**RWTH Aachen** **October 26th, 2015**
**Lehrstuhl für Informatik 9**
**Prof. Dr. T. Seidl**

## Exercise 1 for the lecture

# Data Mining Algorithms

## WS 2015/2016

***Hand in your solutions on Nov. 2$^{nd}$ before the lecture. Please form groups of 3-4 students (exceptions will only be granted for this exercise). You can use the L2P discussion board, if you didn't find a group yet. The first tutorial will be held on Nov. 6$^{th}$.***

**Note: All commands for the R-exercises are required to be provided with comments, indicating which task the commands belong to. All R script files should contain a comment-line with the names and matriculation numbers of the group-members. Send all R-files to yifeng.lu@informatik.rwth-aachen.de or faerber@informatik.rwth-aachen.de. The subject of the mail must start with "[DMA1]".**

### Exercise 1.1) Preliminary tasks for R **7 points**

Download and install R (http://www.r-project.org/). We suggest to use the Integrated Development Environment **R-Studio** (http://www.rstudio.com/). You may also choose any other IDE or just use R console for development.
Install the following packages in R as you may need them later on:

- MASS
- rgl
- lattice

(Tools->Install Packages in R-Studio or take a look at ?install.packages in the R console)

Save the commands for all tasks below into an **R Script** file.

a) Create a vector $v_1 = (4, 9, 10)$ using the **c()** command.
b) Create a vector $v_2 = (12, 9, 6, 3, 0)$ using the **seq()** command.
c) Create a matrix $M = \begin{pmatrix} 4 & 5 & 6 & 9 & 4 \\ 2 & 91 & 15 & 8 & 6 \\ 3 & 42 & 9 & 6 & 2 \\ 4 & 17 & 2 & 7 & 7 \end{pmatrix}$.
d) Extract a sub-matrix $M_1 = \begin{pmatrix} 91 & 15 & 8 \\ 42 & 9 & 6 \\ 17 & 2 & 7 \end{pmatrix}$ out of M.
e) Compute the result of the matrix-vector multiplication: $M_1 \cdot v_1$.
f) Compute the solution for the vector x in the equation: $M_1 \cdot x = v_1$.
g) Create a data frame **shopping** with columns {"item", "price", "amount"} and the following rows:
("bread", 2.99, 1)
("milk", 0.99, 2)
("sugar", 0.99, 1)
("bread", 1.99, 1)
("sugar", 0.99, 3)
("rice", 1.79, 2)

h) Extract the sub-data frame that only contains the columns “item” and “price”.

i) Create a vector **v** that contains the values of the last column of the data frame **shopping** (the command **typeof()** should return either “integer” or “double” and not “list”). You can use the operator **[[ ]]**.

**Exercise 1.2) R and the KDD process** **8 points**

Two data sets of iris flowers *iris1.csv* and *iris2.csv* were collected by different biologists. For information about the data sets see *iris.names* (all files are available on the L2P).

Save the commands for the tasks b) – g) into an **R Script** file.

a) Download the files from the L2P.

b) Create one data frame that contains the data from both data sets.

c) Some of the entries contain missing values. Replace each missing value in each column with the column-wise arithmetic mean.

d) Create a data frame named **test** that contains the attributes *petal length* and *petal width*.

e) Use the built-in function **kmeans()** in R to cluster the data frame **test** with different parameters $k \in \{2, 3, 4\}$ (the corresponding parameter is called *centers*). Store the results in different variables.

f) Use the function **plot()** to plot the data frame **test** for each clustering from part e). Each point should be colored according to the cluster assignment.

g) Use the function **plot()** to plot the data frame **test**, but colour the points according to the original classes.

h) Map the tasks a) – f) to the steps of the KDD process (see lecture).

**Hint: To color the data points you can pass an assignment vector to the function** plot() **by passing it to the “col” parameter, e.g.** col = clustering$cluster**.**

**Remark: The original iris data set can be downloaded from the UCI Machine Learning Repository.** Bache, K. & Lichman, M. (2013). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

**Exercise 1.3) Incremental Aggregation:** **5 points**

Given a Data Warehouse with e.g. 10 million entries, additional 1000 entries arrive each day. Rather than recomputing the desired aggregates, an incremental adaptation to the new data should be supported. In order to accelerate the (re-)computation, precomputed intermediate results shall be stored and intermediate results for the new entries shall be computed. What (and how many) values suffice when considering the following aggregates? For each measure note whether it is an algebraic, holistic or distributive measure.

a) mean value,

b) 25% quantile of the values,

c) variance