

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



The Abdus Salam
International Centre
for Theoretical Physics



UNIVERSITY OF
LIVERPOOL

Copyright © 2019

Kurt Rinnert <kurt.rinnert@cern.ch>, Kate Shaw <kshaw@ictp.it>
Copying and distribution of this file, with or without modification,
are permitted in any medium without royalty provided the copyright
notice and this notice are preserved. This file is offered as-is, without
any warranty.

Some of the figures in this presentation are taken from "An Introduction
to Statistical Learning, with applications in R" (Springer, 2013)
with permission from the authors: G. James, D. Witten, T. Hastie and
R. Tibshirani

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).

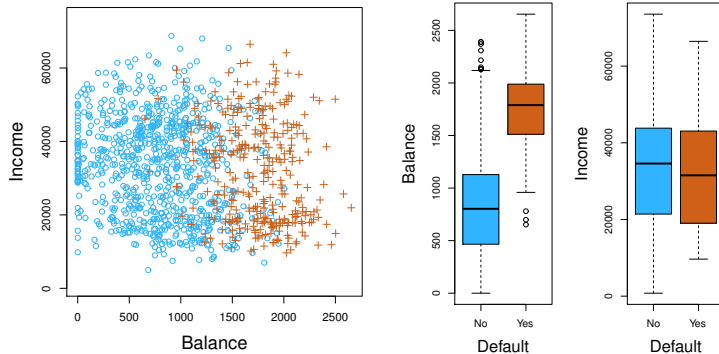
Many ideas from the regression scenario will carry over.

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*:

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

- 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 `default = Yes`:

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

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

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

we can of course choose different *working points*.

The Logistic Model

- Our goal is to model the relationship

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

- 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

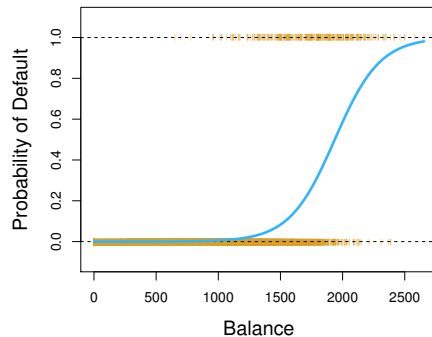
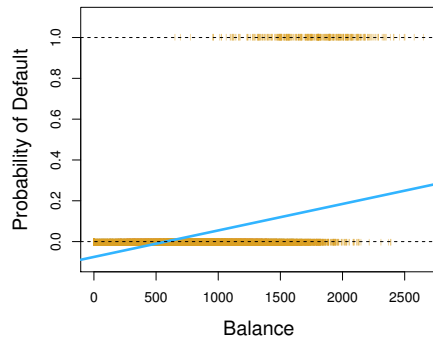
$$p(X) \in [0, 1] \quad \forall X$$

- There are 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 by 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 least 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

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

<i>Term</i>	<i>Name</i>
$P(Y X, I)$	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

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

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

We 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

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

- 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 β 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 $P(A|B) \neq P(B|A)$.