Machine Learning
Exercise Sheet 5

Prof. Dr. Dr. Lars Schmidt-Thieme, Osman Akcatepe
Information Systems and Machine Learning Lab (ISMLL)
University of Hildesheim

06. December 2011
Deadline 12. December, 14.15

Linear discriminant analysis (5 points)

Scientists compared the soil of Iowa to the other soil, which contains a certain bacterium (class 1) and does not contain bacterium (class 2), respectively. At the same time they observed the variables x_1 (pH-value) und x_2 (nitrogen content). Given the number of the instances per class, the mean of the vectors and the covariance matrices for the two types of soils as the following:

$$n_1 = 13, \qquad n_2 = 10$$
$$\mathbf{m}_1 = \begin{pmatrix} 7.8 \\ 45 \end{pmatrix}, \qquad \mathbf{m}_2 = \begin{pmatrix} 5.9 \\ 20.8 \end{pmatrix}$$
$$\mathbf{S}_{V1} = \begin{pmatrix} 0.5 & 4.5 \\ 4.5 & 147.2 \end{pmatrix}, \quad \mathbf{S}_{V2} = \begin{pmatrix} 0.1 & 0.2 \\ 0.2 & 24.2 \end{pmatrix}$$

a) Develop the discriminant functions for the both classes.

b) Allocate the observation $x = \begin{pmatrix} 6 & 52.5 \end{pmatrix}^T$ to one of the two classes.

c) Is it about linear or quadratic discriminant analysis? Name the difference between LDA and QDA.

Data import in R (2 points)

Read capitals 6 and 7 from „An Introduction to R“.

a) What is the difference between a list and an array in R? Name the three possibilities that how they can be accessed over the components of a list. Why are *data frames* especially important constructs in R?

b) Download the data Wein from the *UCI Machine Learning Repository* (`http://archive.ics.uci.edu/ml/datasets/Wine`) and load it in R.

LDA and QDA in R (3 points)

Load the library *MASS* with `library(MASS)`. Create two classification models, which determines the first variable as target variable (class) and the remaining variables as predictor variables: Linear discriminant analysis (`lda`) and quadratic discriminant analysis (`qda`). The functions `lda` and `qda` would be similar as `lm` and `glm`, e.g. `glm(Survived ~., data=Titanic, family=binomial)` for a logistic regression over the dataset *Titanic* included in R.

a) If you call `lda` and `qda` with the parameter `CV=1`, you get a prediction for each entry in your dataset: `result <- qda(Survived ~., data=Titanic, CV=1)`. Compare the methods LDA and QDA, so that you adjust `result$class` with the first column of the dataset for each time.

b) Why can't you establish just a logistic regression model for the dataset Wein? Give reasons.