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

19.05.2022

2. Exercise „Data Mining“

Summer term 2022

1 Preprocessing - General

A mail order company wants to analyse its customers to create an offer for the most active customers. The following sample of customers is given:

Customers

Id	Name	E-Mail-Adress	Street	Place	Postal Code
1	Carla D. Eiffel		Forsthausweg 2	Duisburg	47057
2	F. Ganter	ganter@gxm.de	Geschwister-Scholl-Platz 1	München	80539
3	Jan Klein	jan_klein@gmail.com	Kaiserswerther Str. 16	Berlin	14195
4	Anton BlÃcher	bluecher@gmx.de	Rosengarten 10	Halle/Saale	6132
6	Irving, Hans	hans.irving@web.de	Christian-Albrechts-Platz 4	Kiel	24118
7	Ludwig Mann	lm@lumann.com	Kaiserswerther Strasse 16	Berlin	14195

Purchase data Online-Shop

Id	C-Id	Date	Product-Id	Price	Quantity
1	1	1.1.1970	1	12,99	2
2	1	1.1.1970	5	5,49	1
3	2	12.3.2021	3	15,00	1
4	5	20.3.2022	2	2,00	4
5	3	21.3.2021	5	5,99	1
5	3	21.3.2021	5	5,99	1
6	1	1.1.1970	1	12,99	255

Purchase data phone order

Customer-Id	Date	Product	Price	Quantity
3	3.3.21	2	2	2
3	10.3.21	1	12,99	1
4	4.3.21	2	2,00	1
1	3.3.21	1	12,99	5
7	9.3.21	5	5,99	1

1. Perform all necessary steps of preproccesing on the given dataset! Which of your actions relates to which step of preprocessing?
   Hint: Just describe time-consuming steps. You do not have to perform them!
2. Discuss the remaining problems and name possible ways to solve them!

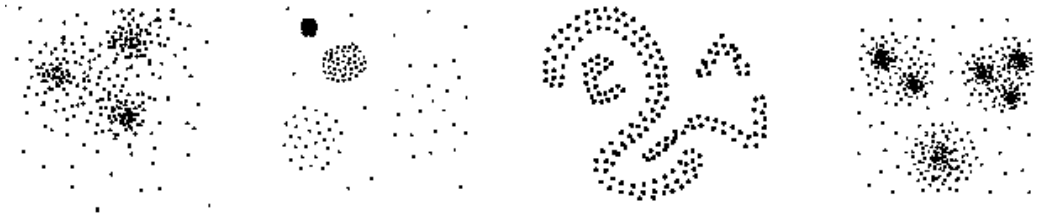
2 Preprocessing - Normalization

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

1. Normalize the values with
    - Rescaling (Min-Max Normalization)
    - Standardization (Z-score Normalization)
2. What are the conceptual differences between rescaling and standardization?
3. Normalize the new value of 35 without recalculating the statistics (min, max, mean, std.dev). Compare the results and describe any issues that you observe.

3 Clustering - Basics

1. Define the term clustering!
2. Consider the different types of clusters and discuss difficulties that may arise if you want to perform clustering with such types of clusters.



4 Clustering - Similarity measures and metrics

1. Define and describe the Manhatten- and L_2-distance.
2. What is the difference between a distance function and a metric?
3. Given are three documents A, B and C. The documents A and B, as well as B and C are similar. Is it possible to derive the similarity of A and C with respect to a metric or distance function? Give reasons for your answer!