Discriminant Analysis

Hans-Peter Helfrich

University of Bonn

Theodor Brinkmann Graduate School

Overview

Classification [Hastie et al., 2009]

General problem

Suppose we have a set of variables $x_1, \ldots, x_p$, and each observation of these variables is assigned to a certain class. We want to make a prediction of the class based on these variables.

Examples

- Prediction of the risk of a heart attack of a person based on the variables, body mass, weight, height and other indicators
- Prediction of health of a plant based on spectral signatures
- Cancer class prediction based on genetic expression measures by microarrays [Efron, 2009]
- Classification of buildings based of roof forms, window form and sizes and so on

Discriminant analysis

Methods

- Linear discriminant analysis
- Quadratic discriminant analysis
- Logistic regression
- Support vector machines

All methods deliver a *decision rule*, which depends on several parameters.

General procedure

The parameters of the decision rule are estimated by a *training set*. For each data vector in this set, the class must be known. The size of the training set should be chosen in relation to the numbers of parameters of the decision rule.

Training set

Data

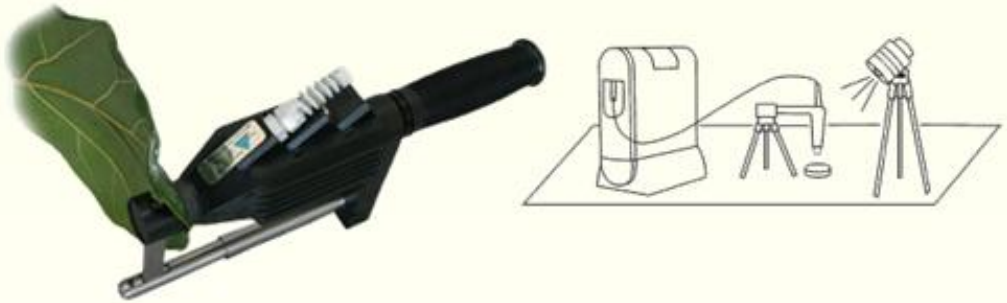
For the training set, we use N data sets. Each set consists of p (numeric) values and one assigned class. The number of classes should be so small such that for each class more data sets than parameters in the decision rule are available.

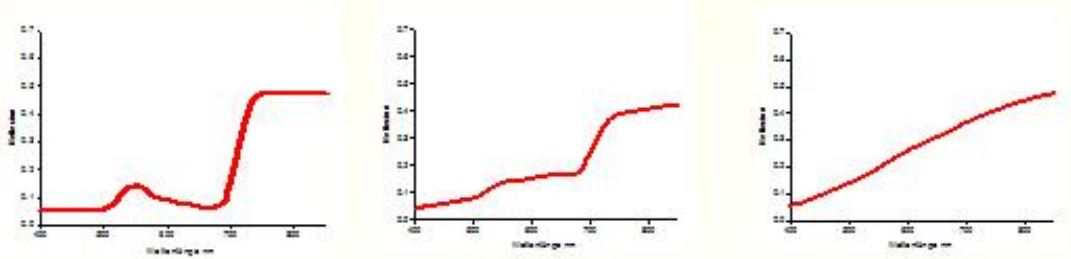
				Class
x_{11}	x_{12}	$\cdots$	x_{1p}	A1
x_{21}	x_{22}	$\cdots$	x_{2p}	A2
$\cdots$				
x_{N1}	x_{N2}	$\cdots$	x_{Np}	AN

Classical methods (suitable for $p \ll N$)

- Linear discriminant analysis
- Quadratic discriminant analysis

Example: Signatures of wheat diseases





CROPSENSe

2

By courtesy of Dr. Kai Schmidt

Coefficients of Weibull expansions

Reflection signatures

Kai Schmidt developed a method to analyze spectral and hyperspectral reflection signatures based on double Weibull functions (Registration no 10 2009 0404 944.0 at the German Patent- and Trade Mark Office)

Plant diseases

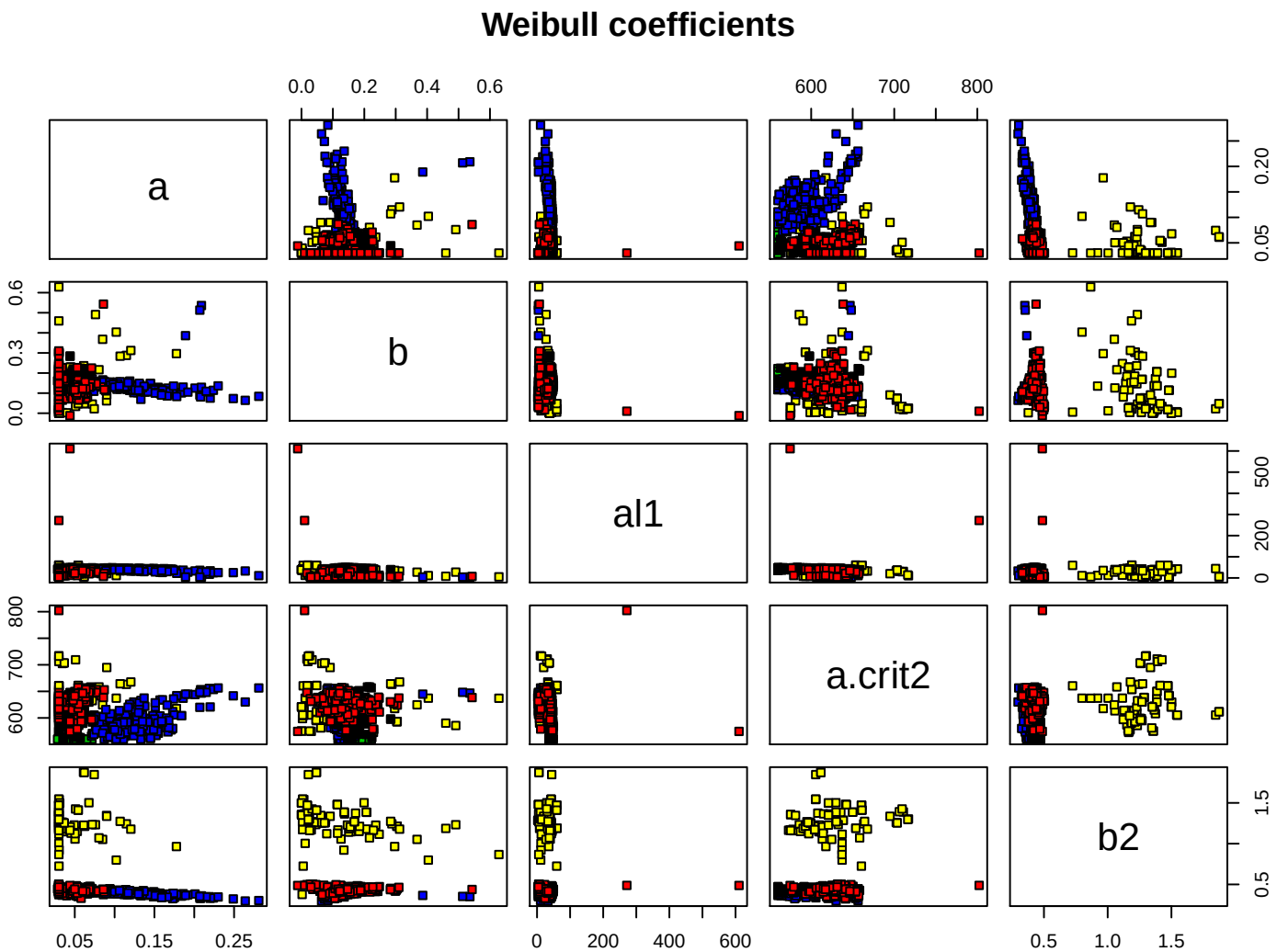
In order to classify plant diseases, Kai Schmidt may uses spectral and hyperspectral signatures of plants with $p = 500$ or more wave lengths.

Weibull expansion

Each signature can be approximated by a parametrized function of Weibull-Type

$$F(\lambda) = A + \sum_{i=1}^{k} B_i \left(1 - e^{-\left(\frac{\lambda}{a_i}\right)^{\alpha_i}}\right) e^{-\left(\frac{\lambda}{b_i}\right)^{\beta_i}}$$

Scatterplot of five Weibull coefficients



Every plant disease is represented by a color

Data: Weibull coefficients

a	b	al1	a.crit2	b2	disease
0.07160833	0.1162924	47.92688	560.000	0.4308042	Control
0.04400000	0.1680000	46.28200	579.915	0.4250000	Rost
0.10000000	0.1300000	39.25000	577.754	0.4100000	Mehltau
0.03000000	0.1380000	6.30800	630.717	0.4430000	Cercospora
0.04600000	0.1180000	12.09500	641.774	0.4190000	Cercospora
0.06100000	0.1730000	43.63100	609.100	0.4610000	Rost
...					

The whole data set consists of $N = 927$ vectors with $p = 5$ elements. Six data vectors are shown. A class is assigned to each data vector. A subset of 200 vectors is taken as training set.

Linear and quadratic discriminant analysis

Decision rules based on probabilities

Assume that the prior probability of class k is π_k with

$$\sum_{j=1}^{m} \pi_j = 1$$

If we do not have prior information, we choose an equal probability for each class, i.e, $\pi_k = 1/m$.

Selection of a class

The likelihood is chosen according to a density function $f_k(x)$. By Bayes theorem, the posterior probability to be allocated to class k based on the observation x (e.g., Weibull coefficients).

$$P(k|x) = \frac{f_k(x)\pi_k}{\sum_{j=1}^{m} f_j(x)\pi_j}$$

Density function

Linear and quadratic discriminant analysis are based on a stochastic model where it is assumed that for each class the data have a multivariate normal density distribution given by

$$f_k(x) = \frac{1}{(2\pi)^{n/2}\sqrt{\det C_k}} \exp \frac{-(x-\mu_k)^T C_k^{-1}(x-\mu_k)}{2}.$$

C_k denotes the covariance matrix for class k,

Covariance matrices

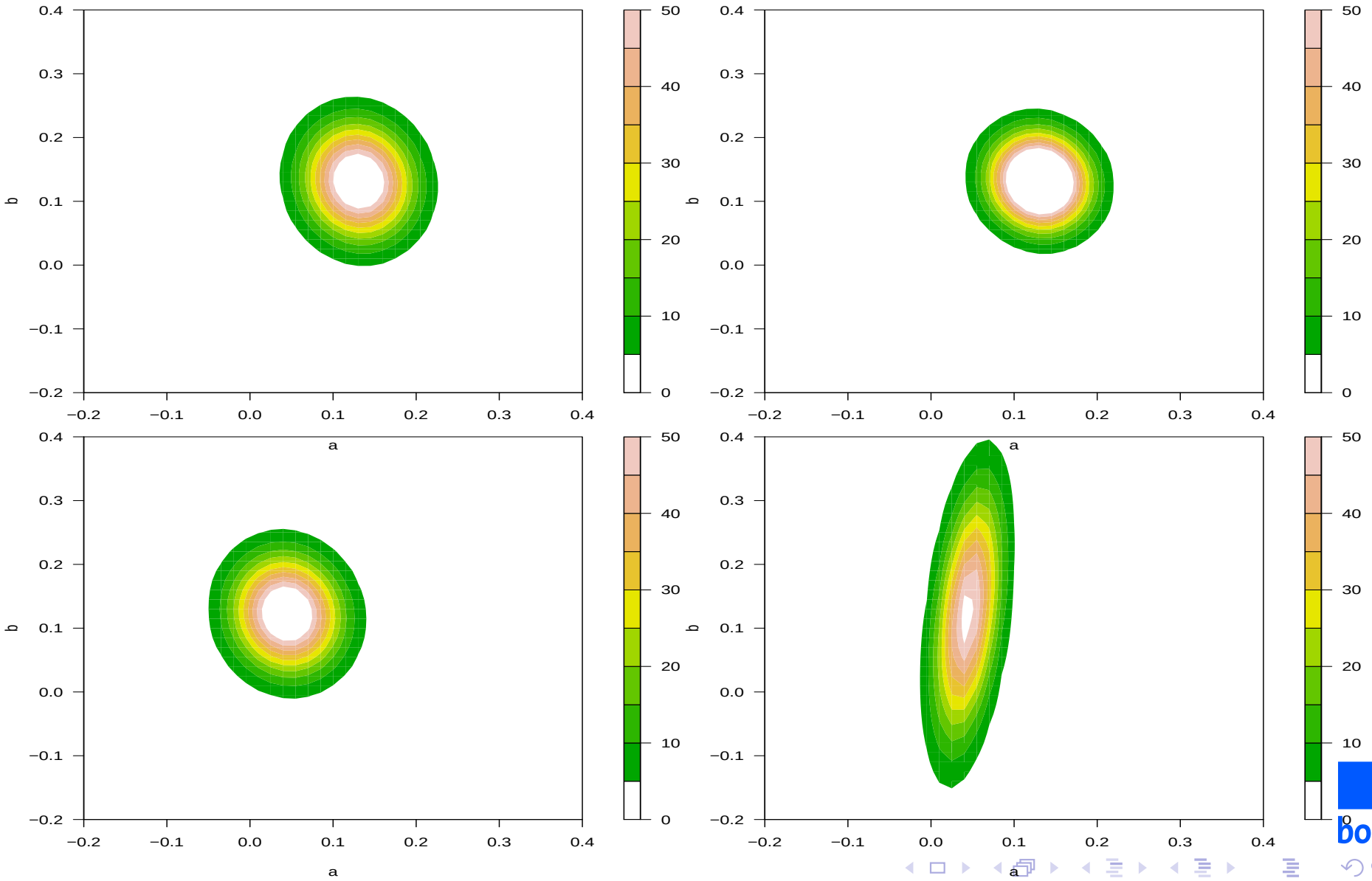
The mean values μ_k are estimated for each class separately. For the covariance matrices, we consider two cases

- Linear discriminant analysis. Only one covariance matrix C for the whole data set is estimated
- Quadratic discriminant analysis. For each class, a covariance matrix C_k is estimated.

Linear and quadratic discriminant analysis

Linear model

Quadratic model



Linear discriminant analysis

The decision function has the form

$$f_k(x) \quad \propto \quad \exp\left(\frac{-(x-\mu_k)^T C^{-1}(x-\mu_k)}{2} + \ln \pi_k\right)$$

The boundary between classes k and l is given by

$$\frac{f_k(x)}{f_l(x)} = 1$$

Boundary between two classes

$$(\mu_k - \mu_l)^T C^{-1} x = \frac{1}{2}\mu_k C^{-1}\mu_k - \frac{1}{2}\mu_l C^{-1}\mu_l + \ln\frac{\pi_k}{\pi_l}$$

The classes are separated by *hyperplanes* (lines in two dimensions, planes in three dimensions).

Linear vs. quadratic discriminant analysis

In the quadratic case, we have for each class a covariance matrix with $n(n+1)/2$ parameters and n parameters for each mean value. In our example, we have 100 parameters, whereas in the linear case only 40 parameters are need. The size of the trainig set is 200, so the fitting is good in both cases and the quadratic discriminant analysis provides better results.

Overfitting

If there are many parameters the fit of the covariance matrices and the mean values may be very good, however, the prediction results may be very poor. This phenomen occurs also in other situations like regression, it is called *overfitting*

R program

```
data <- read.table(file, header = T)
# Selection of variables for classification
spp <- c("a","b", "al1","a.crit2", "b2")

# Random choice of training set
size  <- 200 # sample size of training set
trset <- sample(1:nrow(data),size) # choice of training set
train <-  data[trset, ]

# Quadratic discriminance analysis
zd <- qda(train[,spp], train[,"disease"])
pp <- predict(zd,data[,spp])
lpred <- pp$class
lreal <- data[,"disease"]
```

Results $n = 927$, training set with 200 data

Quadratic Discriminant Analysis

Type	Hit rate	Number
Cercospora	91.87 %	209
Control	97.62 %	210
Mehltau	98.56 %	209
Rhizoctonia	98.89 %	90
Rost	98.09 %	209

Linear Discriminant Analysis

Type	Hit rate	Number
Cercospora	89.00 %	209
Control	100.00 %	210
Mehltau	87.56 %	209
Rhizoctonia	97.78 %	90
Rost	90.43 %	209

Logistic regression

The posteriori probalitities of m classes are modelled via linear functions. The probabilities should between 0 and 1 and they should sum to 1. The model has the form

$$\begin{aligned}
\ln \frac{p(1|x)}{p(m|x)} &= \beta_{10} + \beta_{11}x_1 + \cdots + \beta_{1n}x_n \\
\ln \frac{p(2|x)}{p(m|x)} &= \beta_{20} + \beta_{21}x_1 + \cdots + \beta_{2n}x_n \\
&\vdots \\
\ln \frac{p(m-1|x)}{p(m|x)} &= \beta_{m-1,0} + \beta_{m-1,1}x_1 + \cdots + \beta_{m-1,n}x_n
\end{aligned}$$

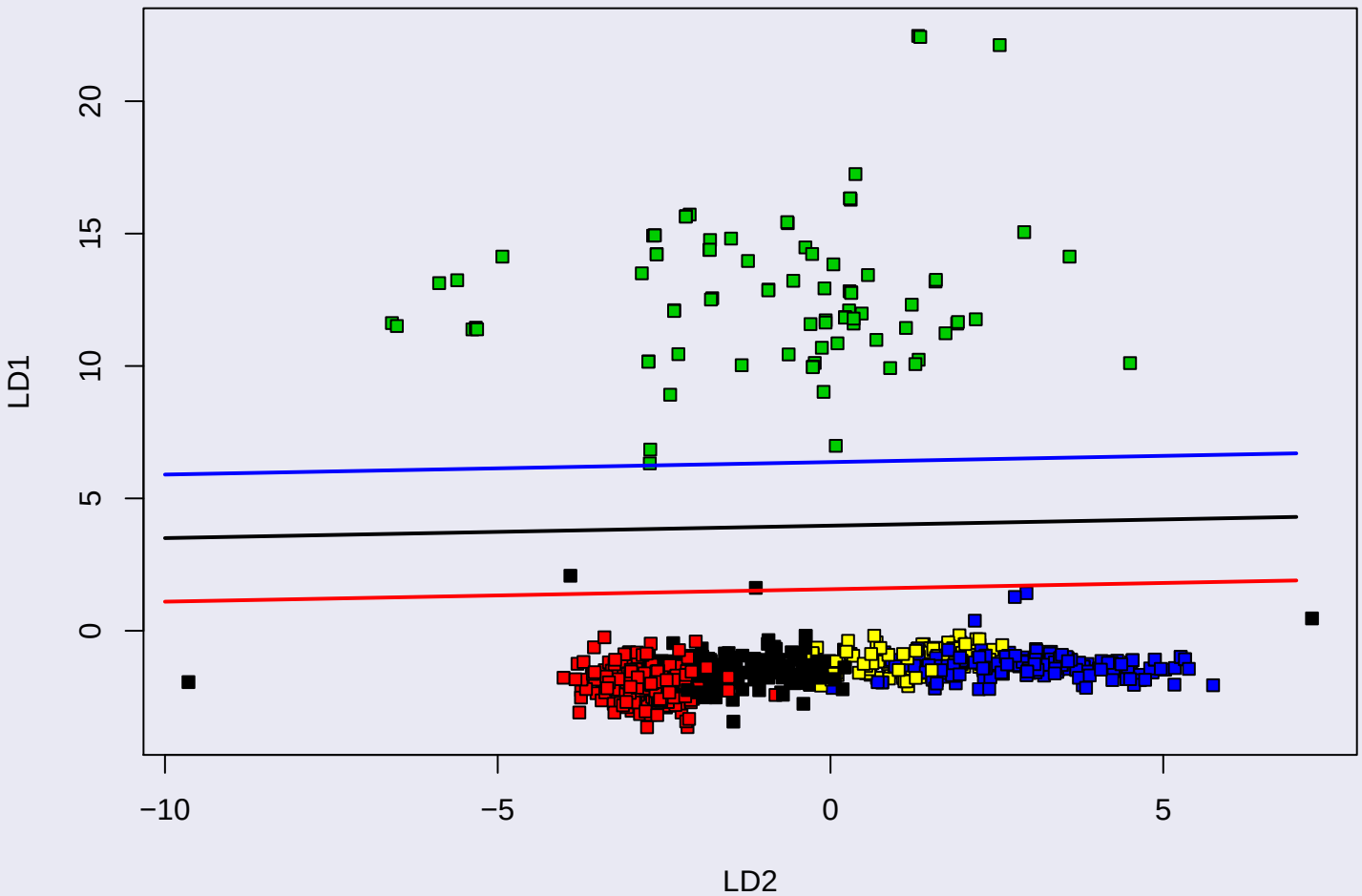
Biostatistical applications

For $m = 2$, the model is often used where binary responses (two classes) occur frequently. For example, patients survive or die, have a heart disease or not.

The support vector classifier

For two classes given, find a hyperplane that separates the two classes.

Main idea: Find the hyperplane that generates the biggest margin between the two classes.



Overlapping regions

In the case that the classes overlap, *slack variables* are introduced that measure the distance of points on the wrong side from the separating planes.

Enlarging the dimension

Hyperplanes are represented by

$$f(x) = c_1 x_1 + \cdots + c_n x_n = b$$

One may consider other nonlinear basis functions

$$h_m(x) = h_m(x_1, \ldots, x_n)$$

and take

$$h(x) = c_1 h_1(x) + \cdots + c_n h_n(x)$$

as classification rule.

Support vector machines

Support vector machines are a class of methods which do not rely on statistical methods. In the simplest case, hyperplanes are used for separating the classes.

Advantages

- Very flexible
- Basis functions can be adapted to many different problems
- Nonparametric methods are also possible

Disadvantages

- Overfitting. Multidimensionaly may give an excellent fitting of the training set, however, the prediction may be very poor
- Uncertainity cannot be easily quantified

References


Efron, B. (2009).
Empirical bayes estimates for large-scale prediction problems.
Journal of the American Statistical Association, 104(487):1015–1028.


Hastie, T., Tibshirani, R., and Friedman, J. (2009).
The elements of statistical learning.
Springer Series in Statistics. Springer-Verlag, New York, second edition.
Data mining, inference, and prediction.