Introduction to Statistical Learning

with applications in Python

Based on "Introduction to Statistical Learning, with applications in R" by Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani

Classification

The classification scenario, Logistic Regression, Linear Discriminant Analysis, Quadratic Discriminant Analysis

Kurt Rinnert

Physics Without Frontiers

ICTP The Abdus Salam International Centre for Theoretical Physics

UNIVERSITY OF LIVERPOOL

Abstract

In the regression scenario the response was quantitative. We now look into the classification scenario in which the response is qualitative. We will look at a few of the many available methods for classification in some detail. More methods will be covered in later lectures.

Overview

- Classification versus Regression.
- Logistic Regression.
- Multiple logistic Regression.
- Linear Discriminant Analysis (LDA).
- Quadratic Discriminant Analysis (QDA).

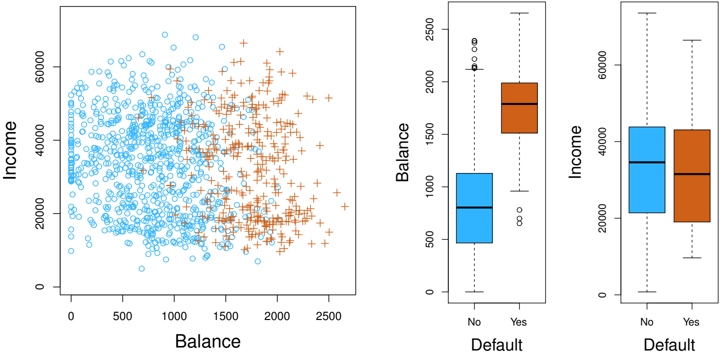
Many ideas from the regression scenario will carry over.

Examples

- Predicting a condition from symptoms in a hospital.
- Fraud detection in online payment systems.
- Predicting the probability of default on debt for credit card holders.

Classifications scenarios are very common.

Example: Default Data Set



Visualisation of the `Default` data set. The classes are color coded.

This is a simulated data set with an unusually high number of defaulters.

Example: Default Data Set

- In this example the response is *binary*:

$$y = \begin{cases} 1 & \text{if } \texttt{default} \text{ is } \texttt{Yes} \\ 0 & \text{if } \texttt{default} \text{ is } \texttt{No} \end{cases}$$

- We encode qualitative responses the same way we encode qualitative predictors.
- Linear regression would work but is not ideal.
- *Logistic regression* is the superior method.

Logistic regression predicts probabilities.

Example: Default Data Set

- For the `Default` data set we would like to predict the probability of $\texttt{default} = \texttt{Yes}$:

$$P(\texttt{default} = \texttt{Yes}|\texttt{balance})$$

- The probability is between 0 and 1.
- We can the *classify* based on P:

$$P(\texttt{default} = \texttt{Yes}|\texttt{balance}) > 0.5 \ : \ \texttt{default} = \texttt{Yes}$$

we can of course choose different *working points*.

The Logistic Model

- Out goal is to model the relationship

$$p(X) = P(Y = 1|X) \longleftrightarrow X$$

- We could use a linear regression model

$$p(X) = \beta_0 + \beta_1 X$$

- This does work but has some problems.
- In particular, the predicted probabilities can be < 0 or > 1.

We prefer a method that does not violate our axioms.

The Logistic Model

- We must model $p(X)$ such that

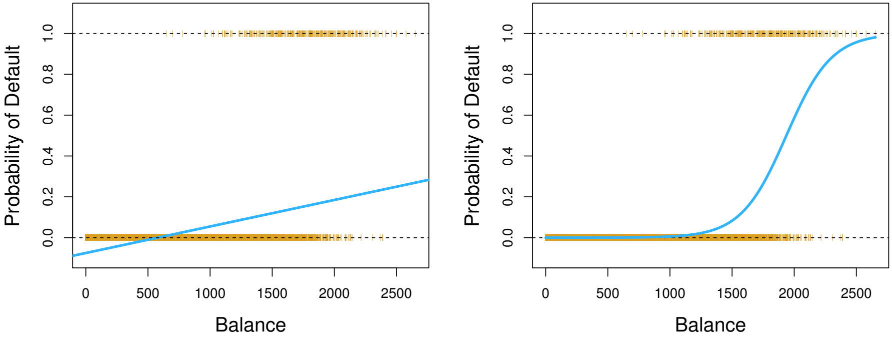
$$p(X) \in [0, 1] \quad \forall X$$

- There a many functions that guarantee that.
- We use the *logistic function*

$$p(X) = \frac{e^{\beta_0 + \beta_1 X}}{1 + e^{\beta_0 + \beta_1 X}}$$

We will need a new fitting method for this.

The Logistic Model



Left: Linear Regression, Right: Logistic Regression

The logistic model satisfies our axioms.

The Logistic Model

- The model looks complicated.
- How can we fit it?
- Some simple rearrangement yields the *odds*:

$$\frac{p(X)}{1-p(X)} = e^{\beta_0+\beta_1 X}$$

- For example:

$$p(0.2) \rightarrow \frac{1}{4} \quad \text{and} \quad p(0.9) \rightarrow 9$$

Odds originate from betting on horse races.

The Logistic Model

- How can the expression for the odds help us?
- How can we fit it?
- We take the logarithm of both sides to obtain

$$\log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X$$

- The left-hand side is called the *log odds* or *logit*.

The model for the logit is linear in *X*.

The Logistic Model

- Recall that in linear regression β_1 describes the increase of Y for a one-unit change in X.
- Here the interpretation is slightly more complicated.
- Changing X by one unit changes the *logit* by β_1.
- Or equivalently, it multiplies the odds bt e^{β_1}.
- This does *not* imply a change of β_1 of $p(X)$!
- However, any *tendency* is preserved.

The logistic model has a nice interpretation.

Maximum Likelihood

- Given

  $$\log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X$$

  we could fit the logistic model with a linear lest square fit.
- Instead, we will use a maximum likelihood fit.
- We have done this before without mentioning it.

Least Squares is just a special case of maximum likelihood.

Maximum Likelihood & Bayes' Theorem

$$P(Y|X,I) = \frac{P(X|Y,I) \times P(Y|I)}{P(X|I)}$$

Term	***Name***
$P(Y\|X,I)$	posterior probability
$P(X\|Y,I)$	likelihood
$P(Y\|I)$	prior probability
$P(X\|I)$	evidence

We are going to derive the maximum likelihood method.

Maximum Likelihood & Bayes' Theorem

- We aim to produce the best estimate of $\boldsymbol{\beta}$.
- These are the *most probable* values of β_0 and β_1 given the training data.
- That is, we seek to maximise

$$P(\boldsymbol{\beta}|X, I)$$

- A priori, we do not know how to construct this probability.

And here comes the power of Bayes' theorem…

Maximum Likelihood & Bayes' Theorem

- Given the training data *and* a model description we *can* construct the likelihood

$$P(X|\boldsymbol{\beta}, I)$$

- We can then use Bayes' theorem to obtain the posterior probability

$$P(\boldsymbol{\beta}|X, I) = \frac{P(X|\boldsymbol{\beta}, I)P(\boldsymbol{\beta}|I)}{P(X|I)}$$

- Or, up to a normalisation factor

$$P(\boldsymbol{\beta}|X, I) \propto P(X|\boldsymbol{\beta}, I)P(\boldsymbol{\beta}|I)$$

All we need is a prior and we are all set.

Maximum Likelihood & Bayes' Theorem

- We choose the prior to reflect our knowledge (or rather lack thereof) *without* taking the training data into account.
- A good start is to assume complete ignorance.
- In many cases a good ignorant prior is a flat distribution:

$$P(\boldsymbol{\beta}|I) = \text{const.}$$

The influence of the prior quickly becomes negligible for large data sets.

Maximum Likelihood & Bayes' Theorem

- The uniform (constant) prior can be absorbed in the normalisation and we obtain

$$P(\boldsymbol{\beta}|X,I) \propto \underbrace{P(X|\boldsymbol{\beta},I)}_{\text{likelihood}}$$

- In practice we often use the logarithm of the likelihood

$$\log\left(P(X|\boldsymbol{\beta},I)\right)$$

- Sometimes this allows for nice & easy analytical solutions.
- More importantly in practice it is numerically much more stable.

Now maximising the likelihood does maximise the posterior!

Maximum Likelihood & Bayes' Theorem

- Under the assumption that the x_i are independent we have

$$P(X|\boldsymbol{\beta}, I) = \prod_{i=1}^{n} P(x_i|\boldsymbol{\beta}, I)$$

- This follows from the product rule

$$P(x_i, x_k|\boldsymbol{\beta}, I) = P(x_i|x_k, \boldsymbol{\beta}, I)P(x_k|\boldsymbol{\beta}, I)$$

and the independence assumption

$$P(x_i|x_k, \boldsymbol{\beta}, I) = P(x_i|\boldsymbol{\beta}, I)$$

We still need a model, though.

The Binomial Distribution

- We need a model to describe a qualitative response variable with two classes (levels).
- The binomial distribution describes this situation

$$P(r|n,I) = \frac{n!}{n!(n-r)!} p^r (1-p)^{n-r}$$

This is the probability to observe *r* "successes".

The Likelihood Function

- We can now construct the *likelihood function* for the logistic regression:

$$\begin{aligned} P(X|\boldsymbol{\beta}, I) &= \prod_{i=1}^{n} P(x_i|\boldsymbol{\beta}, I) \\ &= \prod_{y_i=1} p(x_i) \prod_{y_i \neq 1} (1 - p(x_i)) \end{aligned}$$

with

$$p(x_i) = \frac{e^{\beta_0 + \beta_1 x_i}}{1 + e^{\beta_0 + \beta_1 x_i}}$$

Different problems require different likelihood functions.

Maximum Likelihood Estimate

- In practice we often us the logarithm of the likelihood function.
- The logarithm is a strictly monotonic function, so the extrema a preserved.
- Sometimes we minimise the negative logarithm.
- This is simply because of the abundance of minimisation libraries.

Wo do not care whether there is an analytical solution.

Logistic Regression Results

	Coefficient	Std. Error	Z-statistic	p-value
`Intercept`	-10.6513	0.3612	-29.5	< 0.0001
`balance`	0.0055	0.0002	24.9	< 0.0001

- Where the Z-statistic associated with β_1 is

$$Z = \frac{\hat{\beta}_1}{\mathrm{SE}(\hat{\beta}_1)}$$

- For large samples the Z-statistic approaches the t-statistic.

In logistic regression we use the *Z*-statistic and the associated *p*-value.

Hypothesis Testing

- The logistic null hypothesis is

$$H_0 : \beta_1 = 0 \implies p(X) = \frac{e^{\beta_0}}{1 + e^{\beta_0}}$$

- And the alternative hypothesis is

$$H_a : \beta_1 \neq 0 \implies p(X) = \frac{e^{\beta_0 + \beta_1 X}}{1 + e^{\beta_0 + \beta_1 X}}$$

We reject the null hypothesis based on a cut on the *p*-value of the *Z*-statistic.

A Word of Warning

- The term “maximum likelihood” suggests that we have found the most probable values of the parameters.
- This is in general *not* the case!
- We just determined the $\boldsymbol{\beta}$ that makes the *training data* most probable.
- It is important to keep this in mind.
- Consider the probability of rain given there are clouds versus the probability of clouds given it is raining.

In general $P(A|B) \neq P(B|A)$.

Multiple Logistic Regression

- We can generalise the logistic approach using the logit:

$$\log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_P$$

where

$$X = (X_1, \ldots, X_p)$$

are the p predictors as ususal.

We start from the logit to stress the similarity to OLS.

Multiple Logistic Regression

- We can easily translate this back to the logistic function:

$$p(x) = \frac{e^{\beta_0+\beta_1 X_1+\cdots+\beta_p Xp}}{1 + e^{\beta_0+\beta_1 X_1+\cdots+\beta_p Xp}}$$

As in the simple case we use maximum likelihood.

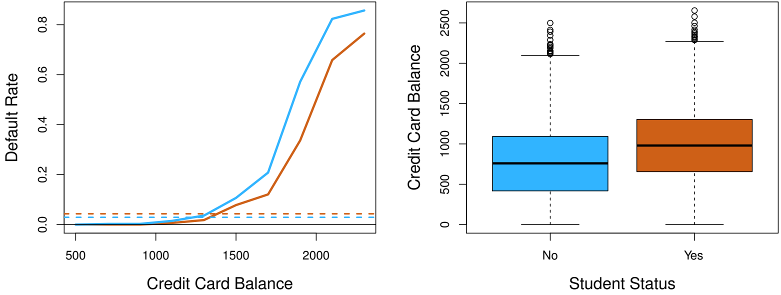
Multiple Logistic Regression

	Coefficient	Std. Error	Z-statistic	p-value
`Intercept`	−10.8690	0.4923	−22.08	< 0.0001
`balance`	0.0057	0.0002	24.75	< 0.0001
`income`	0.0030	0.0082	0.37	0.7115
`student[Yes]`	−0.6468	0.2362	−2.74	0.0062

- The negative coefficient for `student[Yes]` indicates that students are *less* likely to default.
- This holds fir *fixed values* of `balance` and `income`.

We should investigate this a bit further.

Multiple Logistic Regression



- Students (orange) and non-students (blue).
- Solid lines: $\texttt{default} = f(\texttt{balance})$.
- Dashed: overall default rate.
- Students have a higher `balance`.

There is a correlation!

The Bayes Classifier

- Suppose we have K classes with $K \geq 2$.
- The we can specify the posterior probability as

$$P(Y = k|X = x) = \frac{\pi_k f_k(x)}{\sum_{l=1}^{K} \pi_l f_l(x)}$$

where π_k is the prior and

$$f_k(x) = P(X = x|Y = k)$$

- The problem now is the specification of f_k.

The Bayes classifier is the gold standard.

Linear Discriminant Analysis

- For now we assume one predictor, that is $p = 1$
- We want to find $f_k(x)$ in order to find $p_k(x)$.
- Once we have that, we can classify observations by choosing the class with the greatest $p_k(x)$.

We need to make some further assumptions about $\boldsymbol{f_k}$.

Linear Discriminant Analysis

- We now also assume that $f_k(x)$ is *normal* (*Gaussian*).
- In the case of $p = 1$ this means

$$f_k(x) = \frac{1}{\sqrt{2\pi\sigma_k^2}} \exp\left(-\frac{1}{2\sigma_k^2}(x - \mu_k)^2\right)$$

where μ_k and σ_l may vary by class.

Even more assumptions follow…

Linear Discriminant Analysis

- For now we further assume that

$$\sigma_1 = \sigma_2 = \cdots = \sigma_K = \sigma$$

- This yields

$$p_k(x) = \frac{\pi_k \exp\left(-\frac{1}{2\sigma^2}(x-\mu_k)^2\right)}{\sum_{l=1}^{K} \pi_l \exp\left(-\frac{1}{2\sigma^2}(x-\mu_l)^2\right)}$$

We assign the observation $\boldsymbol{X = x}$ to the class with the largest $\boldsymbol{p_k(x)}$.

Linear Discriminant Analysis

- By taking the logarithm, this is equivalent to choosing the class k for which

$$\delta_k(x) = x\frac{\mu_k}{\sigma^2} - \frac{\mu_k^2}{2\sigma^2} + \log(\pi_k)$$

  is largest.

- For example, if $K = 2$ and $\pi_1 = \pi_2$, the Bayes classifier assigns:

$$\text{class} = \begin{cases} 1 & \text{if } 2x(\mu_1 - \mu_2) > \mu_1^2 - \mu_2^2 \\ 0 & \text{otherwise} \end{cases}$$

In practice we don't have access to the true parameters.

Linear Discriminant Analysis

- We have to *estimate* the parameters taking into account our assumptions.
- LDA does just that and yields

$$\hat{\mu}_k = \frac{1}{n_k} \sum_{y_i = k} x_i$$

$$\hat{\sigma}^2 = \frac{1}{n - K} \sum_{k=1}^{K} \sum_{y_i = k} (x_i - \hat{\mu}_k)^2$$

Let’s talk about the interpretation.

Linear Discriminant Analysis

- The estimate $\hat{\mu}_k$ is the average of all training observations of the kth class.
- The estimate $\hat{\sigma}^2$ is the weighted average of the sample variances of the K classes.
- If we don't know the $\pi_!, \ldots, \pi_K$ we estimate them from their frequencies in the training sample:

$$\hat{\pi}_k = \frac{n_k}{n}$$

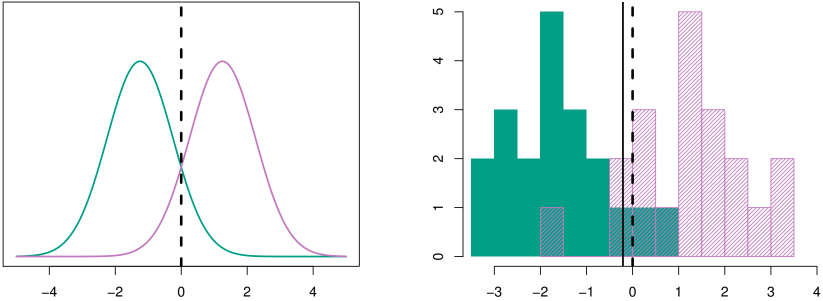
- We can now construct $\hat{\delta}(x)$:

$$\hat{\delta}_k(x) = x\frac{\hat{\mu}_k}{\hat{\sigma}^2} - \frac{\hat{\mu}_k^2}{2\hat{\sigma}^2} + \log\left(\hat{\pi}_k\right)$$

LDA is linear in the sense that $\hat{\delta}_k$ is linear in x

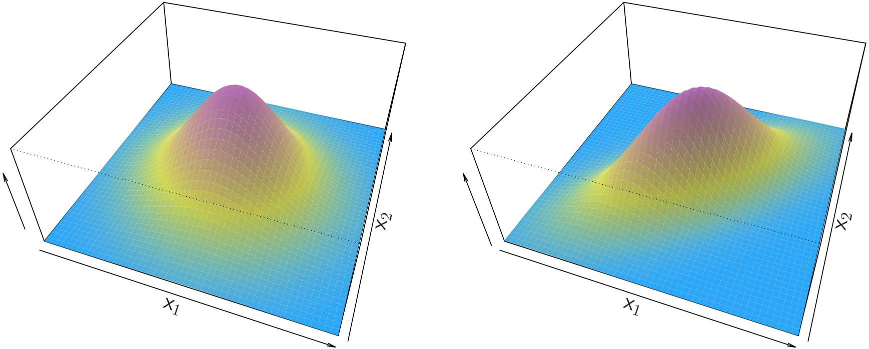
.

Linear Discriminant Analysis



The error rate is only 0.5%. LDA is doing well on this simulation.

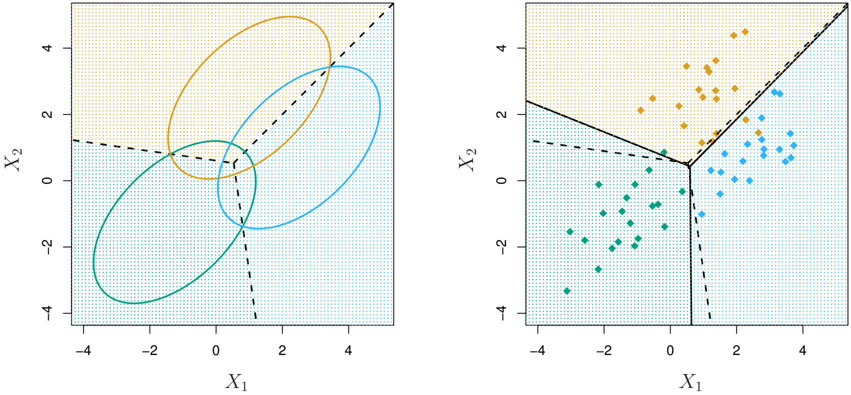
LDA with $p > 1$



- We now have a multivariate Gaussian distribution.
- We now have to estimate the covariance matrix $\boldsymbol{\Sigma}$.

Note that the variances can now be different.

LDA with $p > 1$



- Dashed: Bayes decision boundary.
- Solid: LDA estimate.

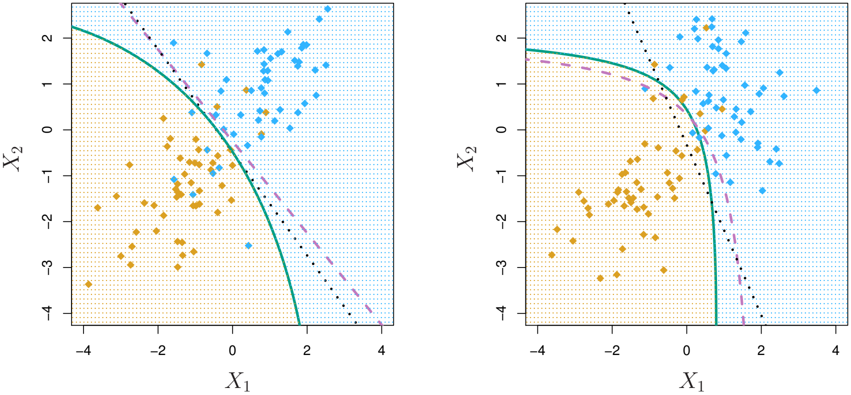
This is a simulated example with three classes.

Quadratic Discriminant Analysis

- QDA is an alternative approach to LDA that allows for curved boundaries.
- We drop the assumption that $\boldsymbol{\Sigma}$ is common to all classes.
- That means QDA provides an estimate $\hat{\boldsymbol{\Sigma}}_k$ in addition to $\hat{\mu}_k$ and $\hat{\pi}_k$.

The resulting $\hat{\boldsymbol{\delta}}_{\boldsymbol{k}}$ is now quadratic in $\boldsymbol{x}$.

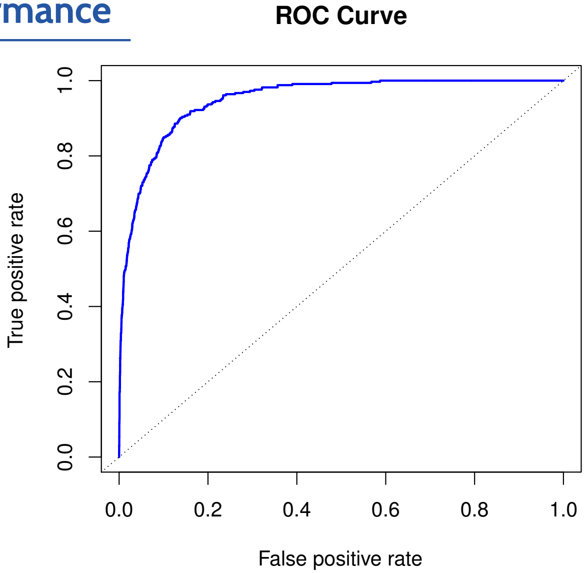
Quadratic Discriminant Analysis



- Left: $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2$
- Right: $\boldsymbol{\Sigma}_1 \neq \boldsymbol{\Sigma}_2$

The QDA performs better when the boundary is curved.

Classifier Performance



we can use this tho choose a *working point*.