

# Umělá inteligence a rozpoznávání (KIV/UIR)

## 7. Příznakové metody rozpoznávání - doplňující informace k přednášce

Ing. Pavel Král, Ph.D.

Katedra informatiky a výpočetní techniky  
Západočeská Univerzita

26. března 2014

Umělá  
inteligence a  
rozpoznávání  
(KIV/UIR)

Ing. Pavel  
Král, Ph.D.

Výběr  
příznaků

Další  
klasifikační  
algoritmy

## 1 Výběr příznaků

## 2 Další klasifikační algoritmy

# Výběr příznaků

(Feature selection)

Umělá  
inteligence a  
rozpoznávání  
(KIV/UIR)

Ing. Pavel  
Kráal, Ph.D.

Výběr  
příznaků

Další  
klasifikační  
algoritmy

## Motivace

- Cena měření, výpočtu a použití (pro klasifikaci) všech příznaků
- Nevhodně zvolené příznaky (tj. i vyšší počet) → snížení ACC

## Poznámky

- Problém **přesnost** vs. **obecnost**
- Vhodná znalost  $P(x|c)$

## Př:

- Klasifikace dokumentů (vhodné vs. nevhodné příznaky)

# Základní metody pro výběr příznaků

Umělá  
inteligence a  
rozpoznávání  
(KIV/UIR)

Ing. Pavel  
Král, Ph.D.

Výběr  
příznaků

Další  
klasifikační  
algoritmy

- Náhodný výběr
- Dokumentová frekvence
- TF-IDF (Term Frequency-Inverse Document Frequency)
- Vzájemná informace (Mutual Information, MI)

# Náhodný výběr

Umělá  
inteligence a  
rozpoznávání  
(KIV/UIR)

Ing. Pavel  
Král, Ph.D.

Výběr  
příznaků

Další  
klasifikační  
algoritmy

- Shora dolů/zdola nahoru
- Postupné ubírání/přidávání příznaků
- Ověření pomocí klasifikátoru

# Dokumentová frekvence

## Document Frequency (DF)

Umělá  
inteligence a  
rozpoznávání  
(KIV/UIR)

Ing. Pavel  
Kráč, Ph.D.

Výběr  
příznaků

Další  
klasifikační  
algoritmy

- Vyjádření, jak často se slovo (term)  $t_i$  vyskytuje v daném dokumentu  $d_j$
- Normalizace délkou dokumentu

$$TF_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}}$$

- $n_{i,j}$  ... počet výskytů slova  $t_i$  v dokumentu  $d_j$
- Dělitel ... součet počtu výskytů všech slov v dokumentu  $d_j$  (délka)

Př:

# TF-IDF

Term Frequency-Inverse Document Frequency

Umělá  
inteligence a  
rozpoznávání  
(KIV/UIR)

Ing. Pavel  
Král, Ph.D.

Výběr  
příznaků

Další  
klasifikační  
algoritmy

$$TF - IDF = TF \cdot IDF$$

- *IDF* ... reprezentace “důležitosti” slova
- Častý výskyt slova  $\rightarrow$  malá důležitost
- Příklad: sloveso “je” - velký výskyt ve všech dokumentech  $\rightarrow$  nedůležitost pro klasifikaci

$$IDF_i = \log \frac{|D|}{|\{j : t_i \in d_j\}|}$$

- $|D|$  ... počet všech dokumentů
- $|\{j : t_i \in d_j\}|$  ... počet dokumentů, kde se vyskytuje slovo  $t_i$ .

Příklad:

# Vzájemná informace (Mutual Information, MI)

Umělá  
inteligence a  
rozpoznávání  
(KIV/UIR)

Ing. Pavel  
Kráal, Ph.D.

Výběr  
příznaků

Další  
klasifikační  
algoritmy

- Dány dvě diskrétní náhodné proměnné  $X$  a  $Y$
- MI = informace, kterou sdílí  $X$  a  $Y$ 
  - Znalost prom.  $X \rightarrow ?$  snížení nejistoty o  $Y$  (a naopak)
  - $X$  a  $Y$  nezávislé proměnné  $\rightarrow MI(X, Y) = 0$

$$MI(X; Y) = \sum_{y \in Y} \sum_{x \in X} p(x, y) \log \left( \frac{p(x, y)}{p(x)p(y)} \right)$$

- $p(x)$  a  $p(y)$  ... hustoty pravděpodobností  $X$  a  $Y$
- $p(x, y)$  ... sdružená hustota pravděpodobnosti  $X$  a  $Y$



# Další klasifikační metody

Umělá  
inteligence a  
rozpoznávání  
(KIV/UIR)

Ing. Pavel  
Král, Ph.D.

Výběr  
příznaků

Další  
klasifikační  
algoritmy

- k-nejbližších sousedů (k-NN)
- Klasifikační a regresní stromy (CART)
- Maximální entropie (Maximum Entropy)

# k-nejbližších sousedů (k-NN)

k-Nearest Neighbors

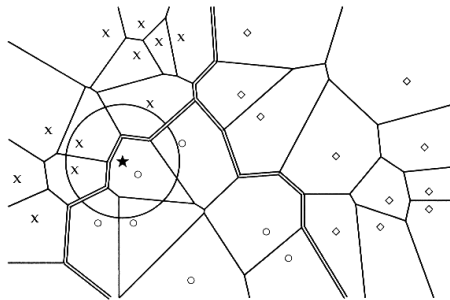
Umělá  
inteligence a  
roznávání  
(KIV/UIR)

Ing. Pavel  
Král, Ph.D.

Výběr  
příznaků

Další  
klasifikační  
algoritmy

- Lokální hranice pro klasifikaci
- 1-NN - třída určena dle nejbližšího souseda
- k-NN - třída s max. počtem nejbližších sousedů (z celk. počtu  $k$ )
- Použití různých metrik



Obrázek: Ukázka [1]

# Klasifikační a regresní stromy (CART)

## Classification and Regression Trees

Umělá  
inteligence a  
rozpoznávání  
(KIV/UIR)

Ing. Pavel  
Král, Ph.D.

Výběr  
příznaků

Další  
klasifikační  
algoritmy

- Popis vzájemných vztahů mezi pozorovanými veličinami pomocí stromu

### Složení

- Kořen, uzly (větvení), listy (terminální uzly), hrany

### Dělení

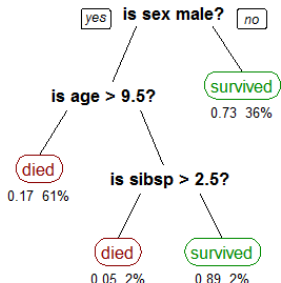
- Binární
- Ternární

### Trénování

- Nastavení kritérií (vah) v uzlech stromu

### Testování

- Průchod stromem dle nastavených kritérií (vah)





Christopher D Manning, Prabhakar Raghavan, and Hinrich  
Schütze,

*Introduction to information retrieval*, vol. 1,  
Cambridge university press Cambridge, 2008.