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!

Customers

Id	Name	E-Mail-Adress	Street	Place	Postal Code
1	Carla D. Eiffel		Forsthausweg 2	Duisburg	47057
2	F. Ganter	ganter@gmx.de	Geschwister-Scholl-Platz 1	München	80539
3	Jan Klein	jan_klein@gmail.com	Kaiserswerther Strasse 16	Berlin	14195
4	Anton Blücher	bluecher@gmx.de	Rosengarten 10	Halle/Saale	06132
~~6~~	~~Irving, Hans~~	~~hans.irving@web.de~~	~~Christian-Albrechts-Platz 4~~	~~Kiel~~	~~24118~~
7	Ludwig Mann	lm@lumann.com	Kaiserswerther Strasse 16	Berlin	14195

a) Remove customer 6, as no transaction have been recorded for this customer (selection)

b) Umlaute, leading zeros in postal code, correct mail of customer 2 (cleaning)

c) Change *Str.* to *Straße* (Consistency)

Purchase data Online-Shop

Id	Customer-Id	Date	Product-Id	Price	Quantity
1	1	11.6.2022	1	12,99	2
2	1	11.6.2022	5	5,99	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	21.3.2022	1	12,99	1

a) Adjust column headings (Consistency)

b) Insert missing dates, adjust price for product 5, adjust amount for transaction 6 (correction, the filled in values are exemplary)

c) Remove transaction 4 as customer 5 is missing; remove duplicate transaction 5 (selection)

Purchase data phone order

Id	Customer-Id	Date	Product -Id	Price	Quantity
7	3	3.3. 2021	2	2,00	2
8	3	10.3. 2021	1	12,99	1
9	4	4.3. 2021	2	2,00	1
10	1	3.3. 2021	1	12,99	5
11	7	9.3. 2021	5	5,99	1

a) Supplement transaction ids (insert missing values)

b) Adjust column headings (consistency)

c) Ensure consistent formats for dates and prices (consistency)

Further steps:

a) Aggregate data

b) Feature selection

c) Feature construction

d) Feature normalization

e) Feature categorization

2. Discuss the remaining problems and name possible ways to solve them!

- Missing values may be imputed from a logfile
- Erroneous values may be replaced with an average value or fully removed
- Strict and specific input mask for entering the customer data, usage of transaction and logging is useful to prevent mistakes or correct them afterwards

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.

1. Results:

Original	53	48	52	56	98	52	40	49	55
Rescaled	0.22	0.14	0.21	0.28	1.00	0.21	0.00	0.16	0.26
Standardized	-0.19	-0.51	-0.25	0.01	2.71	-0.25	-1.02	-0.44	-0.06

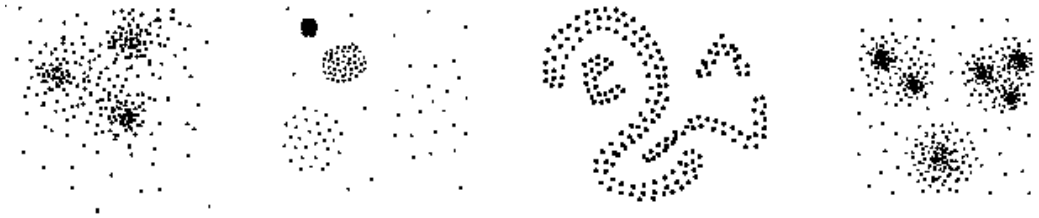
2. Concepts
   - Rescaling: The lowest value (min) is set to 0 and the highest value (max) is set to 1. Each value in between describes the relative position in the min-max-range.
   - Standardization: The mean value is set to 0. A value $\neq 0$ describes the distance from the mean in units of the standard deviation.
3. Rescaled: -0.09
   Standardized: -1.34
   If new values are below/above the minimum/maximum value, rescaling will produce values that are out-of-range. Standardizing new values, does not result in erroneous values

3 Clustering - Basics

1. Define the term clustering!

   - Identification of a finite amount of categories, classes or groups (*Cluster*) in a dataset
   - Objects in the *same* cluster should be as similar as possible
   - Objects from *different* clusters should be as dissimilar as possible

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



   - Convex clusters with many outliers: Influences the centroids and size of clusters
   - Convex cluster with different density: Influences the size & detection of clusters
   - Density-bases, non-convex clusters: Appropriate methods are needed to detect clusters
   - Hierarchical clusters: Outcome depends on number of searched clusters

   Often it is reasonable to analyse the data and systematically select a method.

4 Clustering - Similarity measures and metrics

1. Define and describe the Manhatten- and L_2-distance.

   - Manhattan distance $dist(x, y) = \sum_{i=1}^{d} |x_i - y_i|$ measures the distance between two points on a grid.

- L^2 distance (euclidean distance) $dist(x, y) = \sqrt{\sum_{i=1}^{d} (|x_i - y_i|)^2}$ measures the distance between two points in an euclidean space.

2. What is the difference between a distance function and a metric?

   A distance measure does not satisfy the triangle inequality.

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!

   a) Metric: Yes! With the triangle inequality the following rule applies: the distance between A and C is bounded by the sum of the distances between A and B plus B and C

   b) Distance measure: No! The triangle inequality is not satisfied. Thus, we can not infer a distance.