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

# Foundations of Data Analysis

SS 2019

## Pen and Paper 2: FDA SS 2019

**General Remarks:**

- The deadline for the submission is on 19.06.2019 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_PP2.pdf**.
- If you have problems do not hesitate to contact the tutor or post a question on the Moodle system.

**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 |
|---|---|
| 2.5 | 2.4 |
| 0.5 | 0.7 |
| 2.2 | 2.9 |
| 1.9 | 2.2 |
| 3.1 | 3.0 |
| 2.3 | 2.7 |
| 2 | 1.6 |
| 1 | 1.1 |
| 1.5 | 1.6 |
| 1.1 | 0.9 |

(a) Perform a PCA as has been described in the lecture and find the main component of the data set. (15P)

(b) transform the data set, but use only the main component, i.e. reduce the data to 1D along the main component. (15P)

**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 |

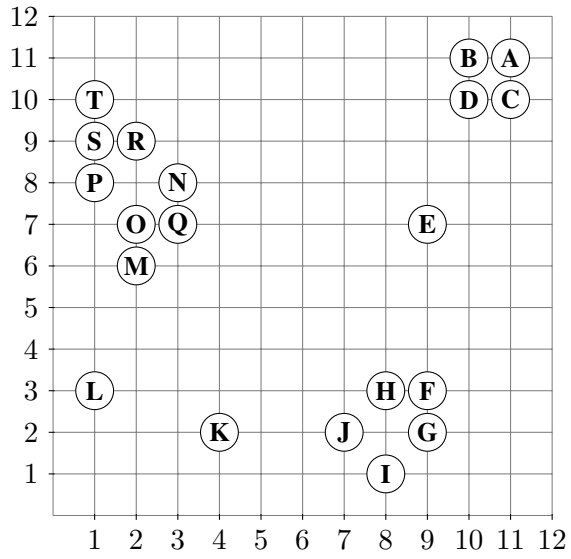
(a) Given minSupport threshold $\sigma = 25\%$ (i.e., 2 transactions), apply the Apriori algorithm and extract all frequent itemsets w.r.t. the given threshold. Please explain in the solution all the steps you took.

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 $\{A, B\} \rightarrow \{E\}$ in the above database of transactions for the given minSupport threshold $\sigma$.

**Aufgabe 2-3 DBSCAN (30P)**

Given the following data set:




As distance function, use the Manhattan Distance:

$$L_1(x, y) = |x_1 - y_1| + |x_2 - y_2|$$

(a) What are core points, border points and noise points? Explain the differences.

(b) Compute DBSCAN and indicate which points are core points, border points and noise points.
Proceed in lexicograchic order. Points cannot change their cluster label once they are assigned to a cluster.

Use the following parameter settings:

- Radius $\varepsilon = 1.1$ and *minPts* $= 2$
- Radius $\varepsilon = 1.1$ and *minPts* $= 4$
- Radius $\varepsilon = 2.1$ and *minPts* $= 4$
- Radius $\varepsilon = 2.1$ and *minPts* $= 10$

(c) When *minPts* $= 2$, what happens to border points?

(d) How does one choose $\varepsilon$ so that there is only one cluster? And how does one choose *minPts*, so that every data point is an outlier?