Machine Learning – WS'12
Exercise-7

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

Linear Discriminant Analysis

Problem-1:

Scientists of Iowa state have acquired the samples of water from state's reservoirs. Some water samples contain a particular bacterium (class 1) while other do not contain (class 2). The samples have two observed variables x_1 (pH) and x_2 (Nitrogen content). The number of instances in each class, the average of the variable vectors and the covariance matrices for the two types of water samples are given as follows:

$$\hat{n}_1 = 13, \quad n_2 = 10$$
$$\hat{\mu}_1 = \begin{pmatrix} 7.8 \\ 45 \end{pmatrix}, \quad \hat{\mu}_2 = \begin{pmatrix} 5.9 \\ 20.8 \end{pmatrix}$$
$$\hat{\Sigma}_1 = \begin{pmatrix} 0.5 & 4.5 \\ 4.5 & 147.2 \end{pmatrix}, \quad \hat{\Sigma}_2 = \begin{pmatrix} 0.1 & 0.2 \\ 0.2 & 24.2 \end{pmatrix}$$

a) Determine discriminant function for the two classes.

b) Assign the observation $x = (6 \quad 52.5)^T$ to one of the classes.

c) Are these LDA or QDA ?

Problem 2:

a) Quadratic discriminant analysis generates non-linear decision boundaries. Why it is sometimes necessary to allow non-linear decision boundaries? What are the disadvantages to use more complex decision boundaries?

b) Why QDA produces non-linear decision boundaries? But LDA does not? As one could transform the data to use LDA rather than complex QDA (see LDA-QDA discussion in lecture), the distinction between two classes Y={0,1} is sufficient? Can the disadvantages of complex QDA be eliminated?

Problem-3:

Suppose we have the following training sample:

	$t=1$			$t=2$		
x_1	2	4	3	5	3	4
x_2	12	10	8	7	9	5

We assume x1 and x2 follow a bivariate normal distribution within each group, where the covariance matrix is assumed to be the same in both groups.

a) Estimate the group means, covariance matrix (unbiased), and group prior probabilities $\pi_k = p(Y = k)$ from this training sample.

b) Estimate the linear discriminant functions a1(x1, x2) and a2(x1, x2) for class 1 and 2 respectively.

c) Give one linear classification rule for this problem and construct a *confusion matrix* by applying to the training sample. What is in-sample error rate?

d) Draw the border lines between the areas that belong to class 1 and 2, respectively, in a scatter plot of the data. Can you find a straight line that has low in-sample error rate?