Machine Learning
Exercise Sheet 6

Prof. Dr. Dr. Lars Schmidt-Thieme, Martin Wistuba
Information Systems and Machine Learning Lab
University of Hildesheim

December 3rd, 2013
Submission until December 10th, 13.00 to wistuba@ismll.de

Discriminant Analysis (5 Points)

Scientists compared the earth of Iowa which contains a specific bacterium (class 1) with other earth that does not contain it (class 2). They observed the variables x_1 (pH value) and x_2 (nitrogen content). The number of instances pro class, the mean of the vectors and the covariance matrix for both kind of earths is given as follows:

$$n_1 = 13, \qquad n_2 = 10$$
$$\mathbf{m}_1 = \begin{pmatrix} 7.8 \\ 45 \end{pmatrix}, \qquad \mathbf{m}_2 = \begin{pmatrix} 5.9 \\ 20.8 \end{pmatrix}$$
$$\mathbf{S}_{W1} = \begin{pmatrix} 0.5 & 4.5 \\ 4.5 & 147.2 \end{pmatrix}, \quad \mathbf{S}_{W2} = \begin{pmatrix} 0.1 & 0.2 \\ 0.2 & 24.2 \end{pmatrix}$$

a) Estimate the discriminant functions for both classes.

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

c) Is this a linear or a quadratic discriminant analysis? Mention differences between LDA and QDA.

Reading Data from Files in R (2 Points)

Read the chapters 6 and 7 of „An Introduction to R“: `http://cran.r-project.org/doc/manuals/R-intro.pdf`

a) What is the difference between a list and an array in R? Mention three possibilities to access the components of a list. Why are *data frames* important constructs in R?

b) Download the Wine data set from the *UCI Machine Learning Repository* (`http://archive.ics.uci.edu/ml/datasets/Wine`) and import it with R. Submit the source code.

LDA and QDA in R (3 Points)

Import the *MASS* library with `library(MASS)`. Create two different classification models for the Wine data set which is using the first variable as target (class) and the others as predictors: linear discriminant analysis (`lda`) und quadratic discriminant analysis (`qda`). The functions `lda` and `qda` are used similar to `lm` and `glm` e.g. `glm(Survived ~ ., data=Titanic, family=binomial)` for a logistic regression on the *Titanic* data set contained in R.

a) If you use `lda` and `qda` with parameter `CV=1` you get for each instance in your data set one prediction: `result <- qda(Survived ~ ., data=Titanic, CV=1)`. Compare the methods LDA and QDA by comparing `result$class` with the first column of the data set.

b) Why cannot you create a logistic regression model for the Wine data set? Explain.