

Classification

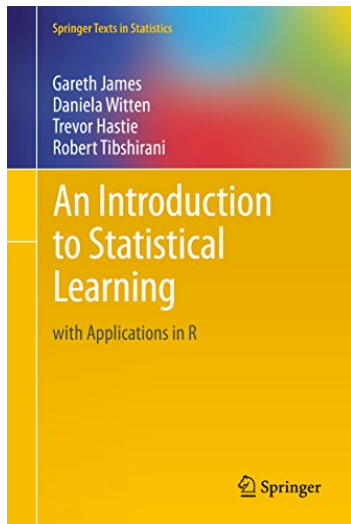
chapter 4 of *An Introduction to Statistical Learning* (ISL)

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

Laboratory of Statistics and Geoinformation (LEG)
UFPR/DEST/LEG



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

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)
- 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)
- 5 K-Nearest Neighbors (KNN)

A classic 'book example dataset relationship'

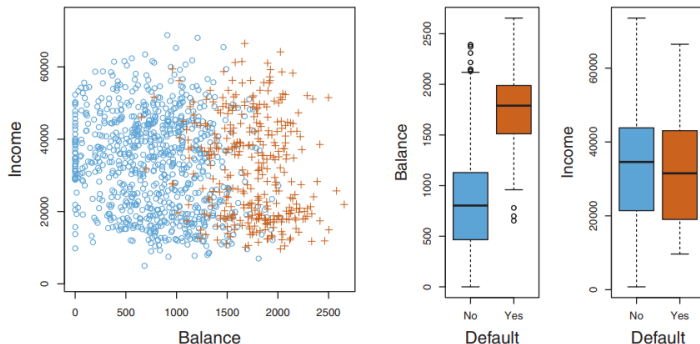


FIGURE 4.1. The **Default** data set. Left: The annual incomes and monthly credit card balances of a number of individuals. The individuals who defaulted on their credit card payments are shown in orange, and those who did not are shown in blue. Center: Boxplots of **balance** as a function of **default** 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)
- 5 K-Nearest Neighbors (KNN)

To start, a comparison with Linear Regression

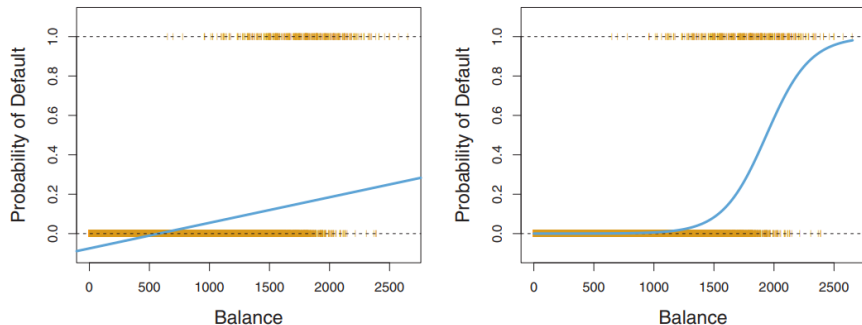


FIGURE 4.2. Classification using the `Default` data. Left: Estimated probability of `default` using linear regression. Some estimated probabilities are negative! The orange ticks indicate the 0/1 values coded for `default` (No or Yes). Right: Predicted probabilities of `default` using logistic regression.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

Some math, but with just one predictor

The model and its relations

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

Some math, but with just one predictor

The model and its relations

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

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)
- 5 K-Nearest Neighbors (KNN)

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)
- 5 K-Nearest Neighbors (KNN)

and...



laureano@ufpr.br