Umělá inteligence a rozpoznávání (KIV/UIR)

Příznakové metody rozpoznávání

Ing. Pavel Král, Ph.D.

Katedra informatiky a výpočetní techniky
Západočeská Univerzita

22. března 2016

Obsah

Výběr příznaků
(Feature selection)

Motivace

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

Poznámky

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

Př:

- Klasifikace dokumentů (vhodné vs. nevhodné příznaky)

Základní metody pro výběr příznaků

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

Náhodný výběr

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

Dokumentová frekvence
Document Frequency (DF)

- 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

$TF-IDF = TF.IFD$

- IDF ... reprezentace "důležitosti" slova
- Častý výskyt slova $\rightarrow$ malá důležitost
- Př: 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.

Př:

Vzájemná informace (Mutual Information, MI)

- 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 $\rightarrow MI(X, Y) = 0$

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

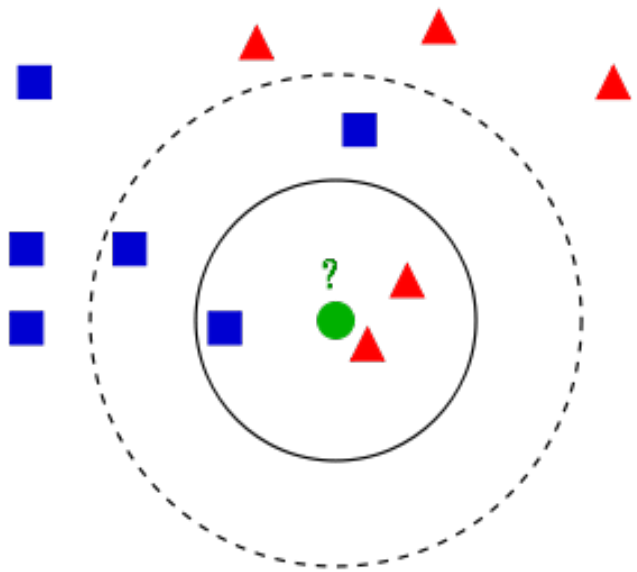
- $p(x)$ a $p(y)$ ... hustoty pravděpodobností X a Y
- $p(x, y)$ ... sdružená hustota pravděpodobnosti X a Y

Další klasifikační metody

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

k-nejbližších sousedů (k-NN)
k-Nearest Neighbors

- 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: Wiki

Klasifikační a regresní stromy (CART)
Classification and Regression Trees

- Popis vzájemných vztahů mezi pozorovanými veličinami pomocí stromu

Složení

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

Dělení

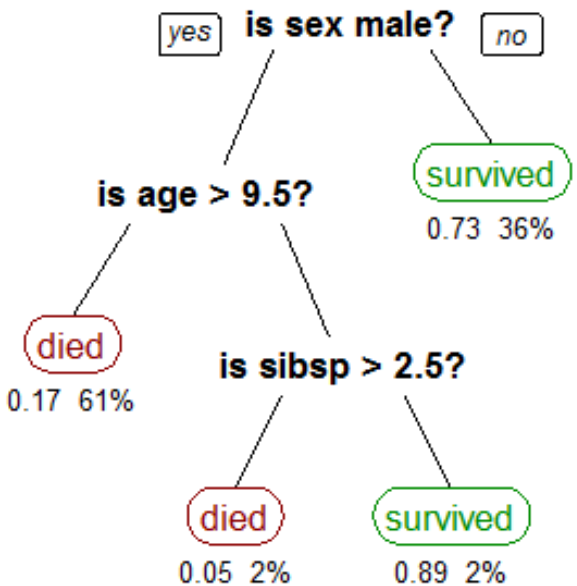
- Binární
- Ternární

Trénování

- Nastavení kritérií (vah) v uzlech stromu

Testování

- Průchod stromem dle nastavených kritérií (vah)



Obrázek: Ukázka [Wikipedia]