11, Bayesovská statistika

- Základní pojmy z teorie pravděpodobnosti
  - $P(A)$ pravděpodobnost nastání jevu
  - $P(A|B) = \frac{P(A,B)}{P(B)}$ (1) podmíněná pravděpodobnost nastání jevu A, nastal-li jev B
  - $P(A,B) = P(A|B) \cdot P(B)$ pravděpodobnost nastání jevu A a B zároveň (tzv. řetězové pravidlo)
  - $P(A|B_1,\dots,B_N) = \frac{P(A,B_1,\dots,B_N)}{P(B_1,\dots,B_N)}$ (2) podmíněná pravděpodobnost nastání jevu A, nastali-li jevy B_1 až B_N
  - $P(A,B_1,\dots,B_N) = P(B_1,\dots,B_N,A)$ (5) jevy můžeme přehodit v pořadí
  - řetězové pravidlo (3)
    $P(B_1,\dots,B_N,A) = P(B_1|B_2,\dots,B_N,A) \cdot P(B_2|B_3,\dots,B_N,A) \dots P(B_N|A) \cdot P(A)$
  - pokud jsou jevy A, B na sobě nezávislé (4):
    1) $P(A|B) = P(A)$
    2) $P(A,B) = P(A) \cdot P(B)$
    3) $P(A,B|C) = P(A|C) \cdot P(B|C)$

- Bayesova věta
  - věta o tom, jak podmíněná pravděpodobnost souvisí s opačnou podmíněnou pravděpodobností pro dva jevy

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}$$

- Naivní Bayesův klasifikátor
  - model pro klasifikaci dat s atributy x_i do tříd y s premisou nezávislosti atributů x_i na sobě (proto naivní)

$X = [x_1, x_2, \dots, x_N]$ vektor atributů y = třída z množiny $Y = [y_1, \dots, y_k]$

$$P(y|X) \overset{(1)}{=} \frac{P(y,X)}{P(X)} \overset{(5)}{=} \frac{P(X,y)}{P(X)} \overset{(2)}{=} \frac{P(x_1,\dots,x_2,y)}{P(x_1,\dots,x_N)} \overset{(3)}{=}$$

$$= \frac{P(x_1|x_2,\dots,x_N,y) \cdot P(x_2|x_3,\dots x_N,y) \cdot \dots \cdot P(x_N|y) \cdot P(y)}{P(x_1|x_2,\dots,x_N) \cdot P(x_2|x_3,\dots,x_N) \cdot \dots \cdot P(x_{N-1}|x_N) \cdot P(x_N)} \overset{(4)}{=}$$

$$= \frac{P(x_1|y) \cdot P(x_2|y) \cdot \dots \cdot P(x_N|y) \cdot P(y)}{P(x_1) \cdot P(x_2) \cdot \dots \cdot P(x_{N-1}) \cdot P(x_N)} = \frac{\prod_{i=1}^{N} P(x_i|y) \cdot P(y)}{\prod_{i=1}^{N} P(x_i)}$$

- hledáme $\hat{y} = y_k = \operatorname{argmax} P(y_k) \cdot \prod_{i=1}^{N} P(x_i|y_k)$ $\forall k$ nezáleží na něm

- Apriorerní a aposteriorní pravděpodobnost

aposteriorní P

$$P(y|x_1,\dots,x_N) = \frac{\prod_{i=1}^{N} P(x_i|y)\cdot P(y)}{\prod_{i=1}^{N} P(x_i)}$$

věrohodnost — apriorní P — evidence

X = data, realita, měření
y = hypotéza

- aposteriorní P: jak pravděpodobná je hypotéza (klasifikace) za daných měření?
- apriorní P: jak pravděpodobná je hypotéza bez ohledu na měření?
- věrohodnost: jak pravděpodobné je naměřit daná data, je-li hypotéza pravdivá?
- evidence: jak pravděpodobné je naměřit daná data za platnosti všech možných hypotéz?

- Typy klasifikátorů Naivního Bayese podle statistického rozdělení X

1) Bernouliho NBK: $P(x_1,\dots,x_N|y_k) = \prod_{i=1}^{N} p_{ki}^{x_i}(1-p_{ki})^{1-x_i}$

p_{ki} = pravděpodobnost, že třída y_k generuje člen x_i

x_i = Booloovské hodnoty (obsahuje / neobsahuje)

2) Multinomiální NBK: $x_i \in \mathbb{N}$ (počty nálezů atributu)

3) Gaussovský NBK: $p(x_1,\dots,x_N|y) = \frac{1}{\sqrt{2\pi\sigma_y^2}}\cdot\exp\left(-\frac{(x_i-\mu_y)^2}{2\sigma_y^2}\right)$

$x_i \in \mathbb{R}$ a $N(\mu_y, \sigma_y)$ (Gaussovské rozdělení)

- Využití NBK

id	hypotéka	rodinný stav	příjem	match
1	ano	svobodný	125k	ne
2	ne	ženatý	100k	ne
3	ne	svobodný	70k	ne
4	ano	ženatý	120k	ne
5	ne	rozvedený	95k	ano
6	ne	ženatý	60k	ne
7	ano	rozvedený	220k	ne
8	ne	svobodný	85k	ano
9	ne	ženatý	75k	ne
10	ne	svobodný	90k	ano

$P(y=ne) = ?$

$\mu = (125+100+70+120+60+220+75)/7 = 110$

$\sigma^2 = \frac{(125-110)^2+(100-110)^2+\dots+(75-110)^2}{7-1} = 2975$

$P(p=120k|y=ne) = \frac{1}{\sqrt{2\pi\cdot 2975}}\cdot\exp\left(-\frac{(120-110)^2}{2\cdot 2975}\right) = 0.0072$

$P(h=ne|y=ne) = 4/7$

$P(rs=\check{z}|y=ne) = 4/7$

$P(X, y=ne) = P(p=120k|y=ne)\cdot P(h=ne|y=ne)\cdot P(rs=\check{z}|y=ne) = 0.0072\cdot 4/7\cdot 4/7 = \underline{0.0023}$ ✓

Jak dopadne $X = [h=ne, rs=ženatý, p=120k]$?

$P(X, y=ano) = 3/3\cdot 0\cdot 1,2E{-}9 = 0$

(ještě nutné vynásobit $P(y_k)$)