(x_j))$$

parametre: (podmiernené) pp. v každom uzle

$$P(N=1) = 0,2$$

$$P(N=0) = 0,8$$

$$P(F=1) = 0,2$$

$$P(P=1 \mid F=1) = 0,15 \Rightarrow P(P=0 \mid F=1) = 0,85$$

$$P(P=1 \mid F=0) = 0,05 \Rightarrow P(P=0 \mid F=0) = 0,95$$

$$P(S=1 \mid N=1, F=1) = 0,2$$

$$P(S=0 \mid N=1, F=1) = 0,8$$

$$P(R=1 \mid F=1) = \dots$$

$$P(R=1 \mid F=0) = \dots$$

$$P(D=1 \mid R=1, P=1) = 0,7$$

$$P(\dots)$$

$$P(D=1 \mid R=0, P=0) = 0,04$$

→ spolu: N... 1
F... 1
S... 4
R... 2
P... 2
D... 4

⇒ 14 parametrov (podmiernené) pp.
potrebujeme táto B. tiež

$$\text{vs. bez siete: } 2^6 - 1 = 64 - 1 = 63$$

2. Ako odhadovať parametre

nut. max. virohodnosti: dáta = X

$$X = \begin{bmatrix} x_{11} & \dots & x_{1p} \\ \vdots & & \vdots \\ x_{n1} & \dots & x_{np} \end{bmatrix}$$

→ uvažuje o n ľuďoch

$$\text{napr. } x_1^T = (1, 0, 1, 1, 1, 1)$$

θ : všetky (podmiernené) pp.

$$\Rightarrow L(\theta) = \prod_{i=1}^n P(X_1=x_{i1}, X_2=x_{i2}, \dots, X_p=x_{ip} \mid \theta) \rightarrow \max_{\theta}$$

v. usky - podmienené pp.

$$\Rightarrow L(\theta) = \prod_{i=1}^n P(X_1=x_{i1}, X_2=x_{i2}, \dots, X_p=x_{ip} | \theta) \rightarrow \max_{\theta}$$

pre x_1 a n : $P(N=1, F=0, \dots, D=1)$: ako modeluje sieť pp, či človek má nadvačku, nie je fajčiar, ... a má dykhanie

Riešenie: $\hat{\theta}$: (podmienená) pp. = pozorované frekvencie

napr. $P(R=1 | F=1) = \frac{\#(R_i=1, F_i=1)}{\#(F_i=1)}$ $\rightarrow$ aké % fajčiarov má nádor

• problém: ak niekto, frekvencia je 0 $\Rightarrow$ pridáme umelí pozorovania $\Rightarrow$ frekvencia = 5

3. Výpočty pp. z B. siete:

• $P(P=1 | F=1) = 0,15$: triválne - priamo parameter siete

• aj ľub. iná (podmienená) pp.

$P(F=1 | D=1)$ PP, čo je fajčiar, keď má dykhanie

$$P(F=1 | D=1) = \frac{P(F=1, D=1)}{P(D=1)} = \frac{\sum_{N, S, R} P(N=n, F=1, S=s, R=r)}{\sum_{N, S, R} P(N=n, F=f)}$$

efektívne

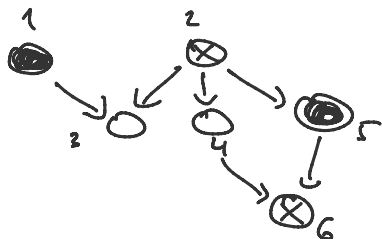
$P=P, D=1$

Vo výp

$S=s, R=r, P=p, D=1$

tiež už vieme vypočítať

NFSRPD



sumy: marginalizácia
 $\rightarrow$ v B. sieťach 3 efektívne algoritmy

vo všeob. : 2 typy uzlov :
● : pozorované
○ : číselná zmietať i.e. pp.
⊗ : ostatné

$$P(\odot | \bullet) = \frac{P(\odot, \bullet)}{P(\bullet)} = \frac{\sum_{\otimes} P(\odot, \bullet, \otimes)}{\sum_{\odot, \otimes} P(\odot, \bullet, \otimes)}$$

• ľahko sa vysporiada s dykhaním prem. (⊗) - napr. dykhanie vyšetrenia
• nie nutne predikujeme jeden Y, ale "niečo, čo nás zaujíma" (bez učiteľa / s učiteľom)

výhody: - reprezentácia pp. rozdelení pomocou menšieho # parametrov $\rightarrow$ vyžaduje menší # dát
- dobre interpretovateľné

nevýhody: - musíme predpokladať / poznať vzťahy (trvať siete)

$\hookrightarrow$ ale aj trvať viem učiť + dáť - ako výber modelu a lin. reg.

$$\left[\begin{array}{l} \text{vo všeob. zápise} \\ P(X_2=a | X_1=b) = \frac{P(X_1=a, X_2=b)}{P(X_1=b)} = \frac{\sum_{x_3, \dots, x_p} P(X_1=b, X_2=a, X_3=x_3, \dots, X_p=x_p)}{\sum_{x_3, \dots, x_p} P(X_1=b, X_2=x_2, \dots, X_p=x_p)} \end{array} \right]$$

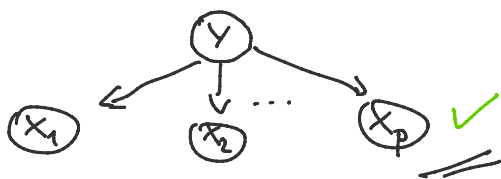
Naivný bayesovský klasifikátor (Naive Bayes classifier)

, x 1

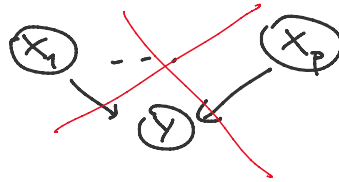
Naivný bayesovský klasifikátor (Naive Bayes classifier)

data $(x_1, y_1), \dots, (x_n, y_n)$ $x_i \in \mathbb{R}^p$ $y_i \in \{1, \dots, C\}$: realizácie $\begin{pmatrix} x_1 \\ \vdots \\ x_p \\ y \end{pmatrix}$

model: (i)



(ii)



napr.: $Y = \text{choroba (áno/nie)}$

$X_j = \text{testy}$

Príes: 1.) "Logický": choroba spôsobuje výsledky testov, nie naopak

2.) Technický: # parametrov (jednoduchosť - všetko binárne)

→ (i) $P(Y=1)$

$$\left. \begin{array}{l} P(X_1=1|Y=1) : 2 \text{ pre } X_1 \\ P(X_1=1|Y=0) : 2 \text{ pre } X_2 \end{array} \right\} 1 + 2p$$

$$\rightarrow (ii) \left. \begin{array}{l} P(X_1=1) \\ \vdots \\ P(X_p=1) \end{array} \right\} P \quad \left. \begin{array}{l} P(Y=1|X_1=1, \dots, X_p=1) \\ \vdots \\ P(Y=1|X_1=0, \dots, X_p=0) \end{array} \right\} 2^p \Rightarrow \text{spolu } 2^p + 1 \text{ param.}$$

3.) Technický: typický - nehoršie vchody pozorujeme a na skôr sa pýtame

- naopak by to bolo triviálne

→ (ii):

$$P(Y=0|X_1=1, X_2=0) = \frac{\#(i: Y_i=0, X_{1i}=1, X_{2i}=0)}{\#(i: X_{1i}=1, X_{2i}=0)}$$

→ Evidencia pridáva hodnotu B. siete

• model: $(X_1 \perp X_2, \dots, X_p) | Y$

$(X_j \perp \text{ostatné } X) | Y$

→ $X_1, \dots, X_p$ sú nez. keď poznáme hodnotu Y

→ $X_1, \dots, X_p$: nez. v každej triede Y príslušnosti

• celá pp.: $P(y, x_1, \dots, x_p) = P(y) \cdot P(x_1|y) \cdot \dots \cdot P(x_p|y)$

• nez. $\Rightarrow P(x_1, \dots, x_p | y) = P(x_1|y) \cdot \dots \cdot P(x_p|y)$
pre spojitú: $f(x_1, \dots, x_p | y) = f_1(x_1|y) \cdot \dots \cdot f_p(x_p|y)$

→ namiesto komplikovanej p-rozmernej pp. (hustoty)

v každej triede $Y=y$ separátne modelujeme pp. (hustoty) alebo sečím marginálnych

→ namiesto komplikovanej p-rozmernej pp. (kustoty)

v každej triede $Y=y$ separátu modelujeme pp. (kustotu) ako súčin marginálnych

odhadly: $\frac{P(x_j | y)}{P(y)}$: frekvencie

otáčky: $P(y | x_1, \dots, x_p) = \frac{P(y, \dots, x_p)}{P(x_1, \dots, x_p)} = \frac{P(y, x_1, \dots, x_p)}{\sum_{l=1}^L P(Y=l, X_1=x_1, \dots, X_p=x_p)}$

to už model
viac vypočítať

ale tu pp. jednotlivých tried, keď poznáme charakteristiky

- podmienená, nez. výskyt x-ov - značne obmedzujúca
- ale napriek tomu často celkom dobre predikuje

Python: Spam? podľa subjektu mailu

→ $P(\text{spam} = \text{True} | \text{text})$

"spammiest words" : s najvyššou $P(\text{spam} = \text{True} | \text{slovo} = 1)$

↑
napr. "assistance" = 1, t.j. to slovo
je v subjekte