Machine Learning II

Prof. Dr. Tim Downie

Week 4 Lecture – 27th October 2023

Classification: Linear and Quadratic discriminant analysis

BHT Berliner Hochschule für Technik

Contents

- Bayes Classifiers
- Linear discriminant analysis
- Quadratic discriminant analysis

Bayes Classifier

Suppose we have a classification problem with K possible classifications: $y_1, y_2, \ldots y_K$.
Y is our discrete or nominal outcome variable taking one of these values and depends on the predictor variables $x_1, \ldots x_p$.

A **Bayes Classifier** is a theoretical ideal. It assigns Y to the class y_j given the values of $x_1, \ldots x_p$ according to the decision rule: Choose the group y_j which maximises the posterior probability

$$P(Y{=}y_k|x_1, \ldots x_p) \qquad \text{for all } k{=}1, \ldots K$$

If $K{=}2$ this simplifies to the rule, assign Y to y_2 if

$$P(Y{=}y_2|x_1, \ldots x_p) > 0.5$$

Bayes Error Rate

Assuming $p=1$ and a fixed $x=x_0$.
The probability that Y is assigned to the wrong class is

$$1 - \max_k P(Y=y_k|x_0).$$

The overall error rate depends on how frequently x takes the value x_0 and all other possible values.

If X corresponds to adult heights, then $x=170$cm occurs more often than $x=200$cm.
It is more important to get the classification right at $x=170$ than at $x=200$.

The Bayesian approach considers X as a random variable with probability density function (pdf) $f_X(x)$. In the height example $f_X(170) > f_X(200)$

Revision: Continuous random variables

The pdf of a continuous r.v. X has integral 1:

$$\int_{\mathbb{R}} f_X(x)dx=1,$$

and expectation

$$E(X)=\int_{\mathbb{R}} x f_X(x)dx.$$

The expectation of a function h of the r.v. X is defined as

$$E(h(X))=\int_{\mathbb{R}} h(x)f_X(x)dx$$

The most common example is $h(x)=x^2$

$$E(X^2)=\int_{\mathbb{R}} x^2 f_X(x)dx$$

The **Bayes error rate** is the integral of the specific error rates weighted by the density of X.

$$1 - \int_{\mathbb{R}} \max_k P(Y{=}y_k|x) f_X(x)\, dx.$$

Treating $\max_k P(Y{=}y_k|x) := h(x)$ as a function of x this is often written as

$$1 - E(\max_k P((Y{=}y_k|X))$$

This measures how often we can expect to make a wrong classification.

Under the special condition that the density of X and $P(Y|X)$ are known, it can be shown that the Bayes Classifier is the best classifier we can hope to achieve and has the smallest possible Bayes error rate.

Not only is this quite a strict condition, the Bayes classifier is not a constructive algorithm.

Linear discriminant analysis (LDA)

LDA is an algorithmic approach to replicating the Bayes classifier assuming the following model.

We assume that the density of x within a given group k follows a normal distribution.

In ML 1 we considered classification on a small data set with 57 Students. For each student the height and sex is known.

In the context to LDA the student height X is modelled depending on their sex.
$X|\text{male} \sim N(\mu_m, \sigma^2)$ and $X|\text{female} \sim N(\mu_f, \sigma^2)$.

An assumption in LDA is that the variance σ^2 is the same in both groups.

We have an *a priori probability* that a person chosen at random is male $\pi_0 = P(male) = 0.5$ for the general population or $\pi_0 = P(male) = 0.66$ if the population consists of engineering students.

If we know a person's height, then we can do better than the prior probability. We can update the prior probability to give the posterior probability $P(\text{male}|x{=}\text{height})$ by using *Bayes' Theorem*.

Let $\pi_0 = P(Y{=}\text{male})$ be the prior probability that a person chosen at random is male.

$\pi_1 = P(Y{=}\text{male}|x)$ is the posterior probability that a person with height x is male.

$$\begin{aligned}\pi_1(x) := P(\text{male}|x) &= \frac{P(x|\text{male})P(\text{male})}{P(x|\text{male})P(\text{male}) + P(x|\text{female})P(\text{female})} \\ &= \frac{P(x|\text{male})\pi_0}{P(x|\text{male})\pi_0 + P(x|\text{female})(1-\pi_0)}\end{aligned}$$

Because $X|\text{male} \sim N(\mu_m, \sigma^2)$ we have a density function instead of a probability for

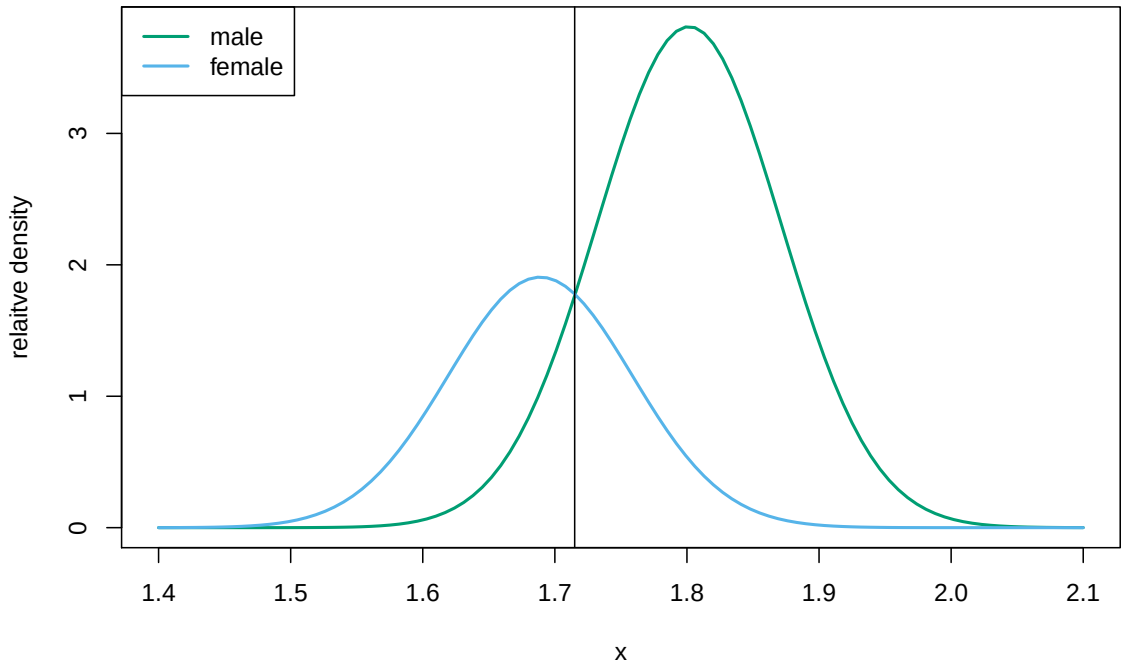
$$P(x|\text{male}) = f_{X|m}(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x-\mu_m)^2}{2\sigma^2}\right\}$$

and $X|\text{female}$ has a corresponding density.
As we assume that the variance is the same in both classes.

$$\pi_1(x) = \frac{\pi_0 f_{X|m}(x)}{\pi_0 f_{X|m}(x) + (1-\pi_0) f_{X|f}(x)}$$

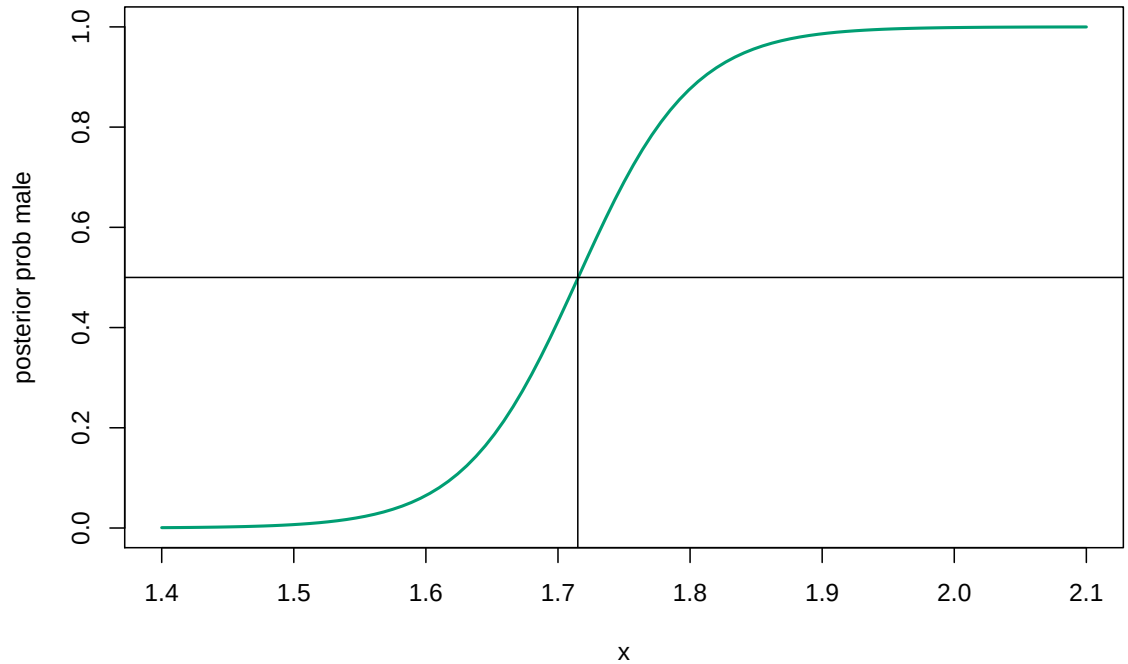
$$= \frac{\frac{\pi_0}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x-\mu_m)^2}{2\sigma^2}\right\}}{\frac{\pi_0}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x-\mu_m)^2}{2\sigma^2}\right\} + \frac{1-\pi_0}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x-\mu_f)^2}{2\sigma^2}\right\}}$$

In practice we set π_0 to be the proportion of males in the dataset and estimate μ_m, μ_f via the sample mean in each class. σ^2 is the estimated variance of mean corrected x-values, either $(x - \mu_m)$ or $(x - \mu_f)$.



For $x > 1.715$ we have $\pi_1(x) > 0.5$ so the predicted value is "male".
This is in practical terms identical to the classifier developed using logistic regression.
Logistic regression and LDA usually give similar results when there is only one predictor variable.

The peak of the green line is higher, because in this data set two-thirds of the students were male, so π_0 is set to $\frac{2}{3}$



The curve shows the posterior probability of male $\pi_1(x)$ as a function of height (x).

If the model is perfect e.g. a normal distribution fits both groups and the mean for the male group μ_m is exactly the sample mean $\overline{x}_m$ etc., then the $\pi_1(x) > 0.5$ corresponds to the Bayes classifier i.e. the best possible classifier.

More than two classes

Extending logistic regression to $K > 2$ outcomes is possible but not easy. It is called multinomial logistic regression.

Linear discriminant analysis extends easily to $K > 2$.

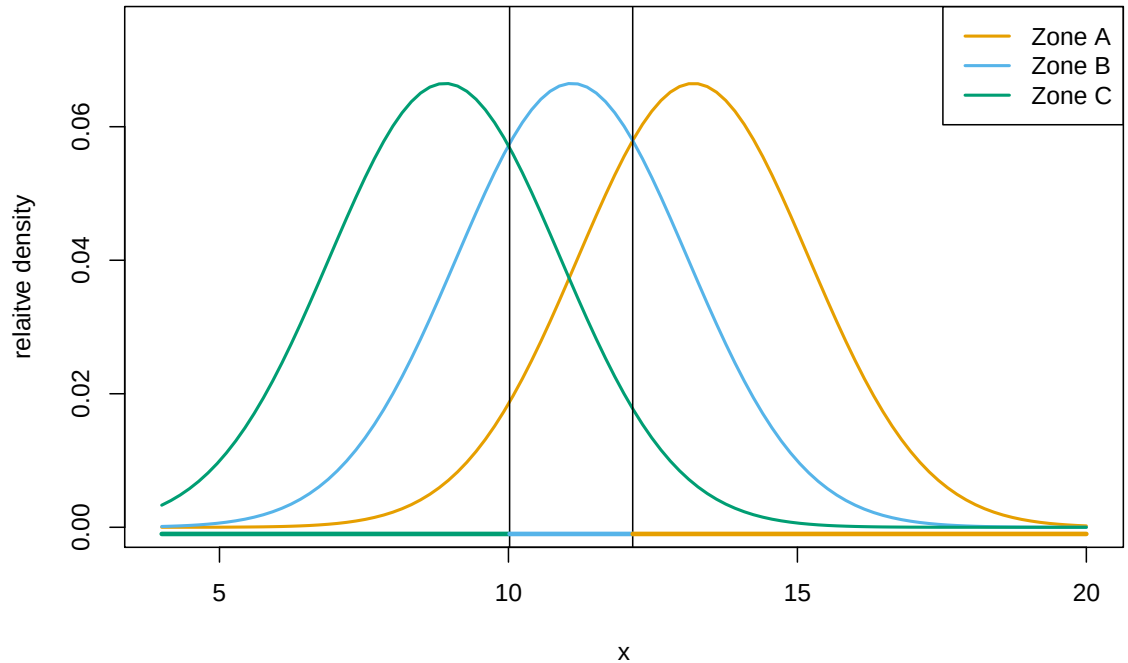
$P(Y{=}y_k|x)$ is the posterior probability our outcome variable is in group y_k given x

$$P(Y{=}y_k|x) = \frac{f_{X|y_k}(x)P(Y{=}y_k)}{\sum_{j=1}^{K} f_{X|y_j}(x|Y{=}y_j)P(Y{=}y_j)}$$

Where $P(Y{=}y_k)$ is the prior probability that our outcome variable is in group y_k.

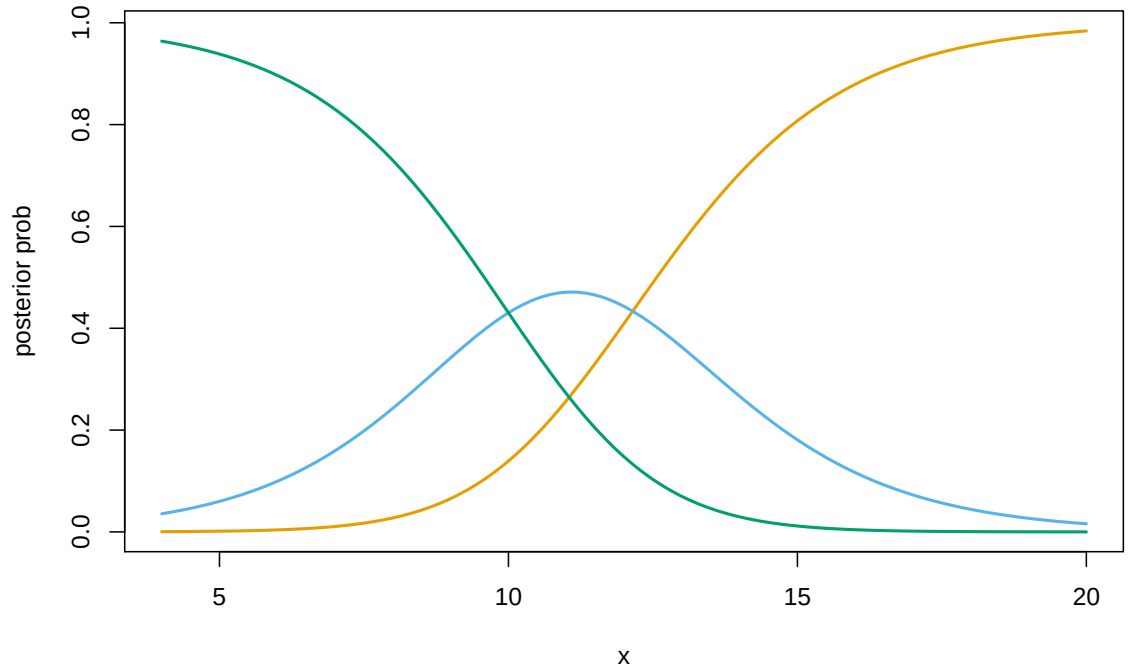
Our LDA classifier rule is to choose the value of k giving the largest $P(Y{=}y_k|x)$

Fictional example: $x=$ The basic rent per square meter for new contracts for flats rented in and around Berlin in 2017. $Y=$ BVG Zone A, B, or C.
The prior probabilities are $\frac{1}{3}$ for each zone.



The Posterior probability of being in each zone given the rent x is plotted and the classifier rule is choose the curve with the highest density for a given value of X.

Zone A	when	$x > 12.15€$
Zone B	when	$10.02€ < x < 12.15€$
Zone C	when	$x < 10.02€$



The curves correspond to $P(Y=y_k|x)$, using the same colour coding as in the previous slide.

More predictor variables

Suppose we have 2 or more predictor variables, $p \geqslant 2$.

The distribution in each group is now modelled using a multivariate normal distribution.[1]

$$(X_1, X_2, \ldots, X_p) | \, Y = y_k \sim N(\boldsymbol{\mu_k}, \Sigma^2)$$

$\boldsymbol{\mu_k}$ is a vector of length p. Σ^2 is the common variance-covariance matrix.

[1] A brief summary of the multivariate normal distribution was given in Lecture 4.

Example for LDA

There are p=2 variables K=2 and two classes for Y.

In the group Y=1, the mean of x_1 is 2.2, and the mean of x_2 is 13.5, so $\boldsymbol{\mu}_1 = \begin{pmatrix} 2.2 \\ 13.5 \end{pmatrix}$

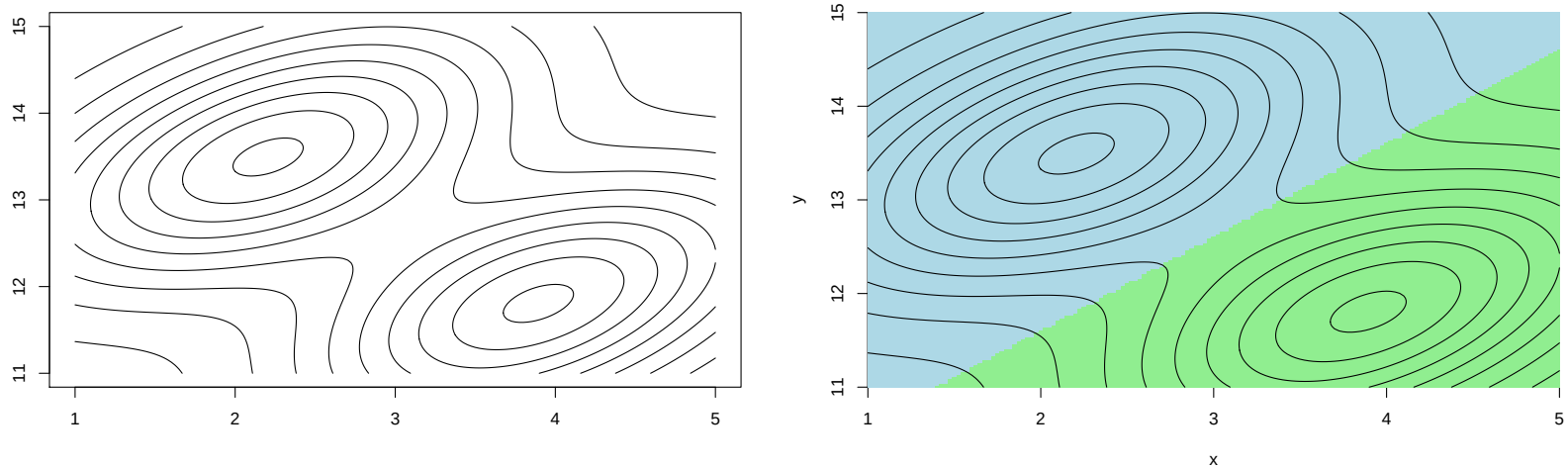
In the group Y=2, the mean vector is $\boldsymbol{\mu}_2 = \begin{pmatrix} 3.9 \\ 11.8 \end{pmatrix}$.

The variance matrix for all observations is $\boldsymbol{\Sigma} = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}$.

The prior probability that Y=1 is $\pi_0 = 0.5$.

The equivalent to the plot on slide 9, when there are two predictor variables.

For a given value of (x_1, x_2), we can calculate the density of (x_1, x_2) given $Y{=}1$ (blue) and the density of (x_1, x_2) given $Y{=}2$ (green) and plot these as a contour plot.



The blue region (right) is where the blue contours (left) are higher than the green contours.

The green region (right) is where the blue contours are lower than the green contours.

The posterior probabilities are $P(Y{=}1|x_1, x_2)$ and $P(Y{=}2|x_1, x_2)$.

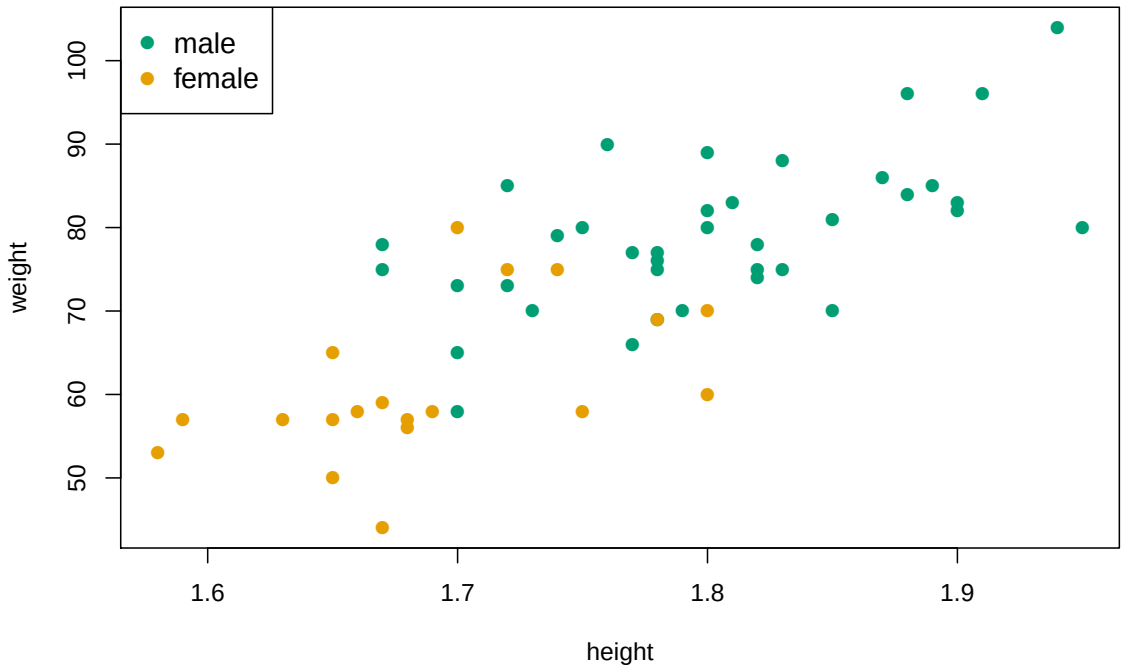
The blue region corresponds exactly to the area where $P(Y{=}1|x_1, x_2)$ is greater than $P(Y{=}2|x_1, x_2)$.

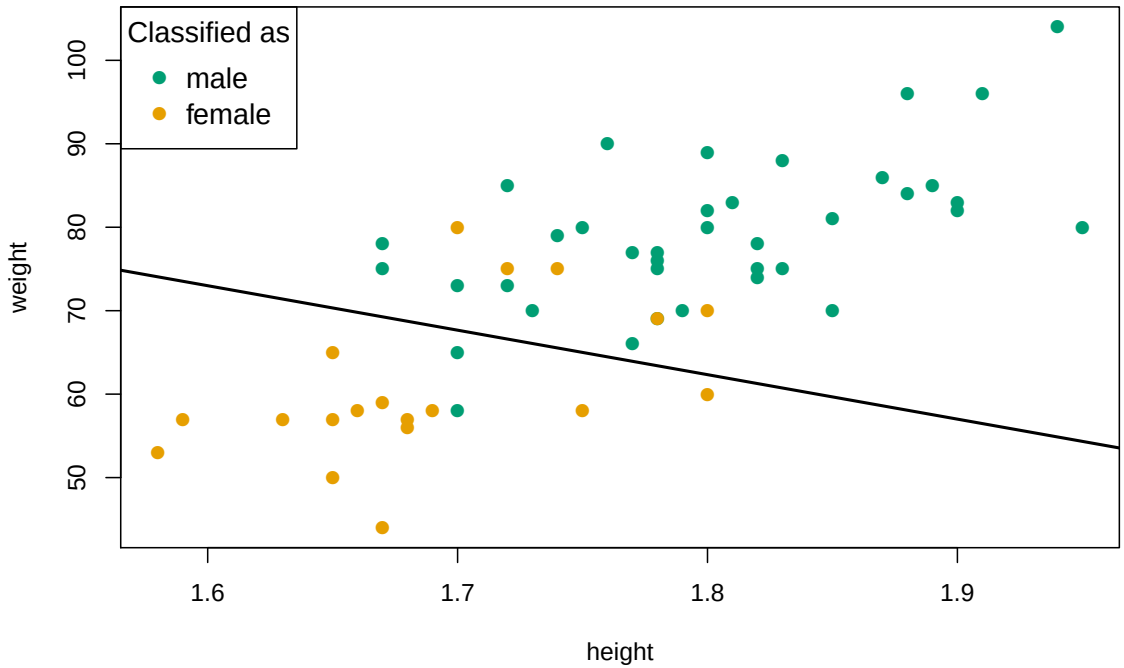
The green region corresponds exactly to the area where $P(Y{=}1|x_1, x_2)$ is less than $P(Y{=}2|x_1, x_2)$.

The boundary between the two is linear. This is why it is called linear discriminant analysis.

LDA on the student data using height and weight.

The original data





The LDA algorithm gives the following coefficients for height and weight

score=-4.56*height-0.086*weight

The predition boundary is determined by whether this score is greater than a cut off value (here -13.6), which is shown in the scatter plot.

Two males and five females are misclassified

Quadratic discriminant analysis (QDA)

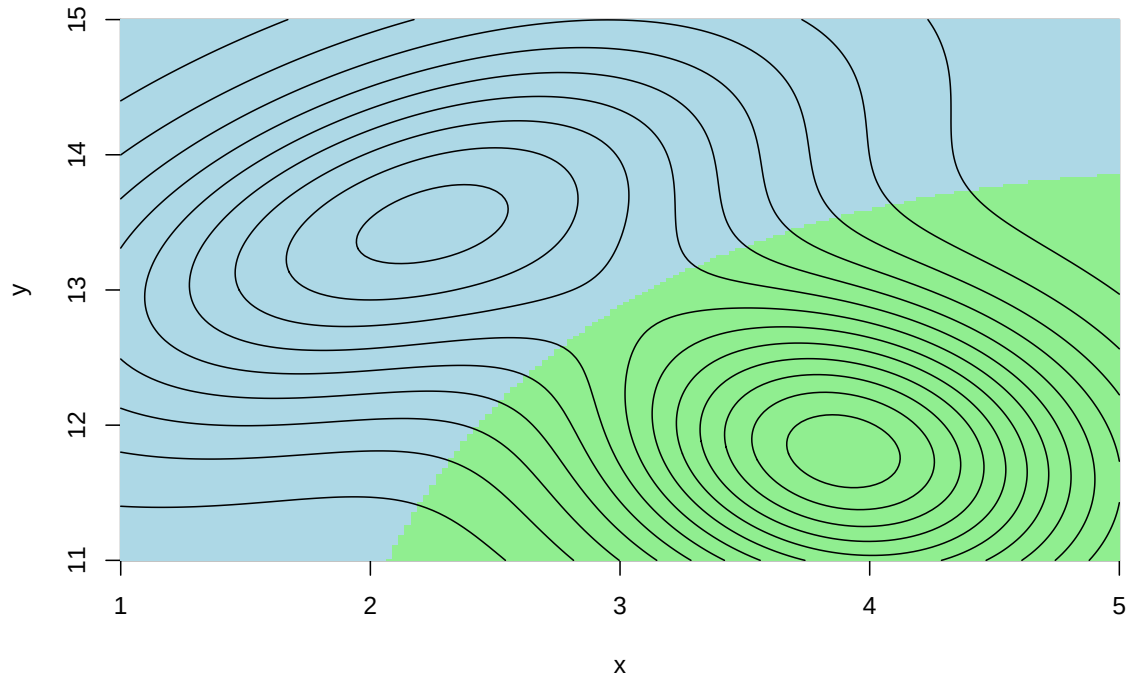
In LDA we assumed that the covariance matrix is the same in each group.

If we relax this assumption, allowing for a different covariance in each group, then we have quadratic discriminant analysis. Each class is still modelled using a using a multinomial distribution.

$$(X_1, X_2, \ldots X_p)|\, Y{=}y_k \;\sim N(\boldsymbol{\mu_k}, \Sigma_k^2)$$

$\boldsymbol{\mu_k}$ is a vector of length p for class k and Σ_k^2 is the $p \times p$ variance-covariance matrix for class k.

The boundary between the two regions is now a curve.



Note that this boundary is *not* quadratic. The log-posterior-probabilities are a 2 dimensional quadratic function. For details see James et al. Section 4.4.3.

Comparison of results using the student data

The classification matrix with 1 variable height using LDA data gives the same results as with logistic regression.

	Classified as	
	Female	Male
Actual	0	1
Female (0)	13	6
Male (1)	5	33

Sensitivity $= 0.868$, Specificity $= 0.684$.

Classification matrix for LDA using height and weight is noticeably better.

	Classified as	
	Female	Male
Actual	0	1
Female (0)	14	5
Male (1)	2	36

Sensitivity $= 0.947$, Specificity $= 0.736$.

QDA gives no further improvement over LDA .