Statistical data analysis
Linear Discriminant Analysis

WS 2016/2017

Introduction

The aim of this tutorial is to get familiar with Linear Discriminant Analysis (LDA). LDA and Principal Component Analysis (PCA) are two techniques for dimensionality reduction. PCA can be decribed as an unsupervised algorithm that ignores data labels and aims to find directions which maximalize the variance in a data. In comparison with PCA, LDA is a supervised algorithm and aims to project a dataset onto a lower dimensional space with good class separability. In other words, LDA maximalizes the ratio of between-class variance and the within-class variance in a given data.

1 Input data

In this tutorial, we will work with a dataset that classifies wines (samples) into three classes using of 13 continuous attributes; for more details see *wine_info.txt* file. The dataset is located at wine.csv.

2 Linear Discriminant Analysis

As we mentioned above, LDA finds directions where classes are well-separated, i.e. LDA maximizes the ratio of between-class variance and the within-class variance. Firstly, assume that C is a set of classes and set D, which represents a training dataset, is defined as $D = \{x_1, x_2, \ldots, x_N\}$.

The **between-classes scatter matrix** S_B is defined as:

$$S_B = \sum_c N_C(\mu_c - \bar{x})(\mu_c - \bar{x})^T$$

where $\bar{x}$ is a vector represents the overall mean of the data, μ represents the mean corresponding to each class, and N_C are sizes of the respective classes.

The **within-classes scatter matrix** S_W is defined as:

$$S_W = \sum_c \sum_{x \in D_c} (x - \bar{\mu_c})(x - \bar{\mu_c})^T.$$

Next, we will solve the **generalized eigenvalue problem** for the matrix $S_W^{-1} S_B$ to obtain the linear discriminants, i.e.

$$(S_W^{-1} S_B)w = \lambda w$$

where w represents an eigenvector and λ represents an eigenvalue. Finally, choose k eigenvectors with the largest eigenvalue and transform the samples onto the new subspace.

3 Step by Step

You should go through the following steps:

1. Load the dataset.
2. Compute the within-scatter matrix (*withinScatterMatrix* function).
3. Compute the between-scatter matrix (*ComputeBetweenScatter* function).
4. Solve the eigenproblem (*SolveEigenProblem* function).
5. Project your data into lower-dimensional subspace, visualize this projection, and compare with PCA (see Fig. 1). Also, try to use scale/unscale version of *prcomp* function in R.
6. Discuss given results.

References

[1] Welling, Max.: *Fisher linear discriminant analysis.*," Department of Computer Science, University of Toronto 3 (2005): 1-4.

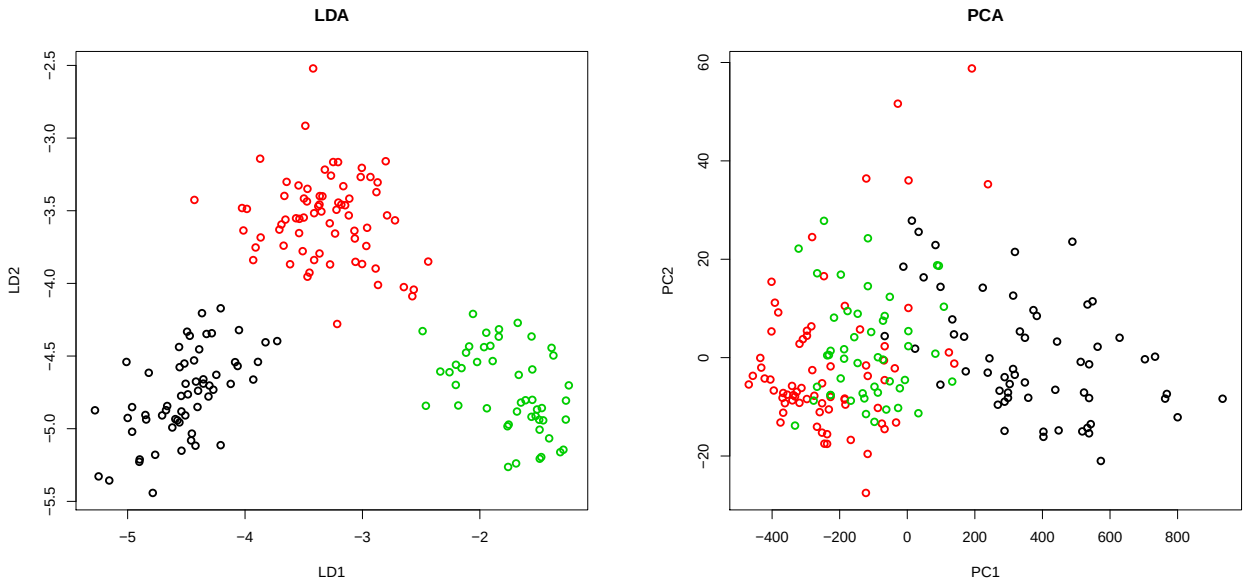

Figure 1: Linear Discriminant Analysis and Principal Component Analysis on the wine dataset.