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

December 7th, 2015

Exercise 7 for the lecture
Data Mining Algorithms
WS 2015/2016

Hand in your solutions on December 14th before the lecture. The tutorial for this exercise will be held on December 18th.

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 faerber@informatik.rwth-aachen.de. The subject of the mail must start with "[DMA1]".

Exercise 7.1) Linear Discriminant Classifier (SSE) **7 = 4+3 points**

a) Download the *trainingData.csv* file from the L2P. This training data includes 2-dimensional points with two possible class labels 1 and 2. Finish the training function *LDCTraining()* in R. Use it to compute the parameters $\mathbf{w}$ and w_0 w.r.t the training data.

b) Download the *newData.csv* file from L2P. This data includes 2-dimensional points without class labels. Finish the classification function *LDCClassify()* in R. Apply the classification function to the data with the parameters you computed in *a)*.

Note: you only need to submit the two R script files. The script we will use to test your functions is also provided on L2P.

Exercise 7.2) m-fold cross validation (R-exercise) **7 = 3+3+1 points**

Download the data set "**m-fold_cross_validation_data.csv**" from the L2P. This dataset consists of 5 classes, each of them contains 90 objects.
Save your solution for part a) as an **R script** file.

a) Train Naïve-Bayes classifiers on the data set (you can use the function **naiveBayes()** from the package 'e1071'). Calculate the accuracy (percentage of correctly classified objects) for naïve Bayes on the given data set using m-fold cross validation with *m=3, 5, 7*.

   Note: If you want to use the function naiveBayes() **from the package 'e1071', make sure to transform column 11 of type** int **into the type** factor **by using the** factor() **command.**

b) Suppose someone implemented the m-fold cross validation from exercise a) poorly. The evaluation of the classifier on the provided .csv-file yielded the following results.

m	Bayes accuracy
3	23.3
5	0
7	61.9
10	62.2

   What was the error in the implementation of the m-fold cross validation? Describe and explain the result for *m=3, m=5* and *m=7* in short and precise sentences.

c) How would you generally choose the value for *m* (in the m-fold cross validation technique) in order to get accurate results? Justify your answer.