Module 2: Statistical Machine Learning

Introduction, Regression, and **Classification**, Decision Trees, Random Forests

Reference: James, G., Witten, D., Hastie, T., Tibshirani, R. (2017). An Introduction to Statistical Learning: with Applications in R., Springer.)

Classification

- Response variable is **qualitative**
- For example, eye color is qualitative - blue, brown, or green
- Often qualitative variables are referred to as **categorical**
- Process of predicting qualitative responses - classification
- Assigning the observation to a category, or class
- Predict the probability of each of the categories of a qualitative variable
- Classifiers - logistic regression, linear discriminant analysis, and K-nearest neighbors

Classification

- Set of training observations $(x_1, y_1), \ldots, (x_n, y_n)$ that we can use to build a classifier.

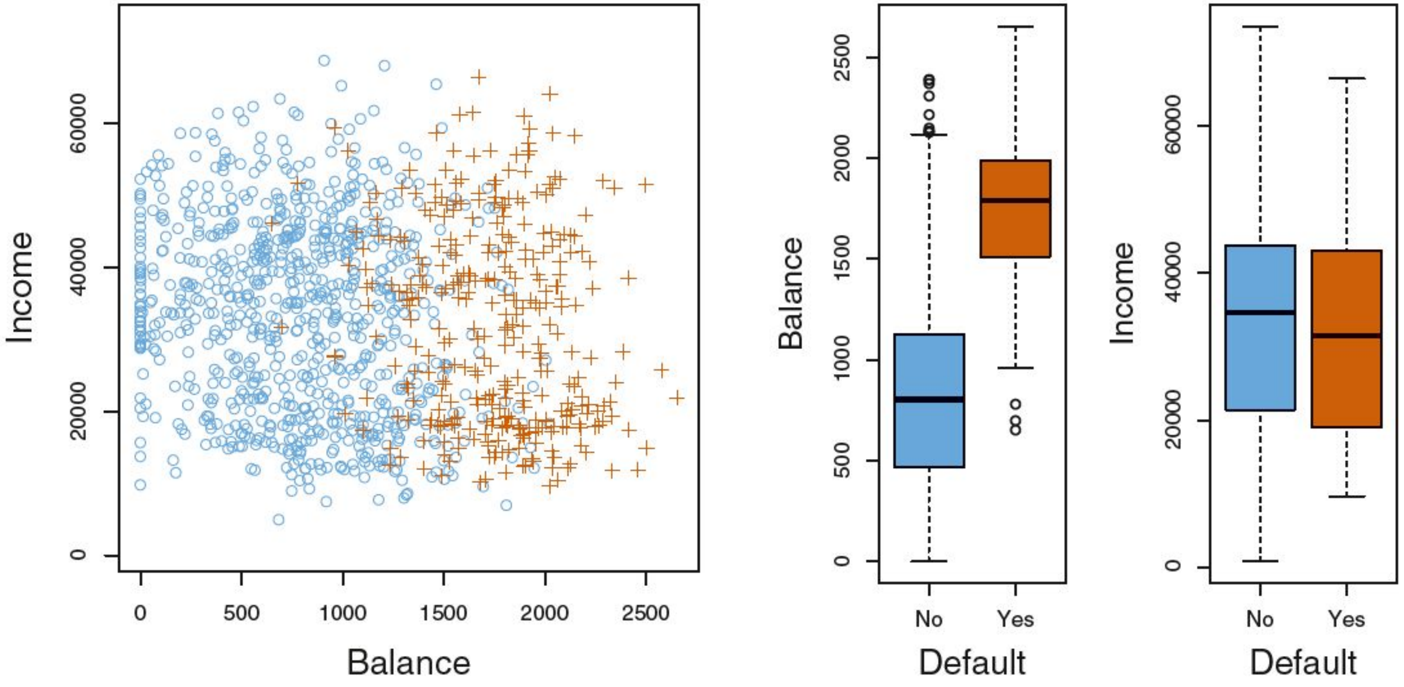

FIGURE 4.1. *The* `Default` *data set.* Left: *The annual incomes and monthly credit card balances of a number of individuals. The individuals who defaulted on their credit card payments are shown in orange, and those who did not are shown in blue.* Center: *Boxplots of* `balance` *as a function of*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

Logistic Regression

Consider the Default data set, where the response default falls into one of two categories, Yes or No.
Rather than modeling this response Y directly, logistic regression models the probability that Y belongs to a particular category.

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

The values of $\Pr(\text{default} = \text{Yes}|\text{balance})$, which we abbreviate $p(\text{balance})$, will range between 0 and 1.

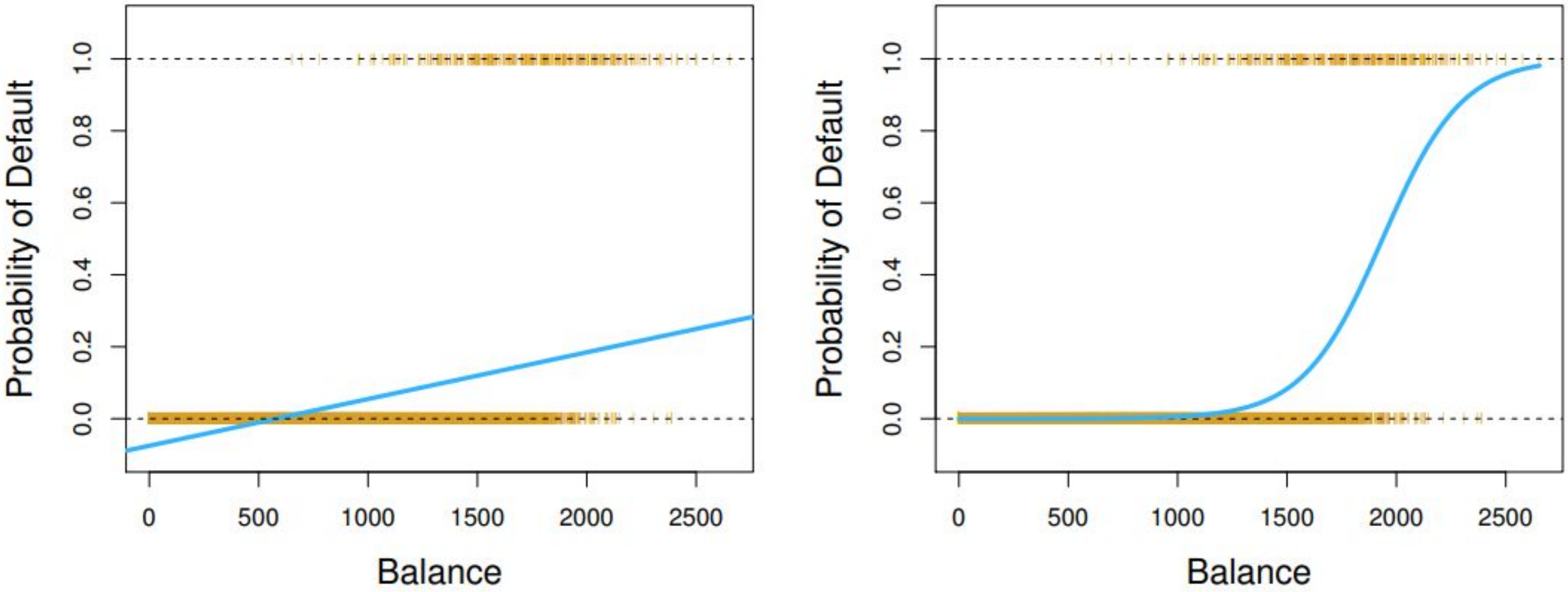

FIGURE 4.2. *Classification using the* `Default` *data.* Left: *Estimated probability of* `default` *using linear regression. Some estimated probabilities are negative! The orange ticks indicate the 0/1 values coded for* `default`*(*`No` *or* `Yes`*).* Right: *Predicted probabilities of* `default` *using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

The Logistic Model

Model the relationship between $p(X) = \Pr(Y = 1|X)$ and X

The logistic function

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

To fit the model, we use a method called maximum likelihood

The logistic function will always produce an S-shaped curve of this form

The Logistic Model

After a bit of manipulation find that

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

The quantity p(X)/[1−p(X)] is called the odds, and can take on any value between 0 and ∞

By taking the logarithm of both sides

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

The left-hand side is called the ***log odds*** or ***logit***. Logit is linear in X

Estimating the Regression Coefficients

The coefficients β0 and β1 are unknown

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

Although we could use (non-linear) least squares to ft the model, the more general method of maximum likelihood is preferred, since it has better statistical properties

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

Estimating the Regression Coefficients

Find $\hat{\beta}_0$ and $\hat{\beta}_1$ - estimates into the model for p(X)

A number close to one for all individuals who defaulted, and a number close to zero for all individuals who did not

This intuition can be formalized using a mathematical equation called a likelihood function:

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

The estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ are chosen to *maximize* this likelihood function.

Estimating the Regression Coefficients

	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

TABLE 4.1. *For the* `Default` *data, estimated coefficients of the logistic regression model that predicts the probability of* `default` *using* `balance`*. A one-unit increase in* `balance` *is associated with an increase in the log odds of* `default` *by* 0.0055 *units.*

Making Predictions

Once the coefficients have been estimated, we can compute the probability of `default` for any given credit card balance. For example, using the coefficient estimates given in Table 4.1, we predict that the default probability for an individual with a `balance` of $1,000 is

$$\hat{p}(X) = \frac{e^{\hat{\beta}_0+\hat{\beta}_1 X}}{1+e^{\hat{\beta}_0+\hat{\beta}_1 X}} = \frac{e^{-10.6513+0.0055\times 1,000}}{1+e^{-10.6513+0.0055\times 1,000}} = 0.00576,$$

which is below 1 %. In contrast, the predicted probability of default for an individual with a balance of $2,000 is much higher, and equals 0.586 or 58.6 %.

Suppose that we wish to classify an observation into one of K classes, where $K \geq 2$. In other words, the qualitative response variable Y can take on K possible distinct and unordered values. Let π_k represent the overall or *prior* probability that a randomly chosen observation comes from the kth class. Let $f_k(X) \equiv \Pr(X|Y = k)$[1] denote the *density function* of X for an observation that comes from the kth class. In other words, $f_k(x)$ is relatively large if there is a high probability that an observation in the kth class has $X \approx x$, and $f_k(x)$ is small if it is very unlikely that an observation in the kth class has $X \approx x$. Then *Bayes' theorem* states that

$$\Pr(Y = k|X = x) = \frac{\pi_k f_k(x)}{\sum_{l=1}^{K} \pi_l f_l(x)}. \tag{4.15}$$

In accordance with our earlier notation, we will use the abbreviation $p_k(x) = \Pr(Y = k|X = x)$; this is the *posterior* probability that an observation $X = x$ belongs to the kth class. That is, it is the probability that the observation belongs to the kth class, *given* the predictor value for that observation.

Linear Discriminant Analysis for $p = 1$