Prof. Dr. Andreas Hotho, Florian Buckermann
Chair of Computer Science X - Data Science

12.05.2022

1. Exercise „Data Mining“

Summer term 2022

1 General

1. Give a definition for Knowledge Discovery in Databases!
2. What is the difference between Data Mining and Data Science? (Use the definitions from the lecture)
3. Give a brief description for four typical tasks that are solved in Data Mining/Data Science!
4. Describe at least three business objectives that can be achieved through the use of Data Mining.
5. Describe three characteristics of Big Data!

2 CRISP-DM Methodology

1. Describe the six phases of the CRISP-DM process!
2. How does the preprocessing step relate conceptually to the other phases?
3. What steps are taken during the evaluation phase?

3 Databases

1. Define a database system and describe its structure!
2. Describe the fundamental difference between the data handling of a classic database application and a data mining application!

4 Statistics

1. In statistics, the features of a given dataset are divided into different data types. Describe the different types and give an example for each!

   What are the effects of the selection of the data type?

2. You retrieved the yearly income table from a list of employees:
   53, 48, 52, 56, 98, 52, 40, 49, 55

   a) Calculate the mean and median for the given data!

   b) What are the conceptual differences between the mean and the median?

   c) How does the mean and median change if you remove the outlier (=98)?

3. Given the numerical features x and y:

Tabelle 1: Features x and y

x	11	12	13	14	15	16	17	18
y	1	4	9	16	25	36	49	64

   a) Sketch the features x and y from table 1 in a cartesian coordinate system.

   b) Calculate the empirical correlation coefficient r for the features x and y!

   c) Why should we expect a higher correlation coefficient for the given values?

   d) Calculate the rank correlation coefficient r_s (Spearman) for the features x and y!

4. On average, the weather forecast for your area predicts 40 % nice weather and 60 % bad weather. The success rate for the prediction of good weather is at 80% and for bad weather at 90 %.

   You arrange an online game day with on of your friends in case of bad weather on thursday. On thursday the weather is in fact bad but your friend does not show up. Your friend points out that forecast from wednesday predicted nice weather for thursday.

   What is the probability that the statement of your friend was just a poor excuse, given the fact that you do not know the weather forecast? (Hint: Bayes' theorem)

5 OLAP

A mobile communication company wants to create an offer that is especially customized for students of different disciplines. Therefore, it should be examined which phone brand (Sungsam, HCT, Blackbear, Egg-Phone and Motololla) the students of different faculties bought in the last 5 months to optimize the offering. The following disciplines are considered: *Medizin*, *BWL*, *Informatik*, *Chemie* and *altorientalische Sprachwissenschaften*.

1. Draw a hypercube that represents the above mentioned dimensions.
2. Highlight the sales numbers for february of the Egg-Phone in the hypercube!
3. Highlight the sales numbers for the last three months of the BWL students in the hypercube!
4. Highlight the corresponding query for the following graph in the hypercube: