EFME 2013 LU Exercise 2

Exercise due: the 15th of December 2013, 23:55

Abstract

The goal of this exercise is to give you an insight into insight in the basic classification algorithms, and the evaluation procedure.

1 How to submit your report

You should make the report of the exercise available as a PDF document. The hand-in for this and later laboratory exercises is done by using TUWEL before the hard deadline (see above) containing the following (please follow this standard):

- In the subject: 'EFME-', your group and exercise number, e.g. 'EFME-Gr0-Ex0' for the zero group and zero exercise.
- As attachments (.7z or .zip): the PDF document with the ALL the MATLAB code necessary to RUN your solution (including your chosen images, and an file that runs ALL your MATLAB code).

You may write the PDF document in English or in German. Include in the document your results and most importantly, a discussion of the results. DO NOT forget to attach the MATLAB code in the same zip-file. Be careful: If the attached MATLAB does NOT run, we will reject your exercise completely. We ask you to put in the MATLAB code also a matlab file (e.g. main.m or exercise.m etc) that runs ALL your solutions (just by one call to it). It is NOT necessary to include a copy of all the code in the PDF document, although key parts necessary to explain a point can be included. It is necessary to comment the code in details. More details are available on the TUWEL web page.

2 Wine Classification (10 points)

2.1 Introduction

The main objective of this exercise is to classify the type of wines. The dataset contains the results of a chemical analysis of wines made in the same region in Italy but derived from three different cultivars (i.e. we have a 3 class problem).

2.2 Data set

The analysis determined the quantities of 13 constituents found in each of the three types of wines, all of them continuous. The dataset consists of 178 instances, such that the class 1 has 59 samples; class 2 has 71 samples; and class 3 has 48 samples. The data are not standardised, thus a data normalization is recommended. This dataset can be found in TUWEL or http://archive.ics.uci.edu/ml/datasets/Wine. Please note that the first entry in the matrix is the class label.

Use the k-NN classifier that you programmed in the first Lab. Answer the questions below:

1. Distinguish between the three classes. Discuss the performance of your classification.
2. Which combination of the 13 features leads to the best results in the questions above? Explain your findings.

BE CAREFUL: You have to use test and training set in this exercise to evaluate your algorithms. Please report on how did you create these datasets. Provide classification errors for $k = 1$ (Nearest Neighbour classifier) as well as for at least two other (best) values of k. Does the classification error depend on the choice of the training and test data set? Show the performance of the classifier by using (at least) two different test and training sets. Is you classifier good enough to be used in real life?

3 Mahalanobis Distance (11 points)

Implement a Mahalanobis distance classifier in MATLAB. The classifier should learn the class mean vectors and covariance matrices from a training set and then use them to classify a test set. For an unknown feature vector, the classifier should calculate the Mahalanobis distance from this vector to each of the class means, and assign the vector to the class corresponding to the smallest Mahalanobis distance. It should be possible to set the classifier to calculate a full covariance matrix for each class, **but your are going to make the following assumptions**:

- The covariance matrices are diagonal. To implement this, the off-diagonal elements of each matrix are simply set to zero.
- All classes have identical covariance matrices. To implement this, calculate the sum in the equation for estimating the covariance matrix for each class, add the sums for all the classes together and take n to be the total number of elements in the training set.

Test your classifier in the data set described in the Exercise 2. Compare the error results with k-NN.

4 Discriminant Functions for the Normal Density (4 Points)

Compute the discriminate function per hand and with MATLAB for the following two sets:

$$A = [1\ 2\ 2\ 3;\ 2\ 1\ 3\ 1];$$
$$\mathrm{B} = [5\ 6\ 4;2\ 3\ 4];$$

When using MATLAB, plot the test set and the discriminate function within one plot. You can use `gscatter` and `ezplot` to visualize the result. Add the mean values of the two classes and their connection line and discuss the result. Compare the MATLAB result with your findings per hand. You can use the MATLAB function `classify` to get the discriminate function. Compare the results when using the identity matrix and the general case when using individual covariance matrices. You should hand in your calculations in handwritten form by scanning pages of your results and adding them into the PDF report.

5 Outcome of this exercise

By the end of this exercise, you should be able to explain the following to someone:

- k-NN and cross validation,
- Mahalanobis distance classification,
- using the training and the test dataset, and
- the evaluation of the pattern recognition algorithms.