

# Rozpoznávanie obrazcov - 9. cvičenie

## Naivný Bayesov klasifikátor

Viktor Kocur

[viktor.kocur@fmph.uniba.sk](mailto:viktor.kocur@fmph.uniba.sk)

DAI FMFI UK

21.4.2020

# Bayesovo pravidlo

## Bayesovo pravidlo

Budeme opäť používať Bayesovo pravidlo:

$$P(\omega_i|\vec{x}) = \frac{P(\vec{x}|\omega_i)P(\omega_i)}{P(\vec{x})} \quad (1)$$

## Naivita

Náš klasifikátor je naivný a predpokladá, že príznaky sú nezávislé:

$$P(\vec{x}|\omega_i) = \prod_k P(x_k|\omega_i) \quad (2)$$

# Klasifikátor

## Klasifikácia

Klasifikujeme pomocou nájdania triedy s najväčšou pravdepodobnosťou:

$$pred_i = \arg \max_i \left( \frac{P(\vec{x}|\omega_i)P(\omega_i)}{P(\vec{x})} \right) \quad (3)$$

$$= \arg \max_i (P(\vec{x}|\omega_i)P(\omega_i)) \quad (4)$$

$$= \arg \max_i \left( P(\omega_i) \prod_k P(x_k|\omega_i) \right) \quad (5)$$

# Klasifikátor

## Výpočet hodnôt

Budeme predpokladať, že máme kategorické príznaky. Teda pre každé  $k$  môže  $x_k$  nadobúdať iba konečne mnoho diskrétnych hodnôt. Označíme celkový počet prvkov trénovacej množiny ako  $N$ . Počet prvkov, ktoré patria do triedy  $\omega_i$  ako  $N_i$ . Počet prvkov, ktoré patria do  $\omega_i$  a pre  $k$ -tý príznak majú hodnotu  $v$  ako  $N_{i,k,v}$ . Potom môžeme definovať:

$$P(\omega_i) = \frac{N_i}{N} \quad (6)$$

$$P(x_k = v | \omega_i) = \frac{N_{i,k,v}}{N_i} \quad (7)$$

## Klasifikátor

age	income	student	credit rating	buys computer
<=30	high	no	fair	no
<=30	high	no	excellent	no
31...40	high	no	fair	yes
>40	medium	no	fair	yes
>40	low	yes	fair	yes
>40	low	yes	excellent	no
31...40	low	yes	excellent	yes
<=30	medium	no	fair	no
<=30	low	yes	fair	yes
>40	medium	yes	fair	yes
<=30	medium	yes	excellent	yes
31...40	medium	no	excellent	yes
31...40	high	yes	fair	yes
>40	medium	no	excellent	no

## Úloha

Spočítajte do ktorej kategórie bude patriť zákazník s náhodnými prediktormi.

age <= 30,  
income = medium,  
student = no,  
credit rating = fair

# Klasifikátor

## Nekategorické dáta

V prípade, že niektorý príznak je numerický, tak nemôžeme aplikovať výpočet z predchádzajúceho slidu. Preto budeme pravdepodobnosť  $P(x_k|\omega_i)$  odhadovať nejakou distribučnou funkciou.

## Parametrické metódy

Pri parametrických metódach odhadneme parametre nejakého dopredu určeného rozdelenia.

## Neparametrické metódy

Pri neparametrických metódach pravdepodobnosť vypočítame na základe bodov z trénovacej množiny v okolí bodu o ktorý sa zaujíname.

# Matlab

## fitcnb

`Mdl = fitcnb(T,'nazov_pola')` - vráti naivný Bayesov klasifikátor pre tabuľku `T` pre klasifikačný cieľ pre stĺpec `nazov_pola`.

## Malab - Table dátový typ

Pre prácu s tabuľkami si pozrite:

<https://www.mathworks.com/help/matlab/tables.html>

A dôležitá je aj časť o prístup k dátam:

[https://www.mathworks.com/help/matlab/matlab\\_prog/access-data-in-a-table.html](https://www.mathworks.com/help/matlab/matlab_prog/access-data-in-a-table.html)

# Naivný Bayes na tabuľkových dátach

## Na dátach

```
load census1994  
Mdl = fitcnb(adulddata, 'salary');
```

## Úloha

Zistite presnosť klasifikátora tak, že ho spustíte (Mdl.predict) na tabuľku adulttest a porovnáte výsledok.



# Matlab

fitcnb

$Mdl = \text{fitcnb}(X,y)$  - vráti naivný Bayesov klasifikátor

Úloha

Otestujte naivný Bayesov klasifikátor na fisheriris dátach.

Úloha

Zobrazte si klasifikátor na dátach zo 6. cvičenia pomocou úpravy skriptu showSVM z toho istého cvičenia.