Statistical Learning. Linear methods for classification

Jose Ameijeiras Alonso
Departamento de Estatística e Investigación Operativa (USC)

Máster Interuniversitario en Tecnologías de Análisis de Datos Masivos: Big Data

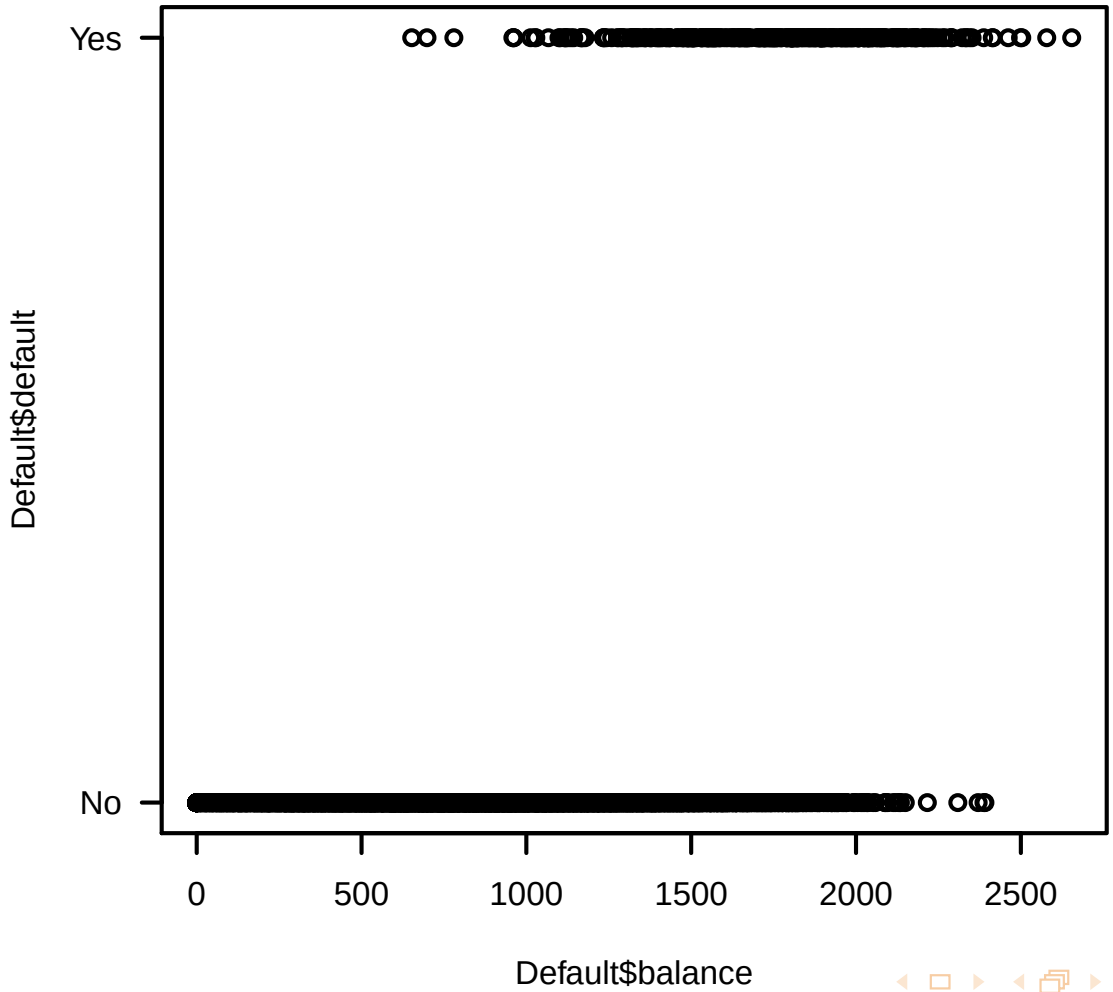
Introduction

Example 3: Suppose we are interested in predicting whether an individual will default on his or her credit card payment, on the basis of credit card balance

The Default data set contains customer default records for a credit card company.

Introduction

```
> library(ISLR)
> data(Default)
> plot(Default$balance, Default$default)
```

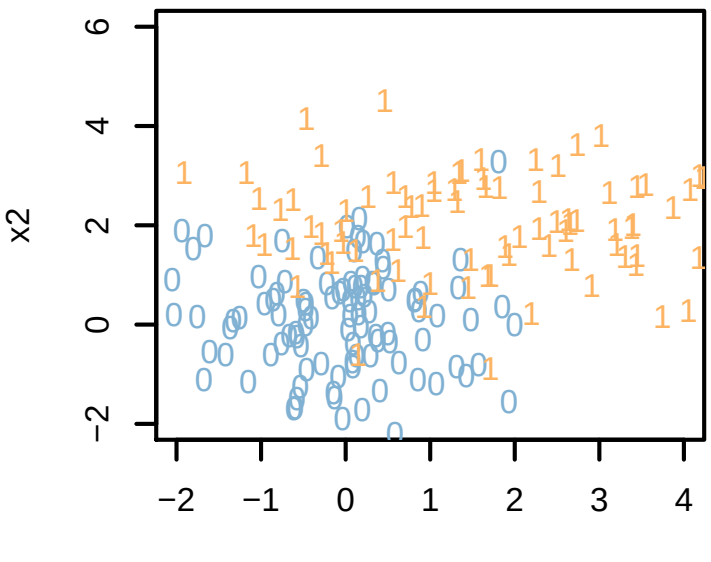


Introduction

- This is an example of supervised learning.
- For each observation of the predictor measurement x_i (balance) there is an associated response measurement y_i (default yes or no), for $i = 1, \ldots, n$. (“right answers”)
- Classification problems: problems with a qualitative response.

Classification problems

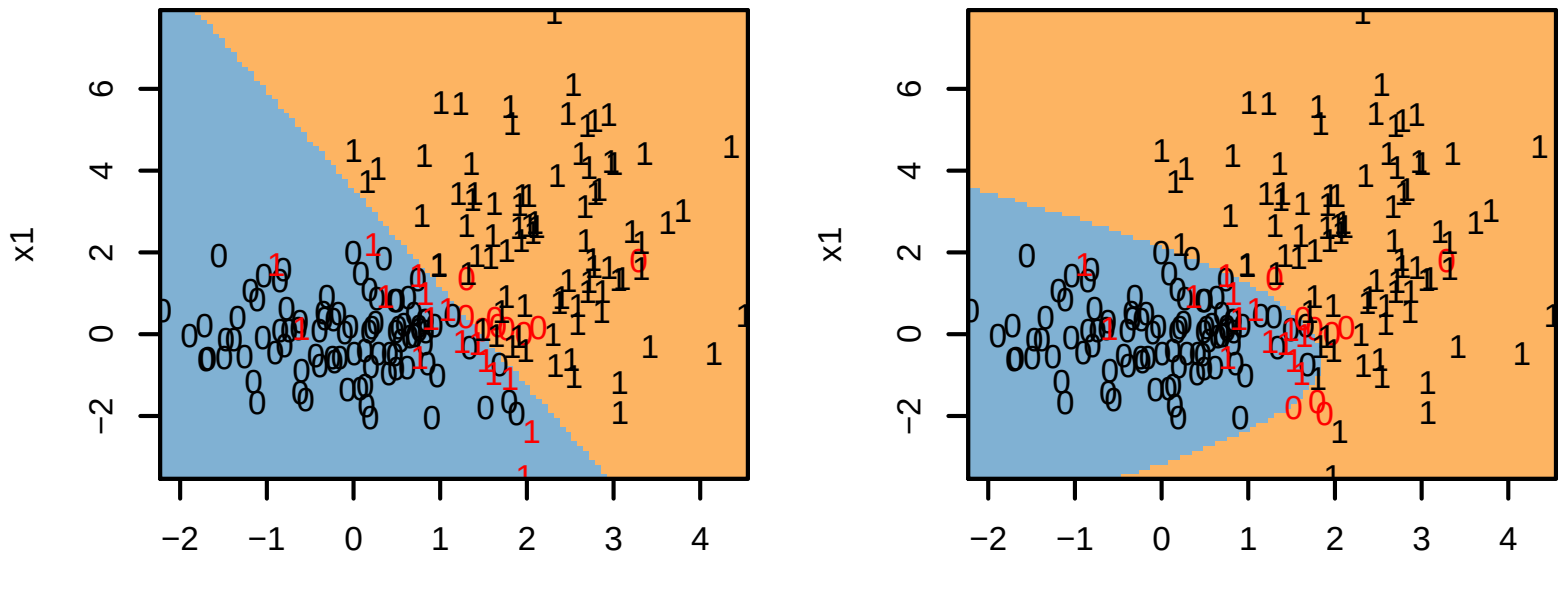
- We observe a qualitative response Y and predictor variables $X = (X_1, \ldots, X_p)$.
- The response variable Y is qualitative, that is, Y can take on K possible values.
- The classification task is to build a function $G(X)$ to predict Y based on X.
- The function $G(X)$ divides the feature vector space into a collection of regions, each labeled by one class.



- We have a training sample $(y_1, x_{11}, \ldots, x_{1p}), \ldots, (y_n, x_{n1}, \ldots, x_{np})$, where y_i can take on K possible values.

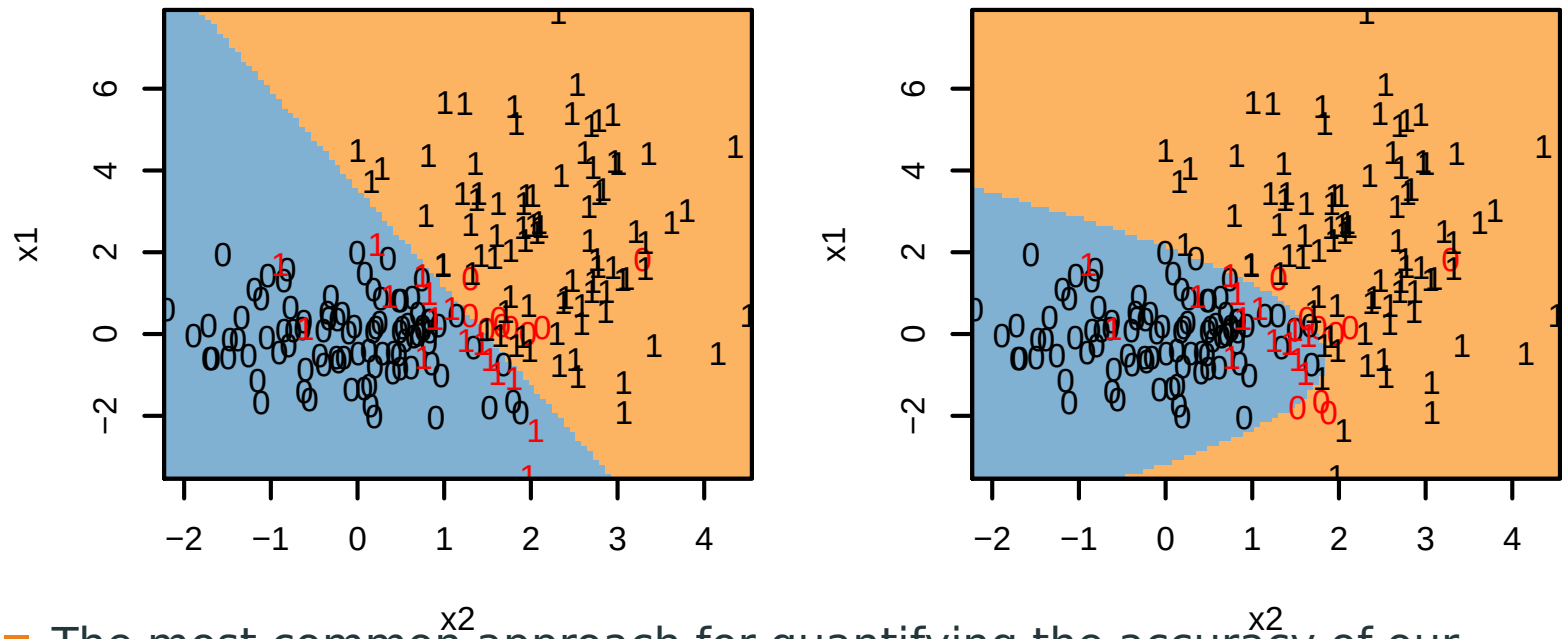
Classification problems

- Linear Methods for Classification: The decision boundaries are linear
  - For binary classification (Y can take only two values): The decision boundary between the two classes is a hyperplane in the feature vector space
  - If Y can take more than two classes: The decision boundary between any pair of classes is a hyperplane



- How do you choose the hyperplanes between classes?

Classification problems



- The most common approach for quantifying the accuracy of our estimate is the training error rate, the proportion of mistakes that are made if we apply our estimate to the training observations:

$$\frac{1}{n}\sum_{i=1}^{n} I\{\hat{y}_i \neq y_i\}$$

Classification problems

Suppose that we observe a qualitative variable Y that can take on K possible values and predictor variables $X = (X_1 \ldots, X_p)$.

- We wish to classify an observation $X = x$ into one of the K classes given by Y
- Often we are more interested in estimating the probabilities that X belongs to each category of Y. For example, it is more valuable to have an estimate of the probability that an individual will default on his or her credit card payment, than just a classification (default yes or no)
- Therefore, we are interested in

$$\mathbb{P}(Y = k/X = x), \quad k = 1, \ldots, K,$$

  the conditional probability that an observation belongs to class $Y = k$ given the predictor value $X = x$
- We will classify the observation $X = x$ to the class for which the value of $\mathbb{P}(Y = k/X = x)$ is greatest

Logistic regression

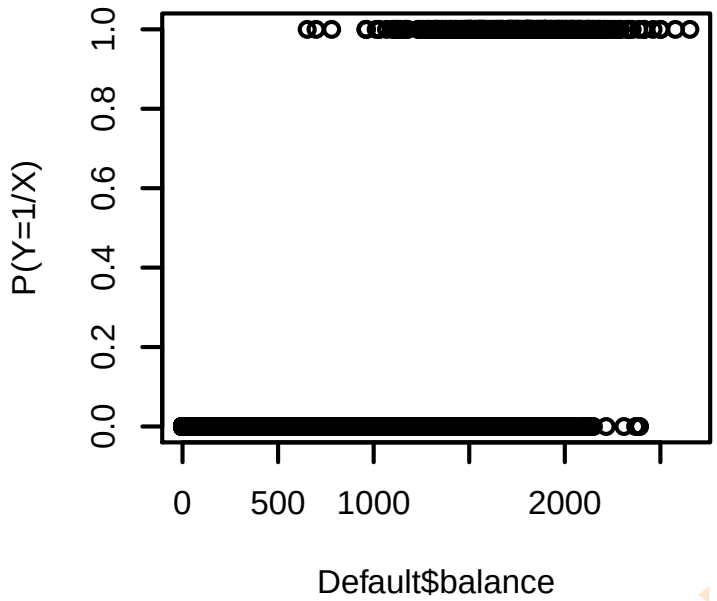
- We will now consider the classification problem for a binary qualitative response Y
- We can use a 0/1 dummy variable to code the response. For instance, for the example of the credit card company

$$Y = \begin{cases} 1, & \text{if default=yes} \\ 0, & \text{if default=no} \end{cases}$$

- We want to model

$$\boxed{p(X) = \mathbb{P}(Y = 1/X)}$$

- Then, we would use a rule that classifies an observation X in the group $Y = 1$ if $p(X) > 0.5$ (otherwise, we would classify X in the group $Y = 0$)

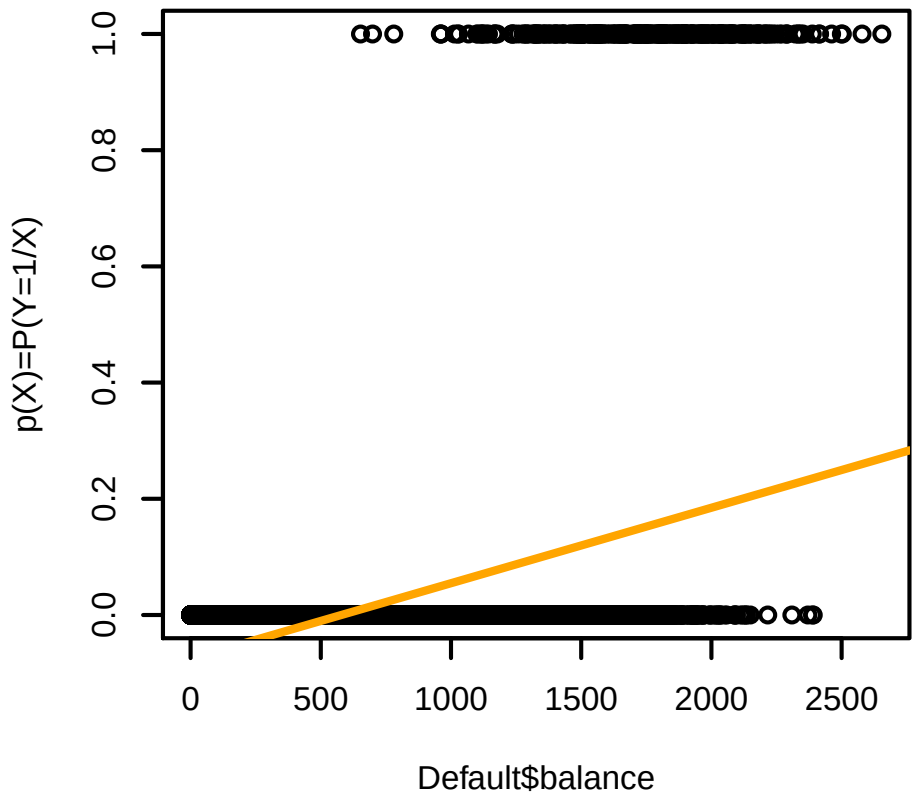


Logistic regression

$$Y = \begin{cases} 1, & \text{if default=yes} \\ 0, & \text{if default=no} \end{cases}$$

- If we use a linear model...

$$p(X) = \mathbb{P}(Y = 1/X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p$$



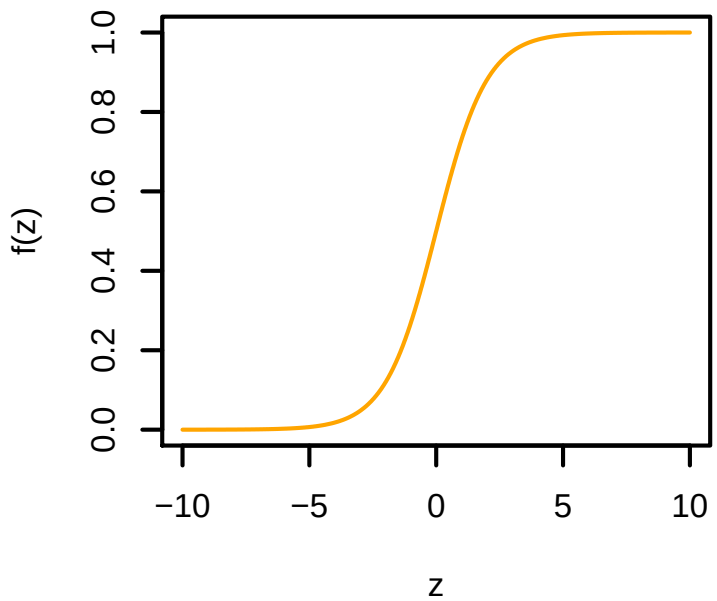
Logistic regression

- In logistic regression, we use the model

$$p(X) = \mathbb{P}(Y = 1/X) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p)}}$$

Logistic regression

$$f(z) = \frac{1}{1 + e^{-z}} = \frac{e^z}{1 + e^z}$$



- $0 \leq f(z) \leq 1$
- $f(z) \to 0$ as $z \to -\infty$ and $f(z) \to 1$ as $z \to \infty$

Logistic regression

- Logistic regression for classification preserves linear classification boundaries. Note that the decision boundary between classes $Y = 0$ and $Y = 1$ is determined by the equation

$$\mathbb{P}(Y = 1/X) = \mathbb{P}(Y = 0/X)$$

or equivalently,

$$\log\left(\frac{\mathbb{P}(Y = 1/X)}{\mathbb{P}(Y = 0/X)}\right) = 0$$

- Now, since $\mathbb{P}(Y = 0/X) = 1 - \mathbb{P}(Y = 1/X)$, we have the condition

$$\log\left(\frac{p(X)}{1 - p(X)}\right) = 0$$

- And using the logistic model we get that

$$\log\left(\frac{p(X)}{1 - p(X)}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p$$

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

Fitting the Logistic Regression Model

- We observe a binary qualitative response Y and predictor variables $X = (X_1, \ldots, X_p)$.
- We assume that

$$p(X) = \mathbb{P}(Y = 1/X) = \frac{1}{1 + \mathrm{e}^{-(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p)}}$$

- Given a training sample $(y_1, x_{11}, \ldots, x_{1p}), \ldots, (y_n, x_{n1}, \ldots, x_{np})$ we want to estimate $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^t$.
- The parameters are chosen using the Maximum likelihood approach

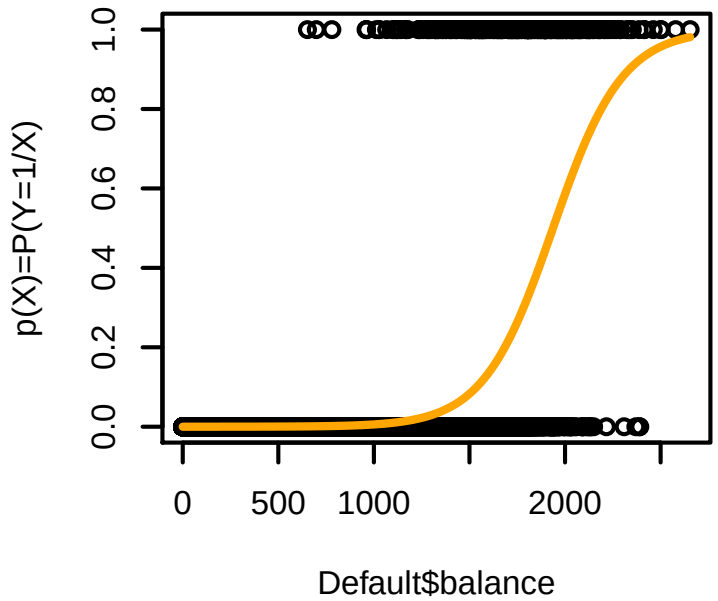
Logistic regression

Example 3: Suppose we are interested in predicting whether an individual will default on his or her credit card payment, on the basis of credit card balance

- We observe a binary response Y and a predictor variable X and assume

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

```
> gl <- glm(default ~ balance, data = Default, family = "binomial")
```



- $\hat{\beta}_0 = -10.65133.$
- $\hat{\beta}_1 = 0.00549.$

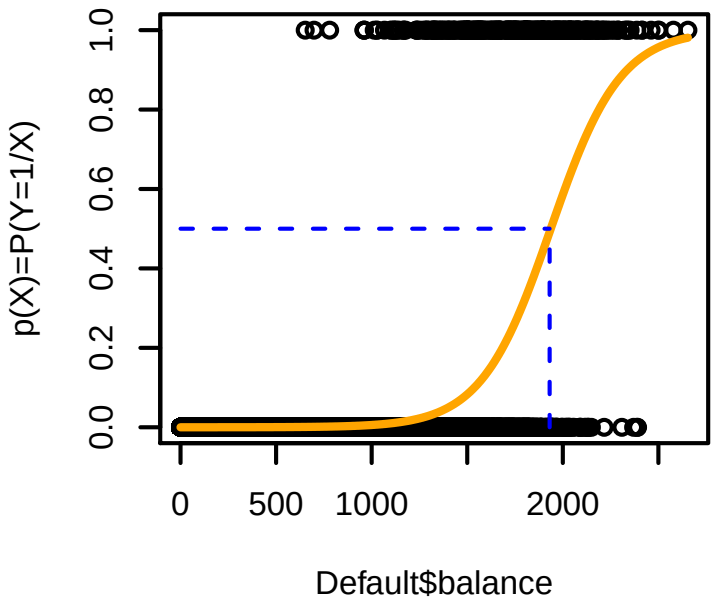
Logistic regression

Example 3: Suppose we are interested in predicting whether an individual will default on his or her credit card payment, on the basis of credit card balance

- Predictions:

$$\hat{p}(X) = \frac{1}{1 + e^{-(\hat{\beta}_0 + \hat{\beta}_1 X)}}$$

```
> gl <- glm(default ~ balance, data = Default, family = "binomial")
```



Logistic regression

```
##
## Call:
## glm(formula = default ~ balance, family = "binomial", data = Defaul
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.2697  -0.1465  -0.0589  -0.0221   3.7589
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.065e+01  3.612e-01  -29.49   <2e-16 ***
## balance      5.499e-03  2.204e-04   24.95   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2920.6  on 9999  degrees of freedom
## Residual deviance: 1596.5  on 9998  degrees of freedom
## AIC: 1600.5
##
## Number of Fisher Scoring iterations: 8
```

Classification based on Bayes rule

Suppose that we observe a qualitative variable Y that can take on K possible values and predictor variables $X = (X_1 \ldots, X_p)$.

- We wish to classify an observation $X = x$ into one of the K classes given by Y.
- We are interested in $\mathbb{P}(Y = k/X = x)$, $k = 1, \ldots, K$ (posterior probability)
- We classify the observation to the class for which $\mathbb{P}(Y = k/X = x)$ is greatest

- Logistic regression involves directly modeling $\mathbb{P}(Y = k/X)$ using the logistic function.
- Logistic regression estimates the posterior probabilities of classes given X without assuming any distribution on X (discriminative modeling)
- We now consider an alternative and less direct approach to estimating these probabilities (generative modeling)
- We estimate the within class density of X given the class label and the posterior probability of Y can be obtained by the Bayes' formula

k-Nearest Neighbors (KNN)

We are interested in $\mathbb{P}(Y = k/X = x)$, $k = 1, \ldots, K$.

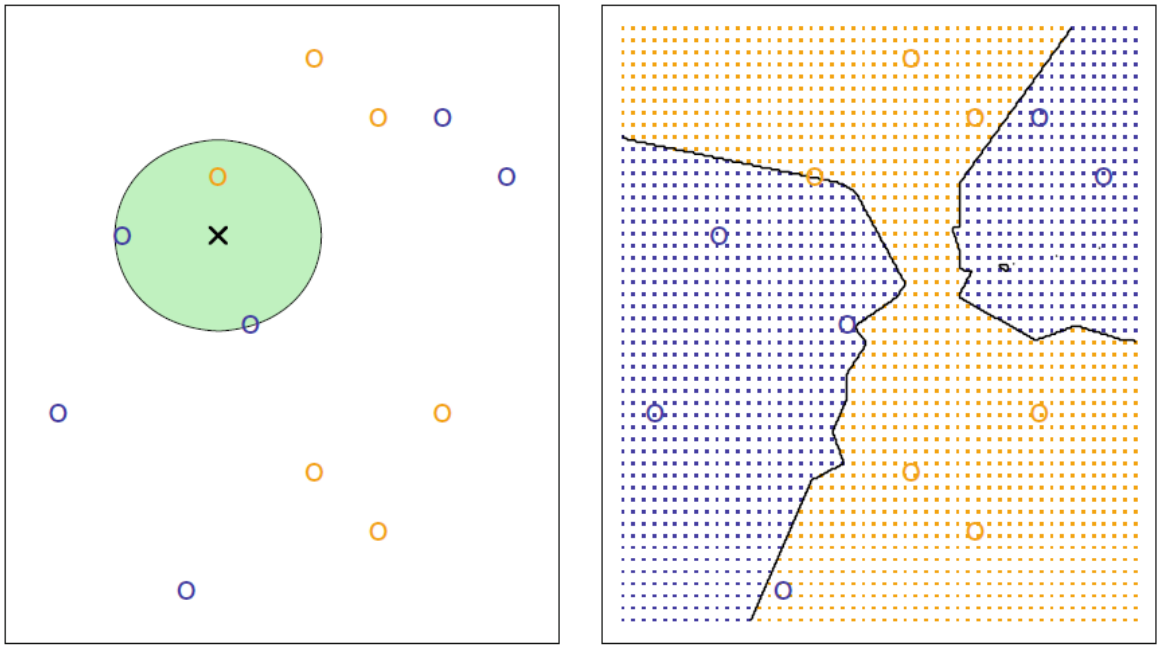
Given the number of neighbors k and a test observation x_0, KNN classifier first identifies the k points in the training data that are closest to x_0, represented by N_0. It then estimates the conditional probability for class j as the fraction of points in N_0 whose response values equal j:

$$\mathbb{P}(Y = j/X = x_0) = \frac{1}{k} \sum_{i \in N_0} I(y_i = j)$$

Finally, KNN applies Bayes rule and classifies the test observation x_0 to the class with the largest probability.

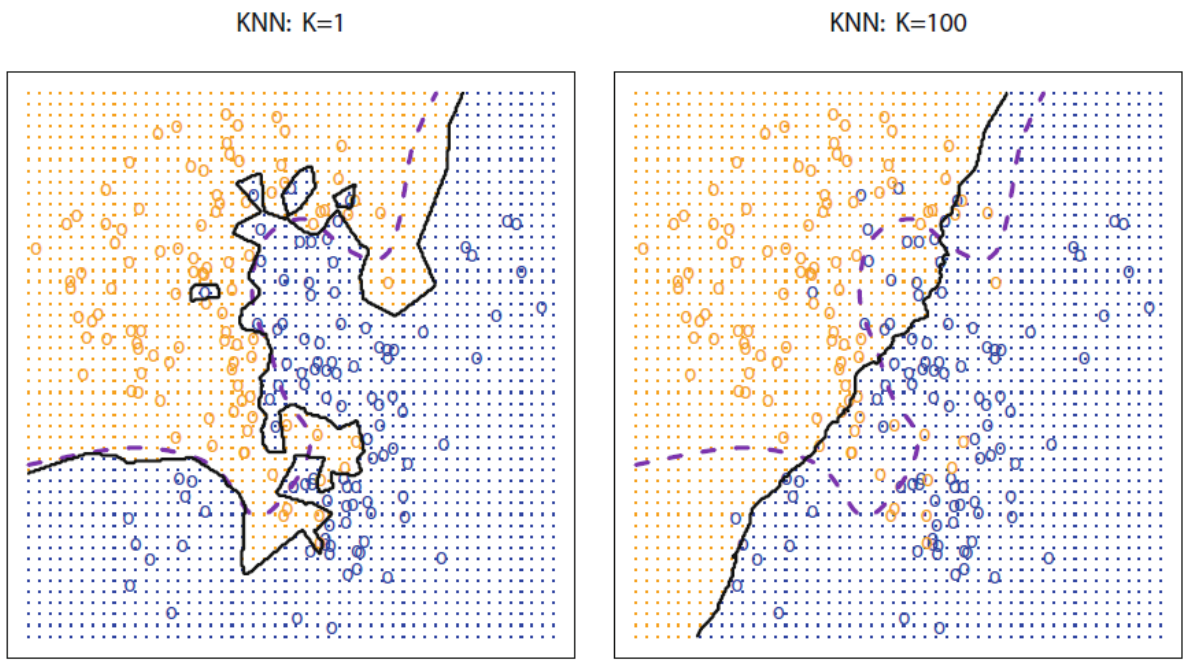
k-Nearest Neighbors (KNN)

Example: training dataset with 2 classes (orange and blue dots). Number of neighbors ($k = 3$).



k-Nearest Neighbors (KNN)

Example: training dataset with 2 classes (orange and blue dots). Number of neighbors: $K = 1$ (left), $K = 100$ (right). Purple: reality.



Linear Discriminant Analysis (LDA)

Bayes rule:

Consider any two events A and B. The Bayes Rule states that to find $\mathbb{P}(B/A)$ (probability that B occurs given that A has occurred) we have that:

$$\mathbb{P}(B/A) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(A)} = \frac{\mathbb{P}(A/B)\mathbb{P}(B)}{\mathbb{P}(A)}$$

Bayes Rule in a Classification Problem:

- Suppose that the π_k denotes the overall or prior probability that a randomly chosen observation comes form the kth class, that is, $\pi_k = \mathbb{P}(Y = k)$
- Let $f_k(x) = \mathbb{P}(X = x/Y = k)$. That is, $f_k(x)$ is the density for X in class k.

$$\mathbb{P}(Y = k/X = x) = \frac{\mathbb{P}(X = x/Y = k)\mathbb{P}(Y = k)}{\mathbb{P}(X = x)} = \frac{f_k(x)\pi_k}{\sum_{j=1}^{K} f_j(x)\pi_j}$$

Linear Discriminant Analysis (LDA)

- Suppose that the π_k denotes the overall or prior probability that a randomly chosen observation comes form the kth class, that is, $\pi_k = \mathbb{P}(Y = k)$
- Let $f_k(x) = \mathbb{P}(X = x/Y = k)$. That is, $f_k(x)$ is the density for X in class k.

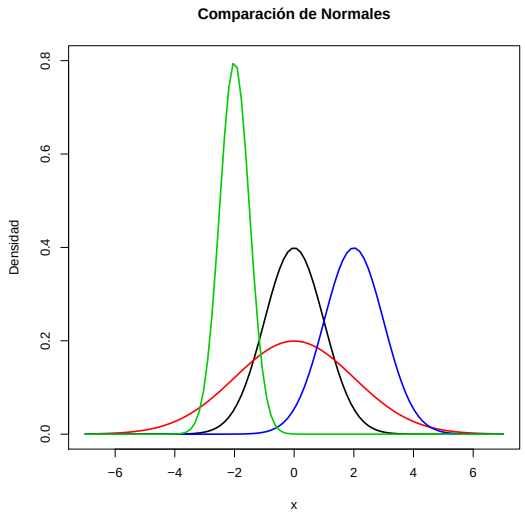
$$\mathbb{P}(Y = k/X = x) = \frac{\mathbb{P}(X = x/Y = k)\mathbb{P}(Y = k)}{\mathbb{P}(X = x)} = \frac{f_k(x)\pi_k}{\sum_{j=1}^{K} f_j(x)\pi_j}$$

- Therefore, instead of directly computing $\mathbb{P}(Y = k/X = x)$, we can plug in estimates of π_k and $f_k(x)$
- In general, we can estimate π_k as the fraction of the training observations that belong to the k-th class
- Estimating $f_k(x)$ tends to be more challenging, unless we assume some simple forms for these densities

Linear Discriminant Analysis (LDA)

- Under Linear Discriminant Analysis (LDA) we assume that the density for X, given every class k is following a Gaussian distribution
- For $p = 1$ a random variable X with normal distribution with mean μ and variance σ^2 has density

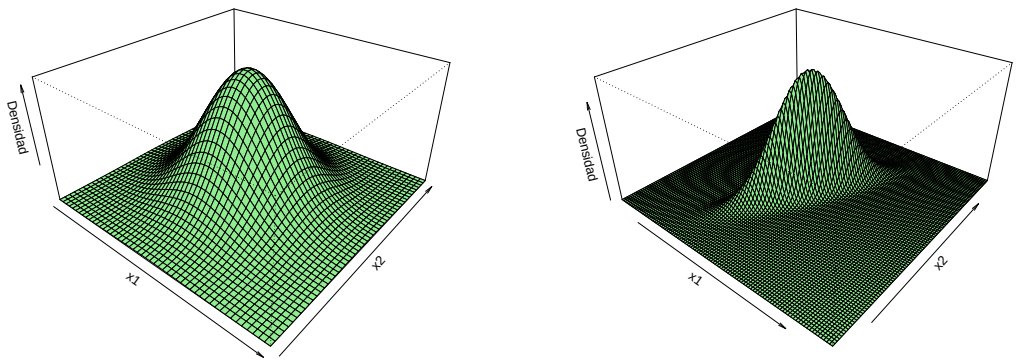
$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}, \quad -\infty < x < \infty.$$



Linear Discriminant Analysis (LDA)

- Under Linear Discriminant Analysis (LDA) we assume that the density for X, given every class k is following a Gaussian distribution
- For $p > 1$, a p-dimensional random vector $X = (X_1, \ldots, X_p)$ with multivariate normal distribution with mean vector μ and covariance matrix Σ has density

$$f(x) = \frac{1}{(\sqrt{2\pi})^p \mid \Sigma \mid^{1/2}} e^{\frac{-(x-\mu)'\Sigma^{-1}(x-\mu)}{2}}, \; x \in \mathbb{R}^p.$$



Linear Discriminant Analysis (LDA)

- Suppose that we observe a qualitative variable Y that can take on K possible values and predictor variables $X = (X_1 \ldots, X_p)$.
- Under Linear Discriminant Analysis (LDA) we assume that:
  - In each class k, the probability density function of X is multivariate normal with mean vector μ_k and covariance matrix Σ_k
  - The covariance matrix is same for all populations $\Sigma_1 = \cdots = \Sigma_K$, that is, there is a shared covariance matrix across all K classes, denoted Σ

$$f_k(x) = \frac{1}{(\sqrt{2\pi})^p \mid \Sigma \mid^{1/2}} e^{\frac{-(x-\mu_k)' \Sigma^{-1} (x-\mu_k)}{2}}, \; x \in \mathbb{R}^p, k = 1, \ldots, K$$

- Then, the Bayes classifier assigns an observation $X = x$ to the class for which $f_k(x)\pi_k$ is largest.
- Because a log transform is monotonic, this equivalent to classifying an observation $X = x$ to the class for which $\log(f_k(x)\pi_k)$ is largest.
- It can be proved that the LDA rule assigns an observation $X = x$ to the class for which $\delta_k(x)$ is largest, where

$$\delta_k(x) = x^t \Sigma^{-1} \mu_k - \frac{1}{2} \mu_k^t \Sigma^{-1} \mu_k + \log \pi_k$$

Linear Discriminant Analysis (LDA)

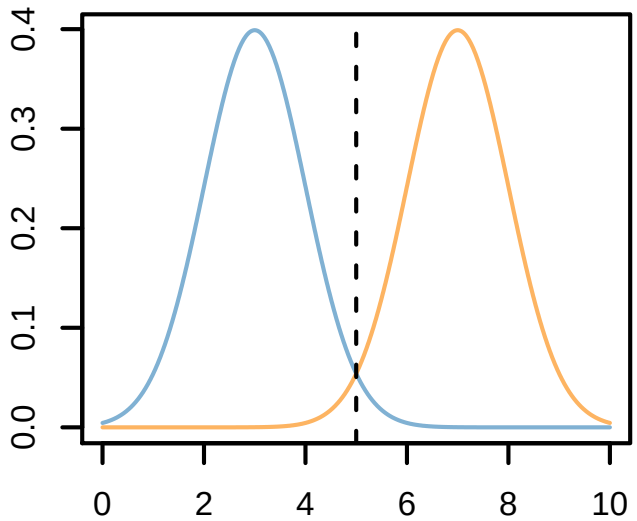
Example: Case $p = 1$, $K = 2$

- LDA assigns an observation $X = x$ to the class for which $\delta_k(x) = x$ is largest

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

- For $\pi_1 = \pi_2 = 0.5$, the decision boundary corresponds to the point where

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



Linear Discriminant Analysis (LDA)

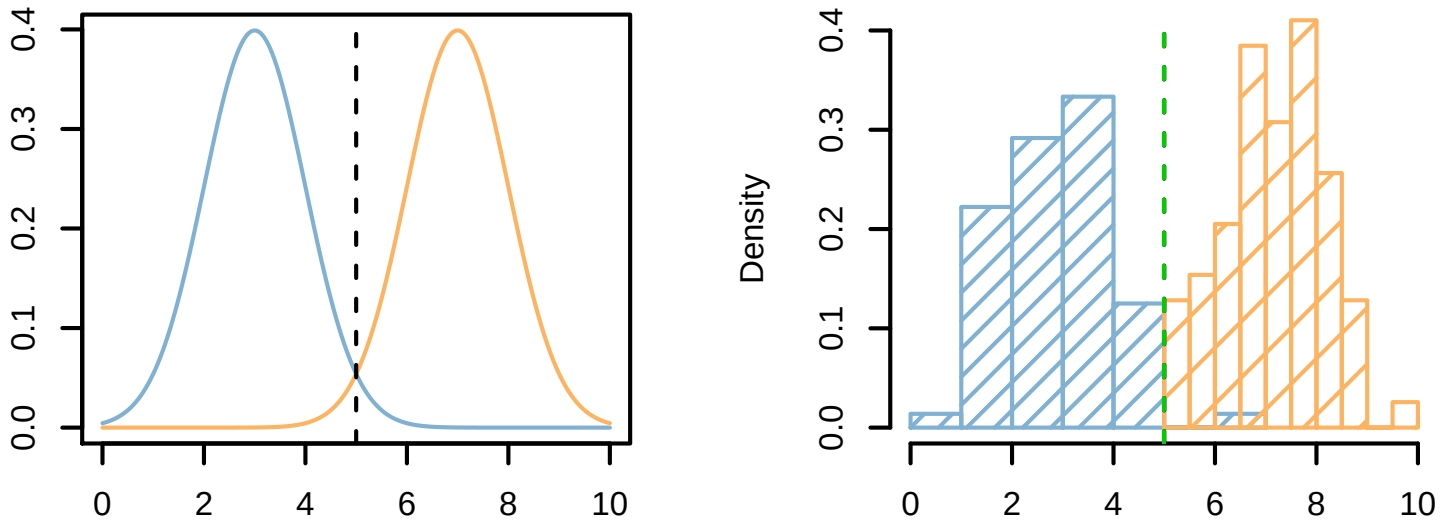
Example: Case $p = 1$, $K = 2$

- LDA assigns an observation $X = x$ to the class for which $\delta_k(x) = x$ is largest

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

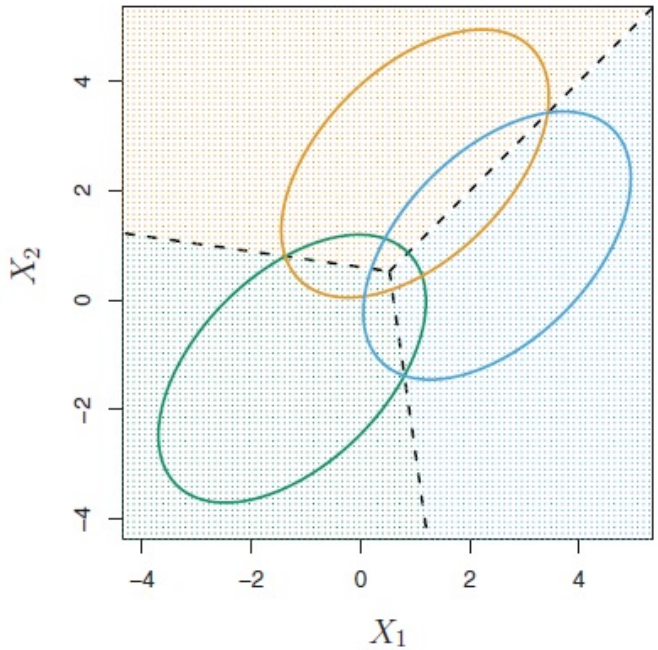
- In practice, we estimate π_k, μ_k and σ and assign an observation $X = x$ to the class for which $\hat{\delta}_k(x)$ is largest, where

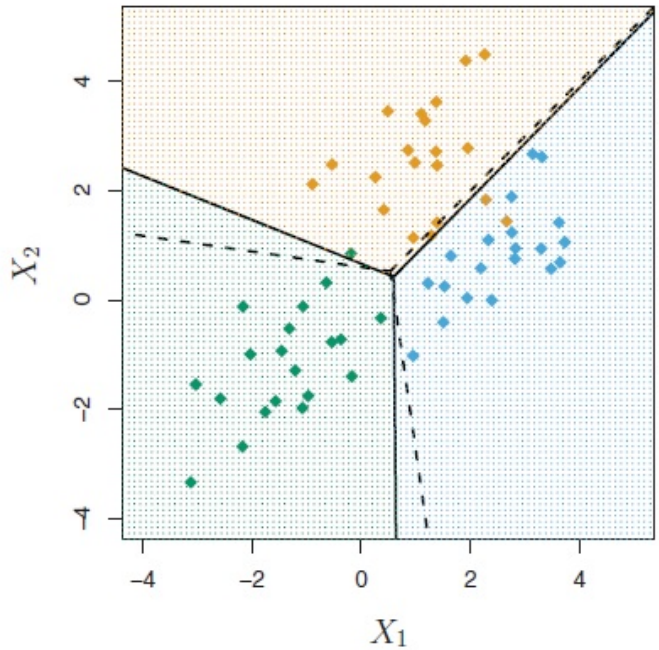
$$\hat{\delta}_k(x) = x\frac{\hat{\mu}_k}{\hat{\sigma}^2} - \frac{\hat{\mu}_k^2}{2\hat{\sigma}^2} + \log \hat{\pi}_k$$



Linear Discriminant Analysis (LDA)

Example: Case $p = 2$, $K = 3$





Quadratic Discriminant Analysis (LDA)

- Suppose that we observe a qualitative variable Y that can take on K possible values and predictor variables $X = (X_1 \ldots, X_p)$.
- Under Quadratic Discriminant Analysis (QDA) we assume that:
  - In each class k, the probability density function of X is multivariate normal with mean vector μ_k and covariance matrix Σ_k
  - The covariance matrix may be different for each population $\Sigma_1 \neq \cdots \neq \Sigma_K$.

$$f_k(x) = \frac{1}{(\sqrt{2\pi})^p \mid \Sigma_k \mid^{1/2}} e^{\frac{-(x-\mu_k)' {\Sigma_k}^{-1} (x-\mu_k)}{2}}, \; x \in \mathbb{R}^p, k = 1, \ldots, K$$

- Then, the Bayes classifier assigns an observation $X = x$ to the class for which $f_k(x)\pi_k$ is largest.
- It can be proved that the QDA rule assigns an observation $X = x$ to the class for which $\delta_k(x)$ is largest, where

$$-\frac{1}{2}(x-\mu_k)^t \Sigma_k^{-1}(x-\mu_k) - \frac{1}{2}\log(|\Sigma_k|) + \log(\pi_k).$$

LDA vs QDA

Example: training dataset with $K = 2$ (orange and blue dots). Classifiers: reality (purple, according to Bayes rule), LDA (black dotted), QDA (green).