Information mining exercise, winter term 2022/23
Dr.-Ing. Benedikt Loepp (benedikt.loepp@uni-due.de)

Exercise sheet 6 Presentation on **08.12.2022**

Exercise 1: Association rules (from previous exercise sheet)

Given the following table that represents which products customers bought in a supermarket ("y" means that the customer represented by this row has bought the product, "n" means that he or she has not bought it):

ID	Beer	Potato chips	Toilet paper	Toothbrush
1	y	n	y	n
2	y	y	y	n
3	n	y	y	y
4	n	y	y	y
5	y	y	n	n
6	y	y	y	y
7	n	n	y	y
8	n	n	n	y
9	n	n	y	y

Calculate confidence and support for the following association rules:

- Potato chips → Beer
- Toilet paper and beer → Toothbrush

Explain the two terms in your own words. Also, from the results, do you think there is a causal relationship between chips and beer?

Exercise 2: Association rules in RapidMiner

Given again the weather data ARFF file from the last exercise sheet. Create a process in RapidMiner that learns association rules. To this end, perform the following steps:

(a) Read the data file.

(b) Prepare the data (hint: discretize, convert nominal to binominal attributes).

(c) Calculate frequent item sets (hint: use `FP Growth`).

(d) Learn the association rules.

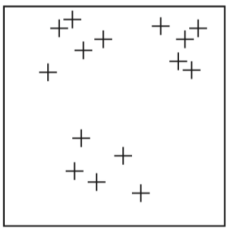
Explain the process and its output.

Exercise 3: Instance-based learning (from previous exercise sheet)

The k-NN method can be used for instance-based classification. Explain briefly the underlying idea. Also, imagine the following situation: You discover high variance in your model and suspect that your data is somewhat noisy. How you would choose k to encounter this problem, and what does this mean for the classification?

Exercise 4: k-means clustering (from previous exercise sheet)

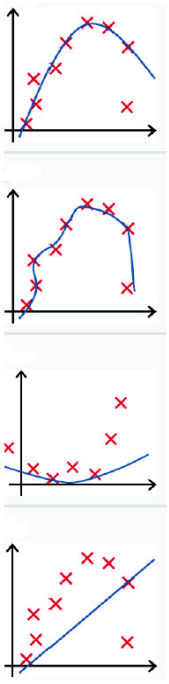
With the help of the example below, please describe briefly the idea of k-means clustering. In particular, elaborate on the meaning of k, the importance of the seeds, and the step-wise process in general. Also, explain what can happen depending on the seed selection and the number of clusters. Finally, explain how you would determine k.



Exercise 5: Overfitting

Please elaborate on the following questions:

(a) In which one of the graphs below do you think the hypothesis has overfit?



(b) Overfitting may also occur because of the discretization of numeric attributes when learning rules. Explain why this can happen and how to circumvent the problem. Please also provide an example.