Universität Wien
Fakultät für Informatik
Prof. Claudia Plant
Prof. Torsten Möller
Dr. Thomas Torsney-Weir

Foundations of Data Analysis

WS 2018/2019

Pen and Paper 2: FDA WS 18/19

General Remarks:

- The deadline for the submission is on 23.01.2019 at 9:45 a.m. Please upload your solution on Moodle. No deadline extension is possible.
- Upload your solutions as PDF format, with the following naming scheme **matrikelnumber_PP_Part2.pdf**.
- If you have problems do not hesitate to contact the tutor or post a question on the Moodle system.
- The solutions for this assignment will be discussed in the lecture on 23.01.2019.

Aufgabe 2-1 PCA (30P)

One of the most important techniques for dimensionality reduction is PCA. We have the following observations of a two-dimensional data set:

x	y
1	1.5
2	1
2.5	1.5
3	2.5
3.5	3.5
4	5.5
5	5.5

(a) Perform a PCA as has been described in the lecture and find the main component of the data set (25P).

(b) Why is Dimensionality reduction performed? Give at least two reasons. (2 keywords are enough) (5P)

Aufgabe 2-2 Frequent Itemsets Mining (30P)

Consider the following transaction database D over the items $I = \{A, B, C, D, E, F, G\}$.

TransID	Items
1	A B C
2	A D G
3	D E F
4	A B D G
5	B C E G
6	A C E F
7	B C E
8	A B C E

(a) Given minSupport threshold $\sigma = 25\%$ (i.e., 2 transactions), apply the Apriori algorithm and extract all frequent itemsets w.r.t. the given threshold. Please explain in the solution all the steps you took.

In particular, please include for each level the candidate set (C_k) after the join step, annotate which objects are pruned and give the explicit reason for pruning. Also give the pruned large/frequent sets (L_k).

(b) Compute the confidence of the rule $\{B, E\} \rightarrow \{C\}$ in the above database of transactions for the given minSupport threshold σ.

Aufgabe 2-3 k-Means and k-Medians (30P)

Let the following dataset be given:

Index	height	weight	class
1	150	65	1
2	150	75	1
3	170	75	2
4	175	70	2
5	175	80	2
6	210	100	2

After visualizing the data please apply k-Means and k-Medians considering the provided parameters. Please evaluate and discuss the results of both clustering algorithms.

Use the Manhatten Distance function for this task:

$$L_1(x,y) = |x_1 - y_1| + ... + |x_n - y_n| = \sum_{i=1}^{n} |x_i - y_i|$$

Write down every step, you can round off fractures for easier calculation. In case of even numbers of samples for the median, pick the smaller one as your middle value.

(a) Draw the data from the table above in $\mathbb{R}^2$

(b) Apply k-Means with the following parametrization:

- $k = 2$
- Centroids $C_1 = (160, 70)$, $C_2 = (190, 85)$

Write down the positions of the cluster centers after convergence

(c) Apply k-Medians with the same parametrization. Write down the positions of the cluster centers after convergence

(d) Create a confusion matrix with your cluster labels in comparison to the class labels for both variants. Calculate the precision, recall and F-Measure of both clustering results.

(e) Discuss your result. Which variant performs better and why? Could you improve k-Means or k-Medians to produce better results?