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!

> Definition: “Knowledge Discovery in Databases (KDD) is the non-trivial process of *identifying valid, novel, potentially useful, and ultimately understandable patterns in data*.” [Osama Fayyad et al., 1996]

2. What is the difference between Data Mining and Data Science? (Use the definitions from the lecture)

> Data Science = Set of fundamental principles that support and guide the principled extraction of information and knowledge from data
> Data Mining = The actual extraction of knowledge from data via technologies that incorporate these principles
> (see Provost et al., Data Science and its Relationship to Big Data and Data-Driven Decision Making, Big Data, Vol. 1, No. 1, 2013)

3. Give a brief description for four typical tasks that are solved in Data Mining/Data Science!

> **Classification:** Mapping of objects into predefined classes (discrete)
>
> **Clustering:** Partitioning of data into groups with similar characteristics
>
> **Regression:** Mapping of objects to continuous target variables
>
> **Association rules:** Discovery of relationships between object, e.g. “Customers who bought Nachos also bought Salsa Dip”
>
> also:
>
> **Outlier detection:** Determining of strongly diverging data points

Segmentation: Separation of data space into cohesive segments

4. Describe at least three business objectives that can be achieved through the use of Data Mining.

   - Customer relationship:
     - Special offers for certain groups of customers
     - Faster reactions to change of customer behavior
   - Target advertising
   - Optimization (e.g. storage costs)
   - Planing, forecasting (e.g. infrastructure for telecommunications)

5. Describe three characteristics of Big Data!

   Volume Increasing amount of data that has to be processed

   Velocity formerly: data is acquired in batches; nowadays: continuous datasteam; "realtime"-analysis

   Variety complex structured und diverse sources (social media, Tweets, . . .)

2 CRISP-DM Methodology

1. Describe the six phases of the CRISP-DM process!

   a) Business Understanding (What is the objective in this case?)
   b) Data Understanding (Which data is available?)
   c) Data Preparation (Preprocessing for modelling)
   d) Modelling (Applying the algorithms)
   e) Evaluation (Interpretation, Have the goals been achieved?)
   f) Deployment (Realization)

2. How does the preprocessing step relate conceptually to the other phases?

a) Selection depends on business goals (Business Understanding)

b) Preparation, modelling and evaluation depends on the selected data

c) Modelling and evaluation influences the selection (iterative process)

3. What steps are taken during the evaluation phase?

- Presentation/Visualization of the found patterns
- Evaluations of the patterns by the user
- poor evaluation $\Rightarrow$ Repeat Data Mining (change parameters, methods, data)
- good evaluation $\Rightarrow$ Integration of the patterns into a knowledge base; Usage of the knowledge for future KDD-process

3 Databases

1. Define a database system and describe its structure!

A database system (DBS) is a software system for the persistent storage of data and the efficient access to it. A DBS consists of a database (DB) that stores the actual data and a database-management-system (DBMS). The DBMS is a software especially designed for managing a database. It allows any software to access the data of the DB through a specified interface.

2. Describe the fundamental difference between the data handling of a classic database application and a data mining application!

For classic database applications, usually many insert-, update-, and delete-operations are performed concurrently on the database. For data mining, the data is usually accessed once and afterwards processed. Therefore, a data mining application usually only needs read-only access to a database.

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?

   Binary features: One fundamental difference between two concepts (yes, no)

   Nominal features: Qualitative difference without a defined order (red, green).

   Ordinal features: Ordered features without interpretable distances (table wine, quality wine, award-winning quality wine).

   Numeric features: Numbers with interpretable distances (metric).

   The selection of the data type is influenced by:

   - Information content
   - Applicability of arithmetic operations

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)?

   a) Mean: 55.89, Median: 52

   b) The median

      ... represents an actual value from the dataset

      ... is only slightly influenced by outliers

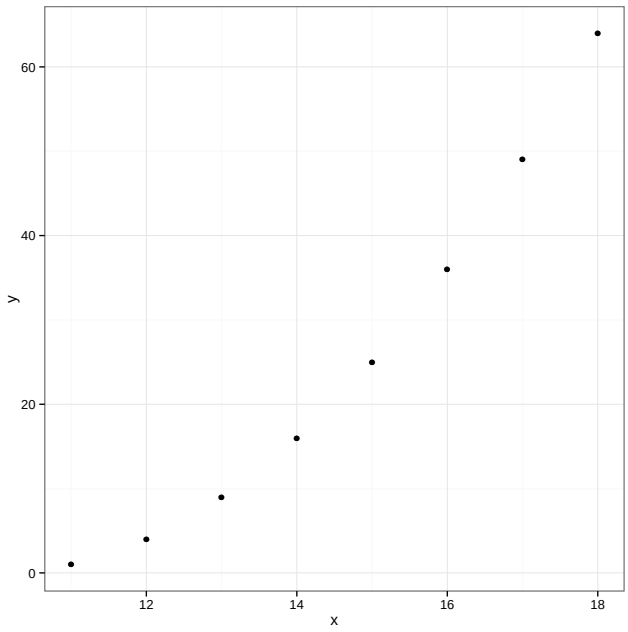
   c) Mean: 50.63, Median: 52 (Median does not change!)

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!

$$x = \frac{11 + 12 + 13 + 14 + 15 + 16 + 17 + 18}{8} = 14,5$$

$$y = \frac{1 + 4 + 9 + 16 + 25 + 36 + 49 + 64}{8} = 25,5$$

$$r_{x,y} = \frac{(11 - 14,5) \cdot (1 - 25,5) + \cdots + (18 - 14,5) \cdot (64 - 25,5)}{\sqrt{(11 - 14,5)^2 + (12 - 14,5)^2 \ldots}}$$

$$\approx 0,976$$

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

The mapping between x and y is defined with $y = (x - 10)^2$. Therewith, we have a quadratic relationship between both variables. The pearson correlation coefficient is a measure of linear correlation. Other correlations measures (like rank correlation coefficients) can detect non-linear relationships.

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

$rg(x)$	1	2	3	4	5	6	7	8
$rg(y)$	1	2	3	4	5	6	7	8

$r_s = 1$

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)

   B: Tomorrow is bad weather

   $\overline{B}$: Tomorrow is nice weather

   A_1: The forecast predicts nice weather

   A_2: The forecast predicts bad weather

   It applies: $A_1 \cap A_2 = \emptyset$

   The following probabilities are given:

   $P(A_1) = 0,4$
   $P(A_2) = 0,6$
   $P(\overline{B}|A_1) = 0,8$ and thus $P(B|A_1) = 1 - P(\overline{B}|A_1) = 0,2$
   $P(B|A_2) = 0,9$

We want to calculate the probability that bad weather has been forecasted given that today is bad weather ($P(A_2|B)$).

Bayes' theorem:

$P(A_2|B) = \frac{P(B|A_2) \cdot P(A_2)}{P(B)}$

$P(B)$ can be calculated from the conditional probabilities: $P(B) = \sum_{i=1}^{n} P(B|A_i) \cdot P(A_i)$
Here: $P(B) = P(B|A_1) \cdot P(A_1) + P(B|A_2) \cdot P(A_2)$

$P(B) = 0,2 \cdot 0,4 + 0,9 \cdot 0,6 = 0,62$

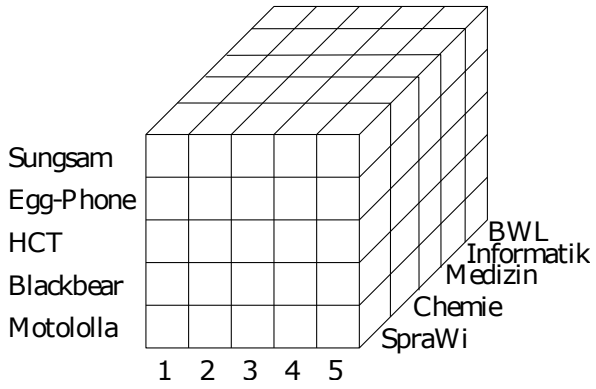
$P(A_2|B) = \frac{0,9 \cdot 0,6}{0,62} \approx 0,87$

With a probability of 87% it is a bad excuse.

5 OLAP

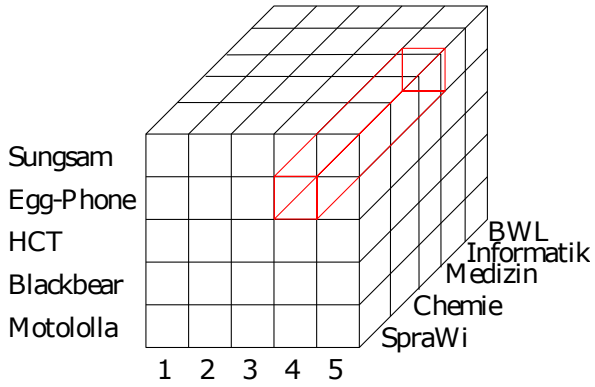
A mobile communication company wants to create an offer that is especially customized for students of different disciplines. Therefore, it should be examined which phone brand (Sungsam, HCT, Blackbear, Egg-Phone and Motololla) the students of different faculties bought in the last 5 months to optimize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

1. Draw a hypercube that represents the above mentioned dimensions.

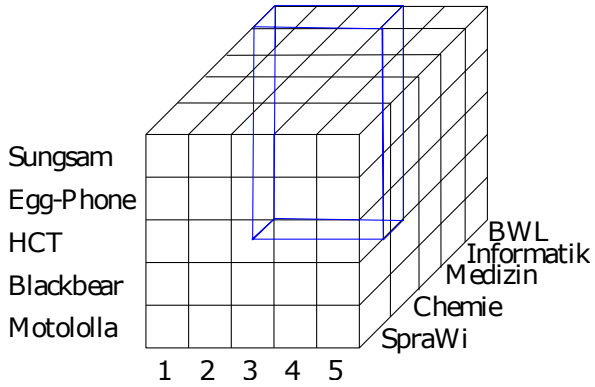


Hint: 1: Month May, 2: Month April, $\cdots$

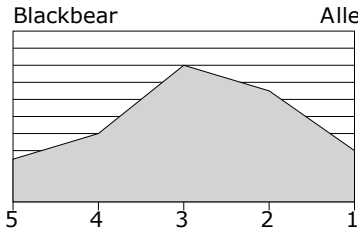
2. Highlight the sales numbers for february of the Egg-Phone in the hypercube!

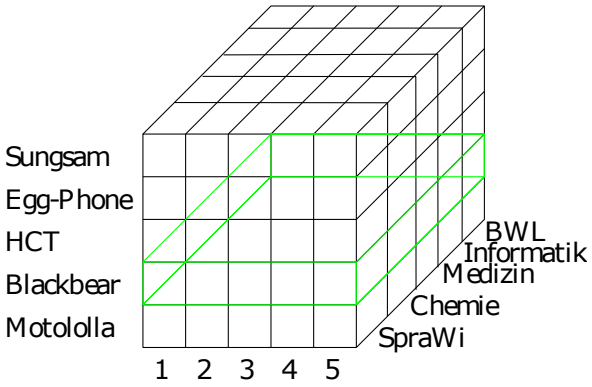


3. Highlight the sales numbers for the last three months of the BWL students in the hypercube!



4. Highlight the corresponding query for the following graph in the hypercube:




Sungsam
Egg-Phone
HCT
Blackbear
Motololla
1
2
3
4
5
BWL
Informatik
Medizin
Chemie
SpraWi