CLASSIFICATION

Classification

- Logistic Regression

Classification

- Qualitative variables take values in an unordered set C, such as:

    eye color $\in$ {brown; blue; green}

    email $\in$ { spam; ham}.

- Given a feature vector X and a qualitative response Y taking values in the set C, the classification task is to build a function C(X) that takes as input the feature vector X and predicts its value for Y ; i.e. C(X) $\in$ C.

- Often we are more interested in estimating the probabilities that X belongs to each category in C.

For example, it is more valuable to have an estimate of the probability that an insurance claim is fraudulent, than a classification fraudulent or not.

Classification

Example: Credit Card Defualt
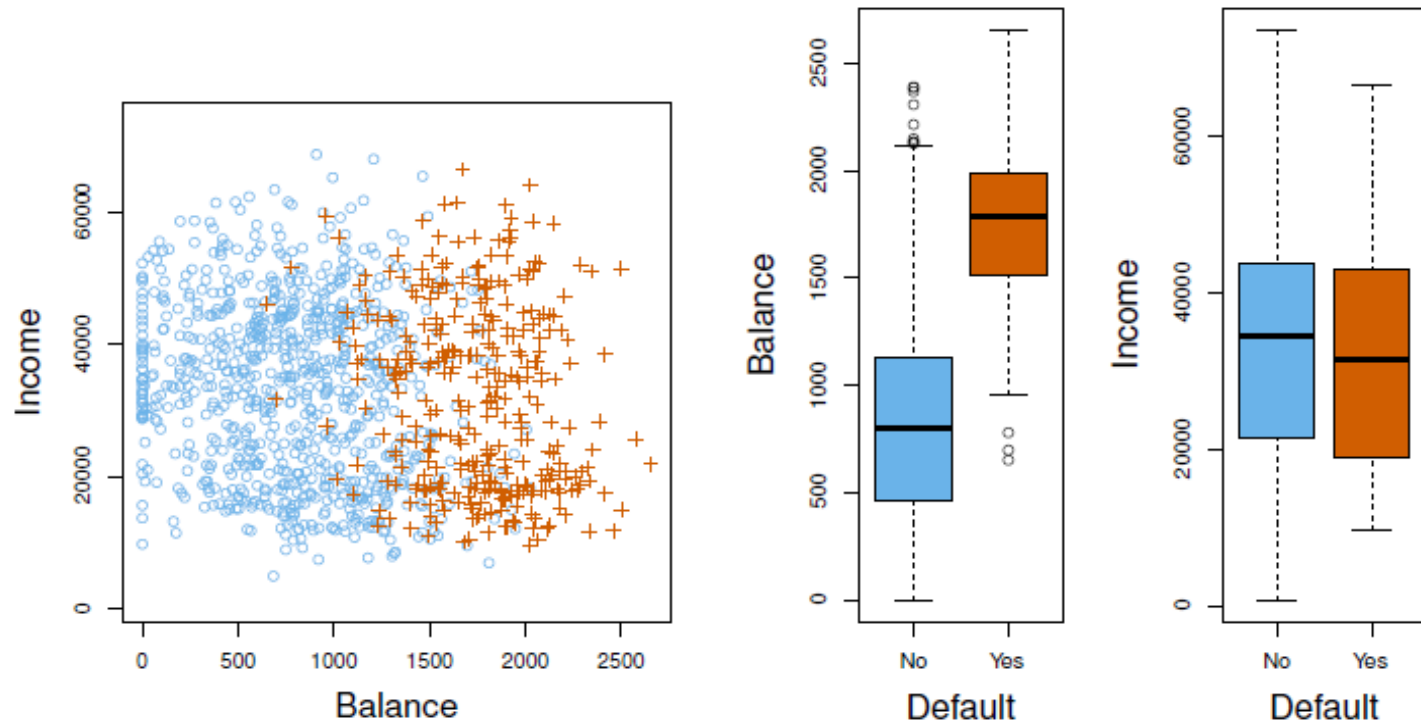
Classification

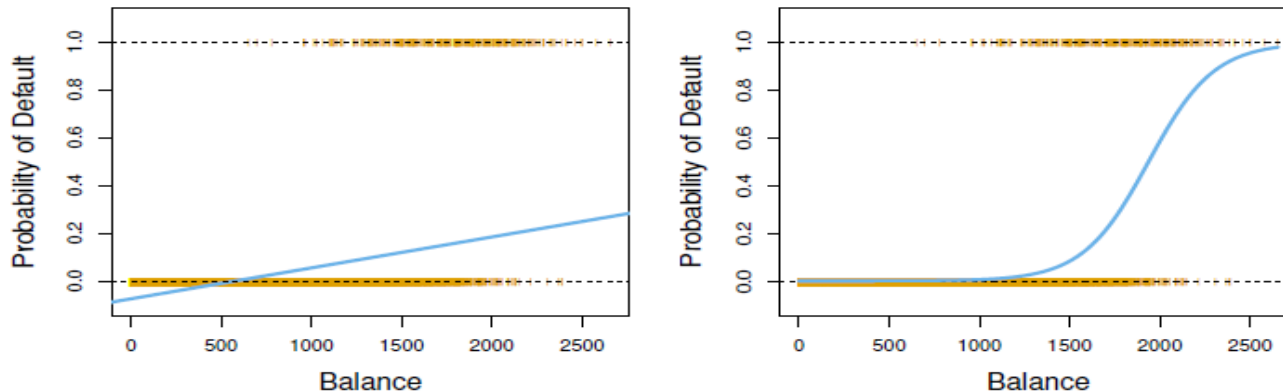Can we use Linear Regression?

- Suppose for the Default classification task that we code

$$Y = \begin{cases} 0 & \text{if No} \\ 1 & \text{if Yes} \end{cases}$$

- Can we simply perform a linear regression of Y on X and classify as Yes if $\hat{Y} > 0:5$?
- In this case of a binary outcome, linear regression does a good job as a classifier, and is equivalent to *linear discriminant analysis* which we discuss later.
- Since in the population $E(Y|X = x) = Pr(Y = 1|X = x)$, we might think that regression is perfect for this task.
- However, linear regression might produce probabilities less than zero or bigger than one. Logistic regression is more appropriate.

Classification

The orange marks indicate the response Y , either 0 or 1. Linear regression does not estimate $Pr(Y = 1|X)$ well. Logistic regression seems well suited to the task.

Classification

Linear Regression…

Now suppose we have a response variable with three possible values. A patient presents at the emergency room, and we must classify them according to their symptoms.

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

This coding suggests an ordering, and in fact implies that the difference between *stroke* and *drug overdose* is the same as between *drug overdose* and *epileptic seizure*.

- Linear regression is not appropriate here.
- Multiclass Logistic Regression or Discriminant Analysis are more appropriate.

Logistic Regression...Model

Let's write $p(X) = \Pr(Y = 1|X)$ for short and consider using `balance` to predict `default`. Logistic regression uses the form

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

($e \approx 2.71828$ is a mathematical constant [Euler's number.]) It is easy to see that no matter what values β_0, β_1 or X take, $p(X)$ will have values between 0 and 1.
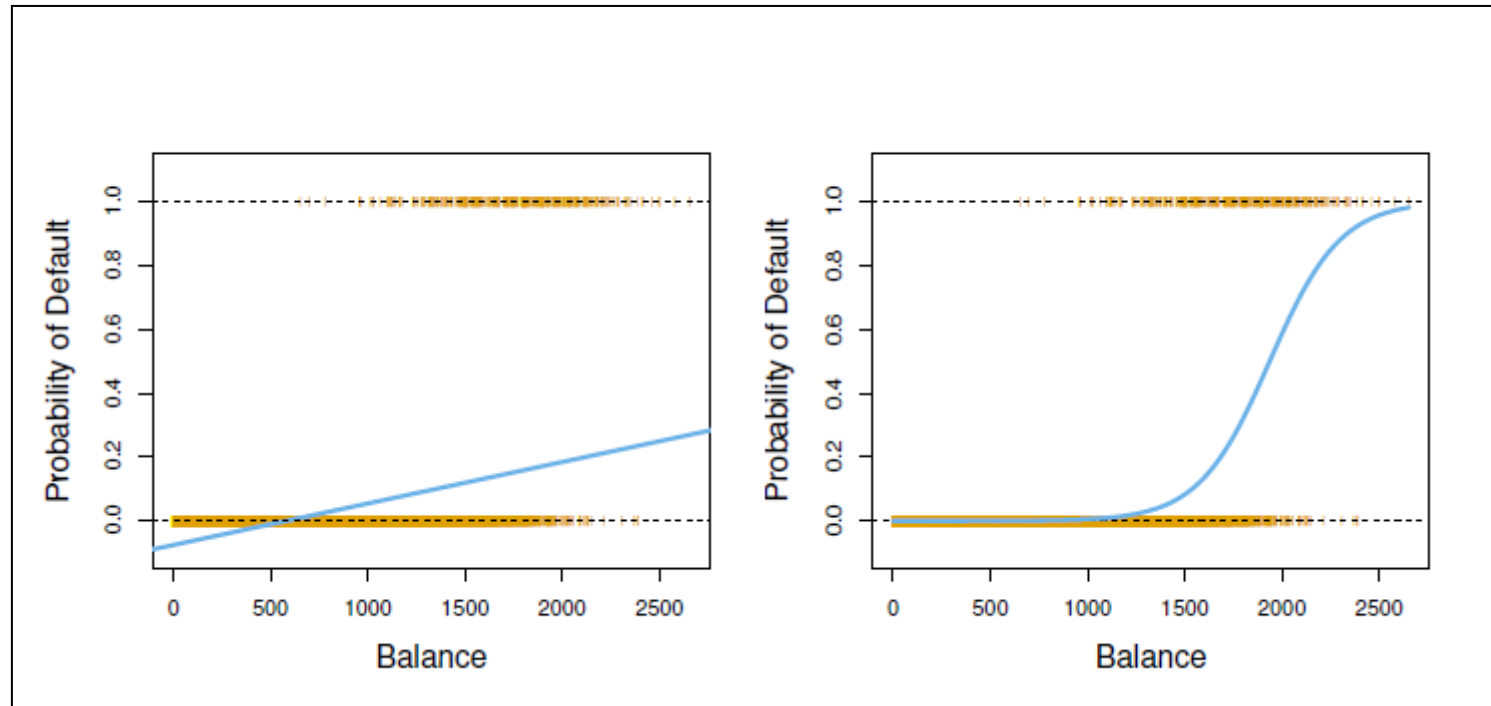
Logistic Regression…Model

A bit of rearrangement gives

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

This monotone transformation is called the *log odds* or *logit* transformation of $p(X)$.

Linear versus Logistic Regression

Logistic regression ensures that our estimate for p(X) lies between 0 and 1.

Logistic Regression...Parameter estimation

Maximum Likelihood

- We use maximum likelihood to estimate the parameters.

$$\ell(\beta_0, \beta) = \prod_{i:y_i=1} p(x_i) \prod_{i:y_i=0} (1 - p(x_i)).$$

- This likelihood gives the probability of the observed zeros and ones in the data. We pick β_0 and β_1 to maximize the likelihood of the observed data.

Logistic Regression…Output of R

Most statistical packages can fit linear logistic regression models by maximum likelihood. In R we use the glm function.

	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

Logistic Regression…Predictions

Making Predictions

What is our estimated probability of default for someone with a balance of $1000?

$$\hat{p}(X) = \frac{e^{\hat{\beta}_0 + \hat{\beta}_1 X}}{1 + e^{\hat{\beta}_0 + \hat{\beta}_1 X}} = \frac{e^{-10.6513 + 0.0055 \times 1000}}{1 + e^{-10.6513 + 0.0055 \times 1000}} = 0.006$$

With a balance of 2000$

$$\hat{p}(X) = \frac{e^{\hat{\beta}_0 + \hat{\beta}_1 X}}{1 + e^{\hat{\beta}_0 + \hat{\beta}_1 X}} = \frac{e^{-10.6513 + 0.0055 \times 2000}}{1 + e^{-10.6513 + 0.0055 \times 2000}} = 0.586$$

Logistic Regression…Predictions

- Lets do this for predictor 'student':

	Coefficient	Std. Error	Z-statistic	P-value
Intercept	-3.5041	0.0707	-49.55	< 0.0001
student[Yes]	0.4049	0.1150	3.52	0.0004

$$\widehat{\Pr}(\texttt{default=Yes}|\texttt{student=Yes}) = \frac{e^{-3.5041+0.4049\times1}}{1 + e^{-3.5041+0.4049\times1}} = 0.0431,$$

$$\widehat{\Pr}(\texttt{default=Yes}|\texttt{student=No}) = \frac{e^{-3.5041+0.4049\times0}}{1 + e^{-3.5041+0.4049\times0}} = 0.0292.$$

Logistic regression with several variables

$$\log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

$$p(X) = \frac{e^{\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p}}{1 + e^{\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p}}$$

	Coefficient	Std. Error	Z-statistic	P-value
Intercept	-10.8690	0.4923	-22.08	< 0.0001
balance	0.0057	0.0002	24.74	< 0.0001
income	0.0030	0.0082	0.37	0.7115
student[Yes]	-0.6468	0.2362	-2.74	0.0062

Logistic regression with more than two classes

So far we have discussed logistic regression with two classes.
It is easily generalized to more than two classes.

$$\Pr(Y = k|X) = \frac{e^{\beta_{0k}+\beta_{1k}X_1+...+\beta_{pk}X_p}}{\sum_{\ell=1}^{K} e^{\beta_{0\ell}+\beta_{1\ell}X_1+...+\beta_{p\ell}X_p}}$$

Multiclass logistic regression is also referred to as multinomial regression.
For running multinomial regression we need to install R package glmnet

Thank You