DSS
Decision Sciences & Systems

Technische Universität München

Tutorial Business Analytics

Tutorial 5: Naïve Bayes

Decision Sciences & Systems (DSS)
Department of Informatics
TU München

Tutorial Business Analytics

Classification

- Given a dataset $D = \{x_1, \dots, x_n\}$ of tuples and a set of classes $C = \{C_1, \dots, C_m\}$
- Each instance x_i consistis of k features (e.g. categorical or numerical)
- The classification problem is to define a mapping $f: D \rightarrow C$ where each instance x_i is assigned to one class

Tutorial Business Analytics

0-Rule

Algorithm:

i. For each class count its absolute frequency

ii. Choose the most frequent one

Tutorial Business Analytics

1-Rule

Algorithm: for each attribute

i. Count the frequency of each class per attribute value
ii. Pick the most frequent class
iii. Define a rule that assigns this most frequent class to the attribute value (rule set)
iv. Calculate error rate

⇒ Choose the attribute with the smallest error rate!

Tutorial Business Analytics

Naïve Bayes – Prerequisites

Bayes Rule is

$$\Pr(h_l|E) = \frac{\Pr(E|h_l) \cdot \Pr(h_l)}{\Pr(E)}$$

and with $E = (e_1, \dots, e_k)$ we have

$$\Pr(h_l|e_1, \dots, e_k) = \frac{\Pr(e_1, \dots, e_k|h_l) \cdot \Pr(h_l)}{\Pr(e_1, \dots, e_k)}$$

Tutorial Business Analytics

Naïve Bayes – Usage and Assumptions

- It is especially appropriate when the dimension of the feature space is high, making density estimation unattractive

- Assumption: Attributes independent and equally important

$$\Pr(h_l|E) = \frac{\Pr(e_1|h_l) \cdot \Pr(e_2|h_l) \cdots \Pr(e_k|h_l) \cdot \Pr(h_l)}{\Pr(E)}$$

$$= \frac{\prod_{i=1}^{k} \Pr(e_i|h_l) \cdot \Pr(h_l)}{\Pr(E)}$$

- These assumptions are rather optimistic, however, Naïve Bayes classifiers often outperform more sophisticated alternatives

Tutorial Business Analytics

Naïve Bayes – Algorithm

i. For each attribute, count the frequency of each class per attribute value (and resolve zero-frequency problem if needed)

ii. Calculate prior $\Pr(h_l)$ and likelihood $\Pr(e_i|h_l)$

iii. Find $\prod_{i=1}^{k} \Pr(e_i|h_l) \cdot \Pr(h_l)$

iv. Normalize the results

$$\Pr(h_l|E) = \frac{\prod_{i=1}^{k} \Pr(e_i|h_l) \cdot \Pr(h_l)}{\Pr(E)}$$

⇒ Choose the class with the highest probability