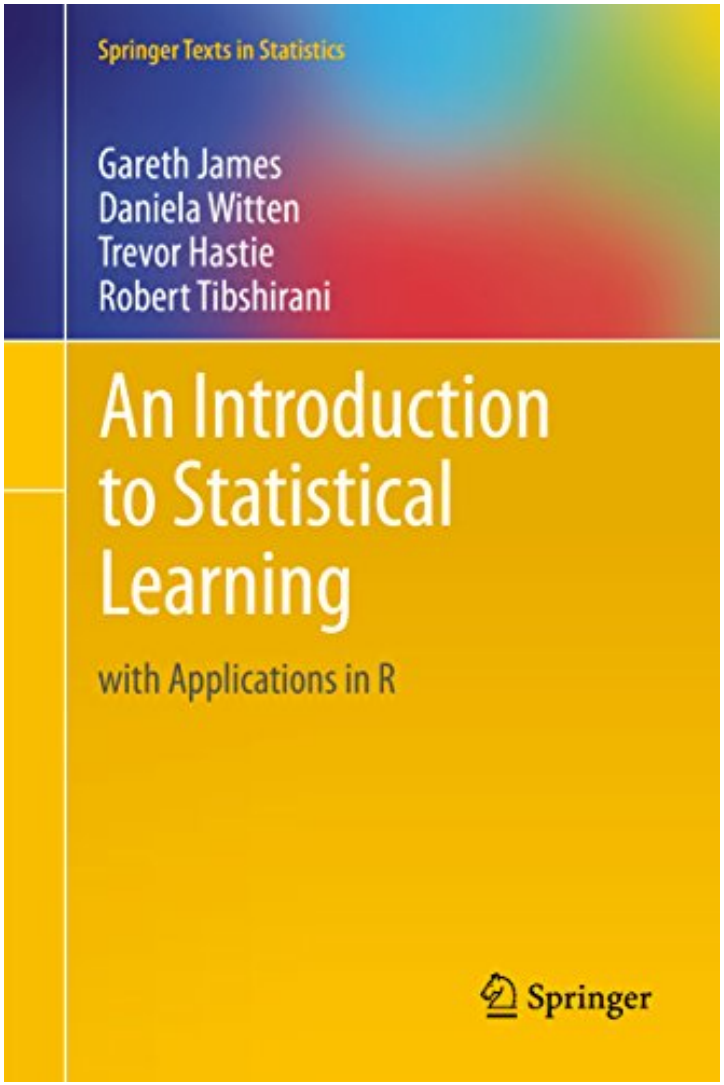
Classification

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)



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

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.

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

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.

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

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

A classic 'book example dataset relationship'

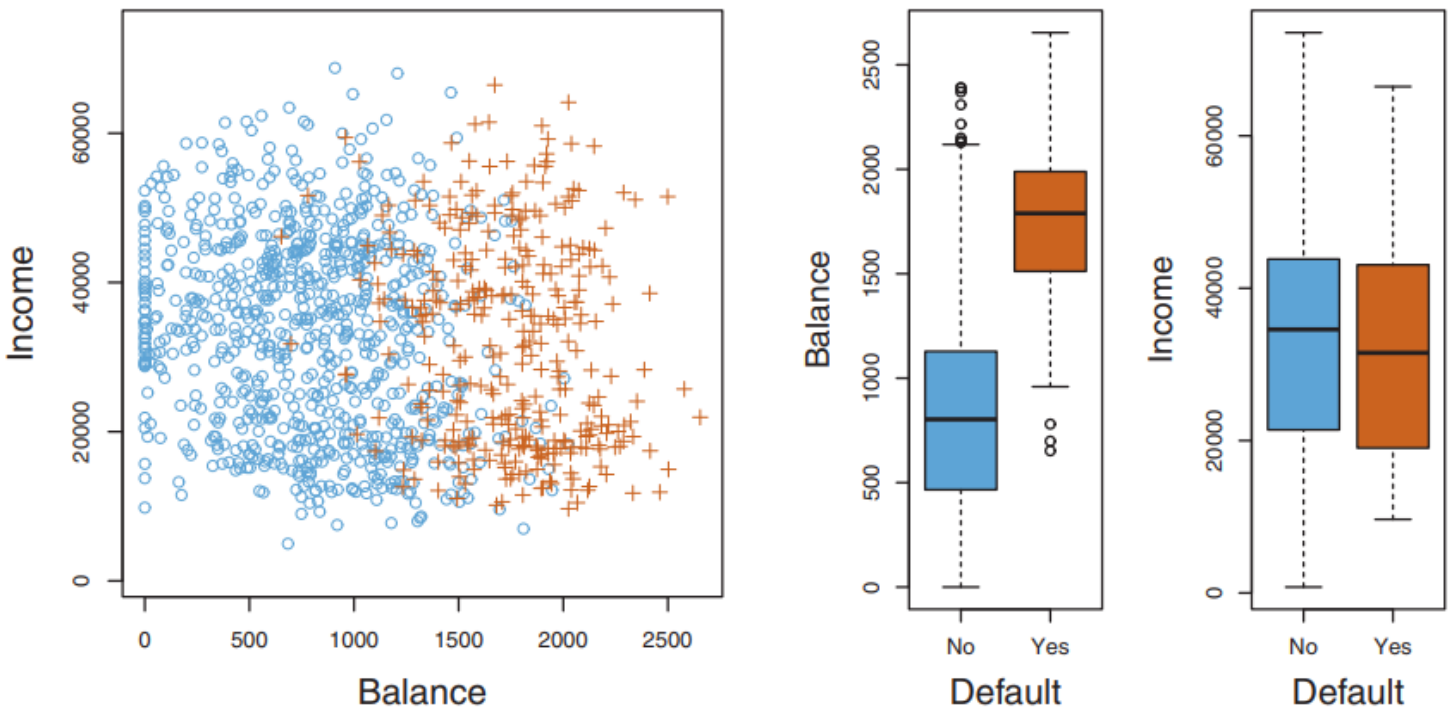


FIGURE 4.1. *The* `Default` *data set.* Left: *The annual incomes and monthly credit card balances of a number of individuals. The individuals who defaulted on their credit card payments are shown in orange, and those who did not are shown in blue.* Center: *Boxplots of* `balance` *as a function of*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

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

On the Agenda

On the Agenda

On the Agenda

and...

laureano@ufpr.br