M. Strohmaier, P. Singer, F. Lemmerich, E. Vasilev
GESIS - Leibniz Institute for the Social Sciences – Computational Social Science Department
Universität Koblenz-Landau

01.02.2016

4. Home Assignment "Machine Learning and Data Mining"

Winter semester 2016 / 2017

You should work on the assignment in groups of 2–3 participants. Send your solution (as .pdf) by email to sent to `vasilev@uni-koblenz.de`, `florian.lemmerich@gesis.org`, and `philipp.singer@gesis.org` until **Tuesday, 7th of February, 23:59h**. Use `[ML-Assignment]` as the email subject and do not forget to denote the name of all contributing students in the pdf as well as in the email. The deadline is strict!

1 Clustering

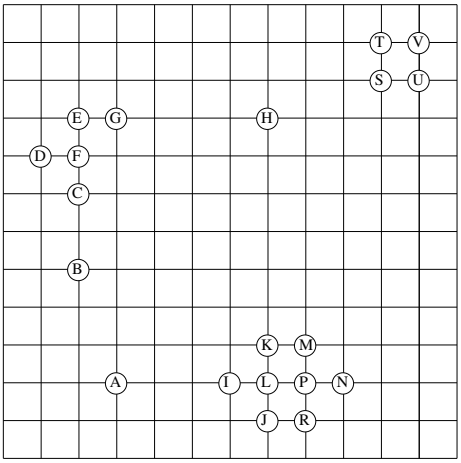
1. Given is the following dataset:

x	1	6	8	3	2	2	6	6	7	7	8	8
y	5	2	1	5	4	6	1	8	3	6	3	7

   Compute a clustering of this two-dimensional dataset using the k-means algorithm with $k = 3$. Apply the L_2 metric (Euclidean distance) and use the first three data points as the initial parameters. Update the centroids only after a full iteration (variation of Lloyd).

   Follow the movement of the centroids in a sketch of the two-dimensional space!

2. Compute for the following dataset two dendrograms, one using single-link and one using average-link clustering. Use the Manhattan distance as distance function between data points.



2 Association Rule Mining

A dataset D consists of the following transactions:

Transactions	Items
t_1	Chips, Apples, Diapers
t_2	Apples, Beer
t_3	Chips, Apples, Beer
t_4	Chips, Diapers, Electronics
t_5	Apples, Electronics
t_6	Chips, Apples, Beer
t_7	Chips, Diapers
t_8	Apples, Beer

Determine for these transactions the frequent itemsets with a minimum support of 25% ! To determine them, use *exactly* the apriori algorithm!

3 Sequence Modeling

1. a) Given the following sequence of three states (R,A,B), learn the parameters of a first-order Markov Chain model (transition matrix) by using Maximum Likelihood Estimation (MLE).

   `R - A - A - B - A - B - B - A - A - A - R`

   b) Given the learned transition parameters, calculate the likelihood for the following sequence.

   `R - B - B - B - A - A - B - A - B - R`

   What we can see is that we have transitions in the data, that we have not yet seen in the training data. Thus, the likelihood is zero. The log-likelihood would be undefined. As this is no reasonable solution, we would apply smoothing to our parameter estimation, e.g., by adding one pseudo count to each transition.

   c) Given the learned transition parameters, predict the most likely next state in the following sequences and explain the process.

   `R - B - B - ?`

   `R - A - B - A - B - A - ?`

   d) Explain what the Markovian assumption for the first-order Markov Chain model is. How can we extend the first-order Markov Chain model when we lessen this assumption? Given these extended models, how can we derive which one is the most suitable for a given sequence dataset? Elaborate your thought process.

   e) What do you think is a "zero-order" Markov Chain model? This has not been covered in the lecture.