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

Exercise sheet 5 Presentation on **01.12.2022**

Exercise 1: Information gain

For the weather data shown on the last exercise sheet, please calculate the information gain for the "windy" attribute. Show the calculation path.

Exercise 2: Information gain in RapidMiner

In the Moodle course, you can find an ARFF file for the weather data. Please use RapidMiner to determine the information gain for the *entire* set of attributes. Also, learn a decision tree with information gain as a criterion for the attribute selection and explain the results.

Exercise 3: Naïve Bayes, once again

With the same ARFF file, create a process in RapidMiner that uses a Naïve Bayes classifier. Next, use RapidMiner to make the same predictions as you did manually for last week's exercise sheet, i.e. for the two following instances:

- `outlook = rainy, temperature = hot, humidity = normal, windy = false`
- `outlook = overcast, temperature = hot, humidity = normal, windy = true`

Explain the output, also taking into consideration the Laplace correction.

Exercise 4: Association rules

Given the following table that shows what 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 5: Instance-based learning

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 6: k-means clustering

Considering the example below, briefly describe 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 briefly what can happen depending on the seed selection and the number of clusters. Finally, explain how you would determine k.

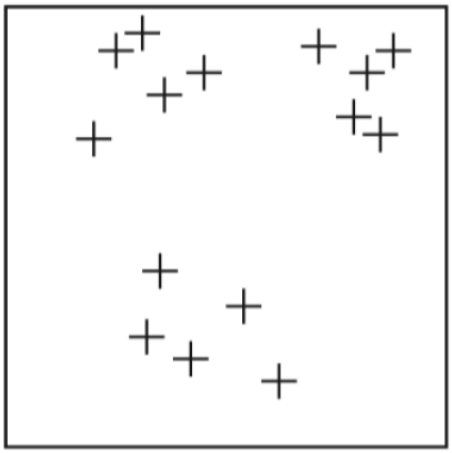

Figure 1: Some data