RWTH Aachen
Lehrstuhl für Informatik 9
Prof. Dr. T. Seidl

November 2nd, 2015

Exercise 2 for the lecture

Data Mining Algorithms

WS 2015/2016

Hand in your solutions on November 9th before the lecture. The tutorial for this exercise will be held on November 13th. Solutions of groups with less than 3 students will get a last warning and will not be graded for Exercise 3.

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 all group-members. Send all R-files to yifeng.lu@informatik.rwth-aachen.de. The subject of the mail must start with "[DMA1]". Additionally hand in a print-out of the code together with the non-R tasks.

Exercise 2.1) Plot & Visualization in R — 6 points

Download the file *original_iris.csv* from the L2P.

a) Create a data frame that stores the data from *original_iris.csv*. The values for the attribute *class* (attribute 5) should be stored as *integers* instead of as *strings* / *factors*.
Hint: You can use the function as.integer() **on a vector of factors to access the internal representation as integers.**

b) Use the data frame you created in part a) to plot the data (attributes 1 – 4) using the function **parcoord()** from the package *MASS*. Color the data according to the classes (attribute 5).
Note: If you were not able to do part a) you can also use the file *iris2.csv* from last week's exercise.

Exercise 2.2) Star and snowflake schema: — 7 Points

Your task is to design a data warehouse application for a university information system. The major focuses are the lectures for students held by professors where time, location, study path, addresses and department chairs are of interest. Professors are part of their respective department chair.

Write down the fact and dimension tables for both the star and the snowflake schema.

Exercise 2.3) Data quantization: — 2 points

Suppose that the data for analysis includes the attribute *age*. The *age* values for the data tuples are (in increasing order): 13, 14, 14, 15, 17, 18, 21, 21, 23, 23, 24, 27, 27, 28, 32, 32, 33, 36, 36, 37, 37, 37, 40, 44, 46, 50, 69.

Apply the binning technique equi-height (also called equi-depth or frequency), using a bin depth of 3, to partition the data.