

# Contents

- 1 Classification
  - Framework and notations

# Contents

## 1 Classification

- Framework and notations
- Confusion matrix

# Contents

## 1 Classification

- Framework and notations
- Confusion matrix
- ROC Curve

# Contents

## 1 Classification

- Framework and notations
- Confusion matrix
- ROC Curve
- Logistic Regression motivation

# Contents

- 1 Classification
  - Framework and notations
  - Confusion matrix
  - ROC Curve
  - Logistic Regression motivation
  - Logistic Regression algorithm

# Contents

## 1 Classification

- Framework and notations
- Confusion matrix
- ROC Curve
- Logistic Regression motivation
- Logistic Regression algorithm
- Other approach K-Nearest Neighbours (KNN)

# Contents

## 1 Classification

- Framework and notations
- Confusion matrix
- ROC Curve
- Logistic Regression motivation
- Logistic Regression algorithm
- Other approach K-Nearest Neighbours (KNN)
- Summary

# Outline

## 1 Classification

- Framework and notations
  - Confusion matrix
  - ROC Curve
  - Logistic Regression motivation
  - Logistic Regression algorithm
  - Other approach K-Nearest Neighbours (KNN)
  - Summary



In fact, this is the same notation as in the linear models chapter, except that now  $Y_i$  is discrete and represent generally a class number, a qualitative variable  $Y_i \in \{0, 1, \dots, K\}$ ,  $Y_i \in \{\text{spam}, \text{ham}\}$

$$Y_i = h(X^i) = \beta_0 + \beta_1 X_{i,1} + \beta_2 X_{i,2} + \dots + \beta_p X_{i,p}$$

Moreover the prediction is a probability to be in the class  $k$ . In a two-class problem, you can decide  $Y_i$  is true if probability is  $> 0.5$

# Outline

## 1 Classification

- Framework and notations
- **Confusion matrix**
- ROC Curve
- Logistic Regression motivation
- Logistic Regression algorithm
- Other approach K-Nearest Neighbours (KNN)
- Summary

Vocabulary :

		Predicted			
		+	-		
Actual	+	TP	FN Type II error	Sensitivity (recall) TP/●	False negative rate FN/●
	-	FP Type I error	TN	False positive rate FP/●	Specificity TN/●
		Precision TP/■	False omission rate FN/■	Accuracy ( TP + TN )/( ● + ● )	
		FDR FP/■	Negative predictive value TN/■	$F_1$ score $2TP/(2TP + FP + FN)$	

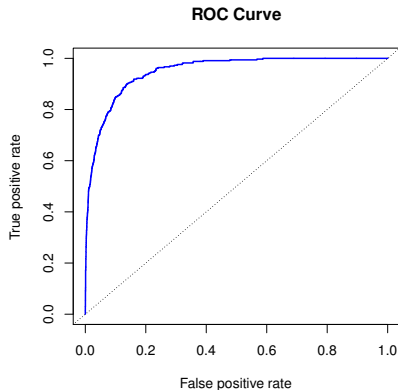
Beware : in the book Predicted is given horizontally

# Outline

## 1 Classification

- Framework and notations
- Confusion matrix
- ROC Curve
- Logistic Regression motivation
- Logistic Regression algorithm
- Other approach K-Nearest Neighbours (KNN)
- Summary

From communication theory Receiver Operator Curve :

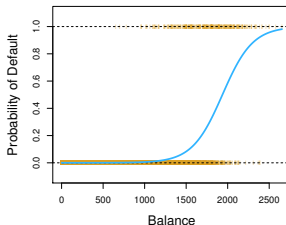
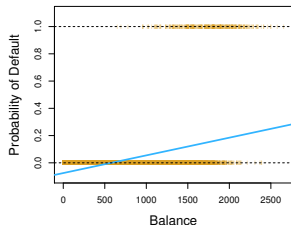


The associated measure is the Area Under the Curve (**AUC**), need to be near 1.

# Outline

## 1 Classification

- Framework and notations
- Confusion matrix
- ROC Curve
- **Logistic Regression motivation**
- Logistic Regression algorithm
- Other approach K-Nearest Neighbours (KNN)
- Summary



- To avoid values outside  $[0, 1]$ , we use a specific function called **sigmoid** or **logistic function**
- $\text{sigmoid}(x) = \frac{1}{1+e^{-x}} = \frac{e^x}{e^x+1} \in [0, 1]$
- The Linear Regression is "filtered" by this function.

# Outline

## 1 Classification

- Framework and notations
- Confusion matrix
- ROC Curve
- Logistic Regression motivation
- **Logistic Regression algorithm**
- Other approach K-Nearest Neighbours (KNN)
- Summary



- After some manipulation the model is :  
 $\ln\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X_1$  where  $p(X)$  is the probability that the response is 1
- Instead of using Least Square algorithm to adjust the  $\beta_i$  parameters, here we use a more statistical way, the Maximum Likelihood algorithm
- Obviously we can have multiple predictors :  
 $\ln\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_i X_p$

# Outline

- 1 Classification
  - Framework and notations
  - Confusion matrix
  - ROC Curve
  - Logistic Regression motivation
  - Logistic Regression algorithm
  - **Other approach K-Nearest Neighbours (KNN)**
  - Summary

- In order to make a prediction for an observation  $X = x$ , the  $K$  training observations that are closest to  $x$  are identified.
- Then  $X$  is assigned to the class to which the plurality of these observations belong. Hence KNN is a completely **non-parametric** approach : no assumptions are made about the shape of the decision boundary.
- There is also a KNN Regression algorithm (book)

# Outline

## 1 Classification

- Framework and notations
- Confusion matrix
- ROC Curve
- Logistic Regression motivation
- Logistic Regression algorithm
- Other approach K-Nearest Neighbours (KNN)
- Summary

- Prediction of probabilities ;
- Same minimization frameworks ;
- In general, all the algorithms have a Regression and a Classification behaviour.