

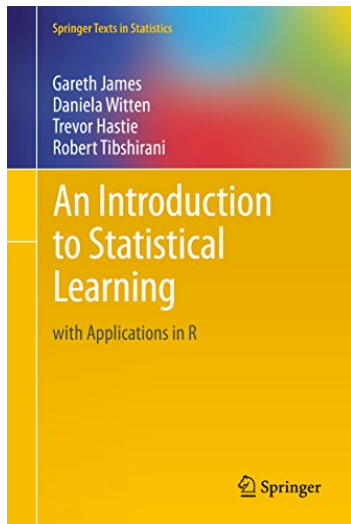
Classification

chapter 4 of *An Introduction to Statistical Learning* (ISL)

Henrique Laureano
<http://leg.ufpr.br/~henrique>



What we read (long description)



4	Classification	127
4.1	An Overview of Classification	128
4.2	Why Not Linear Regression?	129
4.3	Logistic Regression	130
4.3.1	The Logistic Model	131
4.3.2	Estimating the Regression Coefficients	133
4.3.3	Making Predictions	134
4.3.4	Multiple Logistic Regression	135
4.3.5	Logistic Regression for >2 Response Classes	137
4.4	Linear Discriminant Analysis	138
4.4.1	Using Bayes' Theorem for Classification	138
4.4.2	Linear Discriminant Analysis for $p = 1$	139
4.4.3	Linear Discriminant Analysis for $p > 1$	142
4.4.4	Quadratic Discriminant Analysis	149
4.5	A Comparison of Classification Methods	151

Now in a shorter way

What we read (short description)

At chapter 4 are discussed three of the most widely-used classifiers.

- » Logistic Regression
- » Linear Discriminant Analysis (LDA)
- » K-Nearest Neighbors (KNN)

What we didn't read

More computer-intensive methods are discussed in later chapters, such as

- » Generalized Additive Models (GAM)
- » Trees
- » Random Forests
- » Boosting
- » Support Vector Machines (SVM)

On the Agenda

1 Why Not Linear Regression?

2 A typical dataset

3 Logistic Regression

- The model framework
- Estimating the Regression Coefficients

4 Linear Discriminant Analysis (LDA)

- To start... why do we need something different?
- LDA in a nutshell
- Living in a simple and *normal* world
- Now, with more than one predictor

5 K-Nearest Neighbors (KNN)

We could consider encoding the response, Y , as a quantitative variable, e.g.,

Predict the medical condition of a patient on the basis of her symptoms.

$$Y = \begin{cases} 1 & \text{if stroke;} \\ 2 & \text{if drug overdose;} \\ 3 & \text{if epileptic seizure.} \end{cases}$$

We could consider encoding the response, Y , as a quantitative variable, e.g.,

Predict the medical condition of a patient on the basis of her symptoms.

$$Y = \begin{cases} 1 & \text{if stroke;} \\ 2 & \text{if drug overdose;} \\ 3 & \text{if epileptic seizure.} \end{cases}$$

Unfortunately, this coding implies an ordering on the outcomes.

Each possible coding would produce a fundamentally different linear model that would ultimately lead to different sets of predictions.

That leads us to other questions,

- » What if the response variable values did take on a natural ordering, such as mild, moderate, and severe?
- » For a binary (two level) qualitative response, the situation is better.
 - » However, if we use linear regression, some of our estimates might be outside the $[0, 1]$ interval.
 - » However, the dummy variable approach cannot be easily extended to accommodate qualitative responses with more than two levels.

That leads us to other questions,

- » What if the response variable values did take on a natural ordering, such as mild, moderate, and severe?
- » For a binary (two level) qualitative response, the situation is better.
 - » However, if we use linear regression, some of our estimates might be outside the $[0, 1]$ interval.
 - » However, the dummy variable approach cannot be easily extended to accommodate qualitative responses with more than two levels.

For these reasons, it is preferable to use a classification method that is truly suited for qualitative response values, such as the ones presented next.

Curiously,

it turns out that the classifications that we get if we use linear regression to predict a binary response will be the same as for the linear discriminant analysis (LDA) procedure we discuss later.

On the Agenda

- 1 Why Not Linear Regression?
- 2 A typical dataset
- 3 Logistic Regression
 - The model framework
 - Estimating the Regression Coefficients
- 4 Linear Discriminant Analysis (LDA)
 - To start... why do we need something different?
 - LDA in a nutshell
 - Living in a simple and *normal* world
 - Now, with more than one predictor
- 5 K-Nearest Neighbors (KNN)

A classic 'book example dataset relationship'

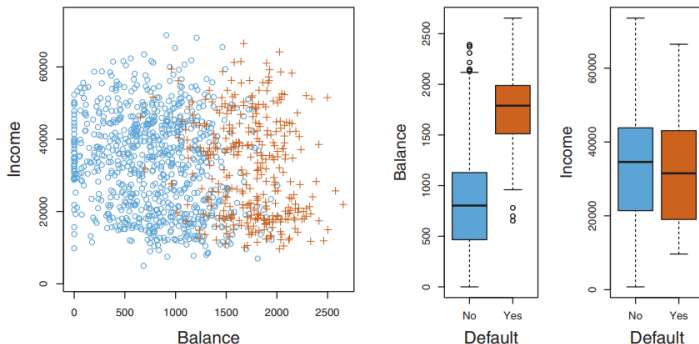


FIGURE 4.1. The **Default** data set. Left: The annual incomes and monthly credit card balances of a number of individuals. The individuals who defaulted on their credit card payments are shown in orange, and those who did not are shown in blue. Center: Boxplots of **balance** as a function of **default** status. Right: Boxplots of **income** as a function of **default** status.

... a very pronounced relationship between balance and default.

On the Agenda

- 1 Why Not Linear Regression?
- 2 A typical dataset
- 3 **Logistic Regression**
 - The model framework
 - Estimating the Regression Coefficients
- 4 Linear Discriminant Analysis (LDA)
 - To start... why do we need something different?
 - LDA in a nutshell
 - Living in a simple and *normal* world
 - Now, with more than one predictor
- 5 K-Nearest Neighbors (KNN)

To start, a comparison with Linear Regression

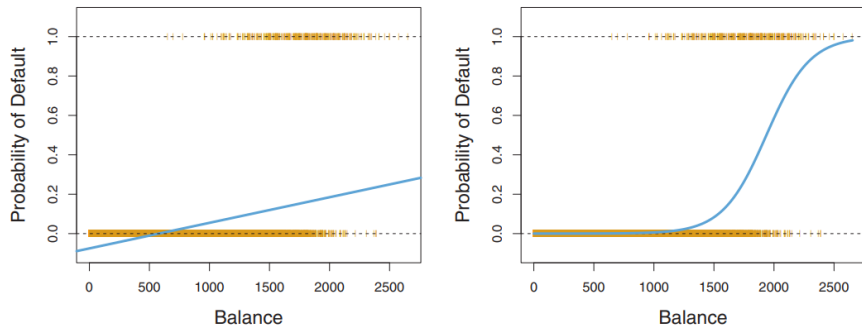


FIGURE 4.2. Classification using the **Default** data. Left: Estimated probability of **default** using linear regression. Some estimated probabilities are negative! The orange ticks indicate the 0/1 values coded for **default**(No or Yes). Right: Predicted probabilities of **default** using logistic regression. All probabilities lie between 0 and 1.

Logistic regression in two slides

Some math, but with just one predictor

The model and its relations (*showing my \LaTeX skills*)

$$p(X) = \underbrace{\frac{e^{\beta_0 + \beta_1 X}}{1 + e^{\beta_0 + \beta_1 X}}}_{\substack{\text{logistic} \\ \text{function} \\ \text{(S-shaped)}}} \Rightarrow \underbrace{\frac{p(X)}{1 - p(X)}}_{\text{odds} \in (0, \infty)} = e^{\beta_0 + \beta_1 X} \Rightarrow \log \underbrace{\frac{p(X)}{1 - p(X)}}_{\substack{\text{log-odds} \\ \text{or} \\ \text{logit}}} = \beta_0 + \beta_1 X$$

Some math, but with just one predictor

The model and its relations (*showing my \LaTeX skills*)

$$p(X) = \underbrace{\frac{e^{\beta_0 + \beta_1 X}}{1 + e^{\beta_0 + \beta_1 X}}}_{\substack{\text{logistic} \\ \text{function} \\ (S\text{-shaped})}} \Rightarrow \underbrace{\frac{p(X)}{1 - p(X)}}_{\substack{\text{odds} \in (0, \infty)}} = e^{\beta_0 + \beta_1 X} \Rightarrow \underbrace{\log \frac{p(X)}{1 - p(X)}}_{\substack{\text{log-odds} \\ \text{or} \\ \text{logit}}} = \beta_0 + \beta_1 X$$

For example,

$$p(X) = 0.2 \Rightarrow \frac{0.2}{1 - 0.2} = \frac{1}{4} \quad \text{and} \quad p(X) = 0.9 \Rightarrow \frac{0.9}{1 - 0.9} = 9.$$

Maximum likelihood

The estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ are chosen to **maximize** a math equation called a *likelihood function*

$$l(\beta_0, \beta_1) = \prod_{i:y_i=1} p(x_i) \prod_{i':y_{i'}=0} (1 - p(x_{i'})).$$

Maximum likelihood

The estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ are chosen to **maximize** a math equation called a *likelihood function*

$$l(\beta_0, \beta_1) = \prod_{i: y_i=1} p(x_i) \prod_{i': y_{i'}=0} (1 - p(x_{i'})).$$

The coefficients $\hat{\beta}_0$ and $\hat{\beta}_1$ are unknown, and must be estimated. The general method of **maximum likelihood** is preferred, since it has better statistical properties.

Maximum likelihood is a very general approach that is used to fit many of the non-linear models examined throughout the book. In the linear regression setting, the least squares approach is in fact a special case of maximum likelihood.

On the Agenda

- 1 Why Not Linear Regression?
- 2 A typical dataset
- 3 Logistic Regression
 - The model framework
 - Estimating the Regression Coefficients
- 4 Linear Discriminant Analysis (LDA)
 - To start. . . why do we need something different?
 - LDA in a nutshell
 - Living in a simple and *normal* world
 - Now, with more than one predictor
- 5 K-Nearest Neighbors (KNN)

Different ideas, sometimes the same results

Different ideas,

The image shows a handwritten comparison between two statistical models. On the left, 'Logistic REGRESSION' is written, followed by the probability expression $\mathbb{P}[Y = K | X = x]$. A red bracket underneath this expression is labeled 'via, logistic function'. In the center, 'vs.' is written in red. On the right, 'LINEAR DISCRIMINANT Analysis' is written, followed by the probability expression $\mathbb{P}[X = x | Y = K]$. A red bracket underneath this expression is labeled 'via BAYES' THEOREM'.

Logistic REGRESSION : $\mathbb{P}[Y = K | X = x]$ vs. LINEAR DISCRIMINANT Analysis : $\mathbb{P}[X = x | Y = K]$

VIA, logistic function VIA BAYES' THEOREM

With LDA we model the distribution of the predictors X separately in each of the response classes (i.e. given Y), and then use Bayes' theorem to flip these around into estimates for $\mathbb{P}[Y = k | X = x]$.

Different ideas,

Handwritten text comparing Logistic Regression and Linear Discriminant Analysis (LDA). The text is written on a light brown background with red underlines and arrows.

Logistic REGRESSION : $\mathbb{P}[Y = K | X = x]$ vs. LINEAR DISCRIMINANT ANALYSIS : $\mathbb{P}[X = x | Y = K]$

Logistic REGRESSION is connected by a red bracket to the text "VIA, logistic function".

LINEAR DISCRIMINANT ANALYSIS is connected by a red bracket to the text "VIA BAYES' THEOREM".

With LDA we model the distribution of the predictors X separately in each of the response classes (i.e. given Y), and then use Bayes' theorem to flip these around into estimates for $\mathbb{P}[Y = k | X = x]$.

Sometimes the same results

When these distributions are assumed to be normal, it turns out that the model is very similar in form to logistic regression.

But, ok... why not continue with logistic regression?

But, ok... why not continue with logistic regression?

Simple, LDA is popular when we have more than two response classes.

Now, a reason more serious: [stability](#)

- » When the classes are well-separated, the parameter estimates for the logistic regression model are surprisingly unstable. LDA does not suffer from this problem.
- » If n is small and the distribution of the predictors X is approximately normal in each of the classes, the linear discriminant model is again more stable than the logistic regression model.

Model framework

$$\underbrace{p_k(x)}_{\text{POSTERIOR}} = \mathbb{P}[Y=k | X=x] = \frac{\overbrace{\pi_k}^{\text{PRIOR}} \overbrace{f_k(x)}^{\text{DENSITY FN}}}{\sum_{L=1}^K \pi_L f_L(x)}, \text{ with } f_k(x) = \mathbb{P}[X=x | Y=k]$$

- » π_k is the overall or **prior** prob. that a chosen obs. comes from k .
- » In general, estimating π_k is easy if we have a sample of Y s: we simply compute the fraction of observations that belong to the k th class. However, estimating $f_k(x)$ tends to be more challenging, unless we assume some simple forms for these densities.

Remember from Chap. 2 that the Bayes classifier has the lowest possible error rate out of all classifiers.

Dealing with just one predictor

Assumptions: $f_k(x)$ is normal with equal variance for the k th classes.

$$p_k(x) = \frac{\pi_k f_k(x)}{\sum_{l=1}^K \pi_l f_l(x)}$$

$$\sim \mathcal{N}(\mu_k, \sigma_k^2)$$

$$\sigma_k^2 = \sigma_l^2$$

BAYES CLASSIFIER

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

with SOME SIMPLE STEPS

if $k=2$ & $\pi_1 = \pi_2$

BAYES DECISION

boundary: $x = \frac{\mu_1 + \mu_2}{2}$

Putting a **hat** (simple average and a weighted average of the sample variances for each class) in everything, the LDA **approx.** this Bayes classifier.

Ok, nice! But... **why** the name **linear discriminant analysis**?

Ok, nice! But... **why** the name **linear discriminant analysis**?

The word **linear** stems from the fact that the **discriminant functions** $\hat{\delta}_k(x)$ are linear functions of x .

Getting bigger

More than one predictor \Rightarrow Multivariate normal distribution,
with a class-specific mean vector
and a common covariance matrix

On the Agenda

- 1 Why Not Linear Regression?
- 2 A typical dataset
- 3 Logistic Regression
 - The model framework
 - Estimating the Regression Coefficients
- 4 Linear Discriminant Analysis (LDA)
 - To start... why do we need something different?
 - LDA in a nutshell
 - Living in a simple and *normal* world
 - Now, with more than one predictor
- 5 K-Nearest Neighbors (KNN)

and...



laureano@ufpr.br