Binary Classification

Mathias Bourel

IMERL - Facultad de Ingeniería, Universidad de la República, Uruguay

June 8, 2021

Bayes classifier

We focus on binary classification
Suppose we have two population groups P_1 y P_2 where elements of P_1 are labelled by 1 and elements of P_2 by 0. Indeed if $\mathbf{X}$ and Y are both random quantities:

$$(\mathbf{X}, Y) \in P_1, \text{with } \mathbf{X} \in \mathcal{X} \subset \mathbb{R}^d \text{and } Y = 1$$

$$(\mathbf{X}, Y) \in P_2, \text{with } \mathbf{X} \in \mathcal{X} \subset \mathbb{R}^d \text{ and } Y = 0$$

The goal is to construct a classifier $F : \mathcal{X} \to \{0, 1\}$:

$$F(\mathbf{x}) = \mathbb{1}_{\{f(\mathbf{x})>0\}}$$

where f is the boundary between two classes, that can be linear or not. If it is linear we call f a *linear classification rule*.
Let define by:

1. $\pi_1 = \mathbb{P}(Y = 1)$ and $\pi_2 = \mathbb{P}(Y = 0)$ with $\pi_1 + \pi_2 = 1$ the *marginal distribution* of Y (prior).
2. $\mathbf{X}|Y = 1 \sim g_1$ and $\mathbf{X}|Y = 0 \sim g_2$ the *conditional density* of $\mathbf{X}$ given Y (we suppose a distribution for the two populations).

Bayes Classifier

Then, applying the well known Bayes formula and recalling that $\mathbb{P}(\mathbf{X} = \mathbf{x} | Y = 1) = g_1(\mathbf{x}) \Delta\, x$:

- The density of $\mathbf{X}$ is the mixture:

$$g(\mathbf{x}) = \pi_1 g_1(\mathbf{x}) + \pi_2 g_2(\mathbf{x})$$

- The conditional distribution of Y given $\mathbf{X}$ are

$$\mathbb{P}(Y = 1 | \mathbf{X} = \mathbf{x}) = \frac{\pi_1 g_1(\mathbf{x})}{g(\mathbf{x})}$$

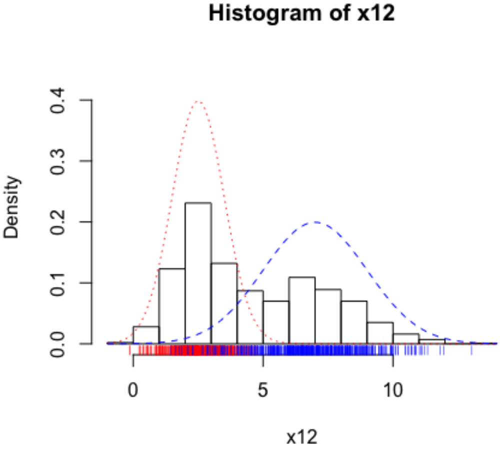
$$\mathbb{P}(Y = 0 | \mathbf{X} = \mathbf{x}) = 1 - \mathbb{P}(Y = 1 | \mathbf{X} = \mathbf{x}) = \frac{\pi_2 g_2(\mathbf{x})}{g(\mathbf{x})}$$

Example

```
pi= .5 # prior
n =1000 #size of the sample

population <- sample(1:0, n, rep = TRUE,
             prob = c(pi, 1- pi))
table(population)
n1 <- table(population)["1"]; n2 <- table(population)["0"]

# parameters
mu.1 <- 2.5; sigma.1 <- 1
mu.2 <- 7; sigma.2 <- 2
x1 <- rnorm(n1, mu.1, sigma.1)
x2 <- rnorm(n2, mu.2, sigma.2)
x12 <- c(x1, x2)

mean(x1); mean(x2); mean(x12)
sd(x1); sd(x2); sd(x12)
```

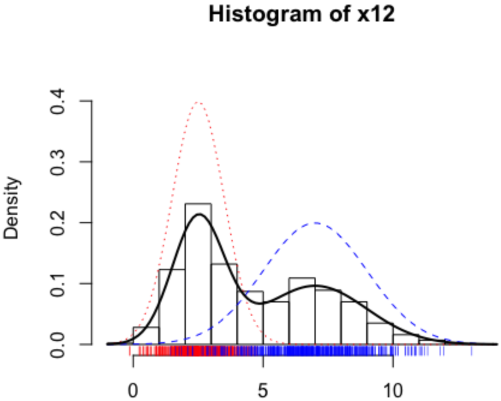
Example

```
hist(x1, freq = F, ylim=c(0,0.45)); curve(dnorm(x, mean = mu.1, sd = sigma.1), add = T, lwd = 2, col = 'red')
hist(x2, freq = F, ylim=c(0,0.45)); curve(dnorm(x, mean = mu.2, sd = sigma.2), add = T, lwd = 2, col = 'blue')
hist(x12, freq = F, ylim=c(0,0.45))
rug(x1, col = 'red')
rug(x2, col = 'blue')
curve(dnorm(x, mean = mu.1, sd = sigma.1), lty = 3, add = T, col = 'red')
curve(dnorm(x, mean = mu.2, sd = sigma.2), lty = 2, add = T, col = 'blue')
```



Example

```
#Density:
g.mixture <- function(x, pi, mu, sigma) {
  g <- pi * dnorm(x, mu[1], sigma[1]) +
       (1 - pi) * dnorm(x, mu[2], sigma[2])
return(g)
}
curve(g.mixture(x, pi = n1/n, c(mu.1, mu.2),
               c(sigma.1, sigma.2)), lwd = 2, add = T)
```



Bayes Classifier

Then a new observation x_0 is classified as following. As

$$\mathbb{P}(\mathbf{Y}=1|\mathbf{X}=\mathbf{x_0})=\frac{\mathbb{P}(\mathbf{X}=\mathbf{x_0}|\mathbf{Y}=1)\pi_1}{\pi_1\mathbb{P}(\mathbf{X}=\mathbf{x_0}|\mathbf{Y}=1)+\pi_2\mathbb{P}(\mathbf{X}=\mathbf{x_0}|\mathbf{Y}=0)}=\frac{\pi_1 g_1(\mathbf{x_0})}{g(\mathbf{x_0})}$$

we classify x_0 in the group with the maximum posterior probability (Bayes rule), indeed if

$$\pi_2 g_2(\mathbf{x_0}) > \pi_1 g_1(\mathbf{x_0})$$

we assign $\mathbf{x_0}$ to group 2, i.e $y_0 = 0$

In the particular case that $\pi_1 = \pi_2$ we assign $\mathbf{x_0}$ to group 2 if

$$g_2(\mathbf{x_0}) > g_1(\mathbf{x_0})$$

Any binary classifier function F can be tested under $0-1$ loss by its risk as follow:

$$R(F)=\mathbb{E}_{(\mathbf{X},\mathbf{Y})}\mathbb{1}_{\{\mathbf{Y}\neq F(\mathbf{X})\}}=\mathbb{P}(\mathbf{Y}\neq F(\mathbf{X}))$$

The Bayes rule is

$$F^*(\mathbf{x_0})=\begin{cases}1 & \text{if}\,\mathbb{P}(\mathbf{Y}=1|\mathbf{X}=\mathbf{x_0})>\mathbb{P}(\mathbf{Y}=0|\mathbf{X}=\mathbf{x_0})\\ 0 & \text{if}\,\mathbb{P}(\mathbf{Y}=1|\mathbf{X}=\mathbf{x_0})<\mathbb{P}(\mathbf{Y}=0|\mathbf{X}=\mathbf{x_0})\end{cases}$$

and it is proved by $F^* = \underset{F}{\text{Argmin}} R(F)$.

The classification boundary of Bayes rule is the set

$$\{\mathbf{x}:\mathbb{P}(\mathbf{Y}=1|\mathbf{X}=\mathbf{x})=\mathbb{P}(\mathbf{Y}=0|\mathbf{X}=\mathbf{x})\}$$

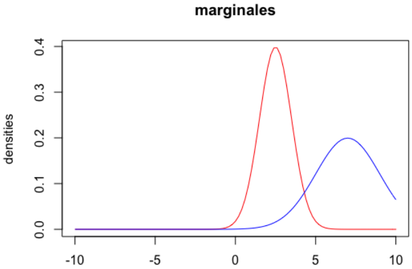
Example

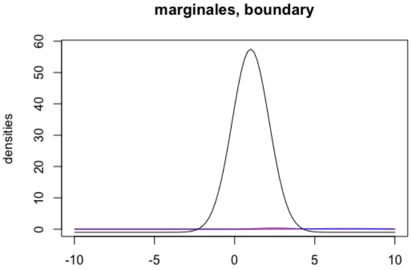
```
x.test <- sample(c(0, 1, 3.5, 8, 10, 12))
clasificacion <- ifelse(pi * dnorm(x.test, mu.1, sigma.1) >
(1 - pi) * dnorm(x.test, mu.2, sigma.2), ’Group 1’, ’Group 2’)
cbind(x.test, poblacion = clasificacion)
```

```
    x.test poblacion
[1,] "8"    "Group 2"
[2,] "3.5"  "Group 1"
[3,] "0"    "Group 1"
[4,] "10"   "Group 2"
[5,] "1"    "Group 1"
[6,] "12"   "Group 2"
```

Example

```
mu.1 <- 2.5; sigma.1 <- 1
mu.2 <- 7; sigma.2 <- 2
x1 <- rnorm(n1, mu.1, sigma.1)
x2 <- rnorm(n2, mu.2, sigma.2)

curve(dnorm(x, mean = mu.1, sd = sigma.1),xlim=c(-10,10), lty = 1, col = ’red’,
ylab=’densities’,main=’marginales’)
curve(dnorm(x, mean = mu.2, sd = sigma.2), lty = 1,add=T, col = ’blue’)
boundary=function(x)
{dnorm(x,mu.1,sigma.1)/dnorm(x,mu.2,sigma.2)-1
}
curve(boundary(x), lty = 1,add=T, col = ’black’)
```





Bayes classifier and classification boundary

As

$$\mathbb{P}(\mathbf{Y}=1|\mathbf{X}=\mathbf{x_0})=\frac{\pi_1 g_1(\mathbf{x_0})}{g(\mathbf{x_0})} \quad \text{and} \quad \mathbb{P}(\mathbf{Y}=0|\mathbf{X}=\mathbf{x_0})=\frac{\pi_2 g_2(\mathbf{x_0})}{g(\mathbf{x_0})}$$

the Bayes decision boundary is

$$\left\{\mathbf{x} : \frac{g_1(\mathbf{x})}{g_2(\mathbf{x})}=\frac{\pi_2}{\pi_1}\right\}$$

Any binary classifier divides the input space $\mathcal{X}$ as $\mathcal{X}=R_1 \cup R_0$ where

$$R_1=\{\mathbf{x}\in\mathcal{X} : F(\mathbf{x})=1\} \qquad \text{and} \qquad R_0=\{\mathbf{x}\in\mathcal{X} : F(\mathbf{x})=0\}$$

```
#Equal prior
boundary=function(x)
{dnorm(x,mu.1,sigma.1)/dnorm(x,mu.2,sigma.2)-1
}

library(rootSolve) #required by the function uniroot.all
raices <- uniroot.all(boundary,c(-100,100))

raices
[1] -2.293689  4.293689
```

So the optimal classification regions are

$$R_1^* = (-2.293, 4.293) \quad \text{and} \quad R_0^* = (-\infty, -2.293) \cup (4.293, +\infty)$$

Bayes classifier and classification boundary

Suppose we have observing a difference at the prior. How does this affect the classification rule? If $\pi_1 = 0.25$ and $\pi_2 = 0.75$, the Bayes decision boundary is

$$\left\{ \mathbf{x} : \frac{g_1(\mathbf{x})}{g_2(\mathbf{x})} = \frac{\pi_2}{\pi_1} \right\} = \left\{ \mathbf{x} : \frac{g_1(\mathbf{x})}{g_2(\mathbf{x})} = 3 \right\}$$

```
boundary=function(x)
{dnorm(x,mu.1,sigma.1)/dnorm(x,mu.2,sigma.2)-3
}

library(rootSolve) #required by the function uniroot.all
raices <- uniroot.all(boundary,c(-100,100))

raices
[1] -1.814038  3.814038
```

Classification error and cost

For any decision function we can make two possible mistakes:

- assign to class 1 an observation when its true label is 0 (false positive)
- assign to class 0 an observation when its true label is 1 (false negative)

We denote by

- $C(1,0)$ the cost of misclassifying an observation of class 1 to 0
- $C(0,1)$ the cost of misclassifying an observation of class 0 to 1

We assume that $C(i,i) = 0 \, \forall \, i$ and $C(i,j) \geq 0 \, \forall i,j$.

We can think about the expected cost risk of classifying an instance $\mathbf{x}$ in class 1 as

$$\mathbb{R}(1|\mathbf{X}=\mathbf{x}) = \sum_j \mathbb{P}(j|\mathbf{X}=\mathbf{x})C(j,1) = \mathbb{P}(0|\mathbf{X}=\mathbf{x})C(0,1)$$

Then the classifier takes the decision of classifying $\mathbf{x}$ in the positive class, if the risk of classifying $\mathbf{x}$ in the negative class is more important of classifying in the positive class:

$$F^*(x) = 1 \Leftrightarrow \mathbb{R}(1|\mathbf{X}=\mathbf{x}) \leq \mathbb{R}(0|\mathbf{X}=\mathbf{x}) \ \Leftrightarrow \ \mathbb{P}(0|\mathbf{X}=\mathbf{x})C(0,1) \leq \mathbb{P}(1|\mathbf{X}=\mathbf{x})C(1,0)$$

$$\Leftrightarrow \frac{\mathbb{P}(1|\mathbf{X}=\mathbf{x})}{\mathbb{P}(0|\mathbf{X}=\mathbf{x})} > \frac{C(0,1)}{C(1,0)}$$

As $\mathbb{P}(0|\mathbf{X}=\mathbf{x}) = 1 - \mathbb{P}(1|\mathbf{X}=\mathbf{x})$ we get a threshold p to assign class 1 to x if

$$\mathbb{P}(1|\mathbf{X}=\mathbf{x}) \geq p = \frac{C(0,1)}{C(1,0)+C(0,1)}$$

Classification error and cost

So

$$F^*(x) = \begin{cases} 1 & \text{if } \frac{\mathbb{P}(\mathbf{Y}=1|\mathbf{X}=\mathbf{x})}{\mathbb{P}(\mathbf{Y}=0|\mathbf{X}=\mathbf{x})} > \frac{C(0,1)}{C(1,0)} \\ 0 & \text{if } \frac{\mathbb{P}(\mathbf{Y}=1|\mathbf{X}=\mathbf{x})}{\mathbb{P}(\mathbf{Y}=0|\mathbf{X}=\mathbf{x})} < \frac{C(0,1)}{C(1,0)} \end{cases}$$

Remark: if $C(0,1) >> C(1,0)$, threshold $p \approx 1$ and the classification is usually 0. And if $C(0,1) << C(1,0)$, threshold $p \approx 0$ and the classification is usually 1.

The boundary is the set

$$\left\{x : \frac{\mathbb{P}(\mathbf{Y}=1|\mathbf{X}=\mathbf{x})}{\mathbb{P}(\mathbf{Y}=0|\mathbf{X}=\mathbf{x})} = \frac{C(0,1)}{C(1,0)}\right\} = \left\{x : \frac{g_1(\mathbf{x})}{g_2(\mathbf{x})} = \frac{\pi_2 C(0,1)}{\pi_1 C(1,0)}\right\}$$

In example above if we assume that $C(0,1) = 2$ and $C(1,0) = 1$, $\pi_1 = \pi_2$ the Bayes boundary bound is

$$\left\{\mathbf{x} : \frac{g_1(\mathbf{x})}{g_2(\mathbf{x})} = 2\right\}$$

```
boundary=function(x)
{dnorm(x,mu.1,sigma.1)/dnorm(x,mu.2,sigma.2)-2
}

library(rootSolve) #required by the function uniroot.all
raices <- uniroot.all(boundary,c(-100,100))

raices
[1] -2  4
```

Bayes classifier, error and cost

In a certain sense we can think:

Prediction \ Reality	Positive 1	Negative 0
Positive 1	True Positive (TP) $C(1,1)$	False Positive (FP) $C(0,1)$ Type II error
Negative 0	False Negative (FN) $C(1,0)$ Type I error	True Negative (TN) $C(0,0)$

Table: Confusion matrix

and

$$\mathbb{P}(1|\mathbf{X}=\mathbf{x}) \geq p = \frac{C(0,1)}{C(1,0)+C(0,1)} = \frac{FP}{FN+FP}$$

Arguments above are very important when we deal with:

- Imbalanced problems: the two groups are not equally represented in the dataset. For example if we suppose we have observing a rare event (a disease for example) that occurs among 5% of the population. If observations of class 1 are infected and observations of class 2 are healthy, we have $\pi_1 = 0.1$ and $\pi_2 = 0.9$.
- The misclassification error is not the same for both class. If we want to give a treatment to the population, it is more serious to say that a patient is healthy when it has a disease, than the contrary.

Linear Regression models

As $\mathbb{E}(\mathbf{Y}|\mathbf{X}=\mathbf{x})=\mathbb{P}(\mathbf{Y}=1|\mathbf{X}=\mathbf{x})$ we suppose that is linear of the form $\beta_0+\mathbf{x}'\beta_{\mathbf{1}}$
We recall that the Minimum Least Square estimator of $\beta=(\beta_{\mathbf{0}},\beta_{\mathbf{1}}')'$ is

$$\beta^{LS}=\underset{\beta}{\text{Argmin}}||\mathbf{y}-X\beta||^2=\underset{\beta}{\text{Argmin}}(\mathbf{y}-X\beta)'(\mathbf{y}-X\beta)$$

where $\mathbf{y}=(y_1,\dots,y_n)$ with $y_i\in\{0,1\}$ and X is the data matrix with all elements of the first column equal 1.
It is well known that in most of case: $\beta^{LS}=(X'X)^{-1}X'\mathbf{y}$ and the classification rule is

$$\widehat{f}(\mathbf{x})=\beta_0^{LS}+\mathbf{x}'\beta_{\mathbf{1}}{}^{LS}-0.5$$

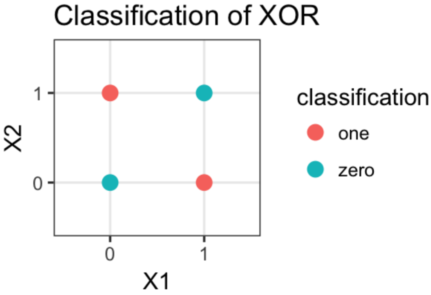
and the prediction is

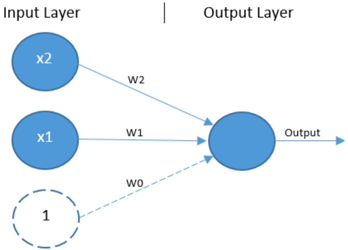
$$\widehat{y}=\widehat{F}(\mathbf{x})=\begin{cases}1 & \text{if } \widehat{f}(\mathbf{x})>0\\ 0 & \text{if } \widehat{f}(\mathbf{x})<0\end{cases}$$

Remarks:

- Very simple
- Low variance, but much bias
- suppose a linear boundary and $\widehat{f}(\mathbf{x})$ should be a probability....

Not all is linear





Not all is linear

- for $(0,0)$ and $(1,1)$ we have $f(x,w,w_0) = w'x + w_0 < 0$
- for $(1,0)$ and $(0,1)$ we have $f(x,w,w_0) = w'x + w_0 \geq 0$

Then

$$\begin{cases} 0w_1 + 0w_2 + w_0 < 0 & (1) \\ w_1 + w_2 + w_0 < 0 & (2) \\ w_1 + w_0 \geq 0 & (3) \\ w_2 + w_0 \geq 0 & (4) \end{cases}$$

Then $(3) + (4) - (2) \Rightarrow w_0 \geq 0$ which is absurd. Then the data is non linearly separable.

ROC curve

Suppose that the population consists of individuals who have a tumor, which can be malignant or benign. It is clear that the rule $p(\mathbf{x}) = \mathbb{P}(Y = 1|\mathbf{X} = \mathbf{x}) > 0.5$ then $Y = 1$ is not appropriate.

- The *sensibility* is the graphic of curve $Se(t) = \mathbb{P}(p(x) > t|Y = 1), 0 \leq t \leq 1$
  Varying t, the curve of Se gives proportion of individuals to whom malignancy is detected. For $t = 0$ all individuals would be malignant, and for t $=$ 1 all would be benign. This is the *True Positive Rate*.
- The *specificity* is the graphic of curve $Sp(t) = \mathbb{P}(p(x) < t|Y = 0), 0 \leq t \leq 1$
  Varying t, the curve of Sp gives proportion of individuals to whom a benign tumor is detected. For $t = 0$ all individuals would be benign, and for $t = 1$ all would be malignant. It is a major problem in medical diagnosis to determine the cut-off value such that it detects the greatest number of malignant tumors, without committing too many errors (deciding that it is malignant when in fact it is benign).

The ROC (Receiving Operating Characteristic) curve summarizes the two sensitivity and specificity curves. It is the curve that results from representing the points

$$\big(1 - Sp(t); Se(t)\big) \quad 0 \leq t \leq 1$$

that is, 1-Specificity (False Positive Rate) on the OX axis, and Sensitivity (True Positive Rate) on the OY axis.

- The ROC curve is not necessary above the diagonal and convex, but it is monotone, and the more it moves away from the diagonal, the better the discrimination ($TPR = f(FPR)$).
- For $t = 0$ we have $TPR = FPR = 1$ and for $t = 1$ we have $TPR = FPR = 0$. As t decreases, TPR and FPR increase.

ROC curve

Prediction \ Reality	Positive 1	Negative 0
Positive 1	True Positive (TP)	False Positive (FP)
Negative 0	False Negative (FN)	True Negative (TN)
	P	N

Table: Confusion matrix

For $t = 0$:

Prediction \ Reality	Positive 1	Negative 0
Positive 1	P	N
Negative 0	0	0

For $t = 1$:

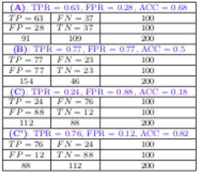
Prediction \ Reality	Positive 1	Negative 0
Positive 1	0	0
Negative 0	P	N

For $t = t^*$ (ideal):

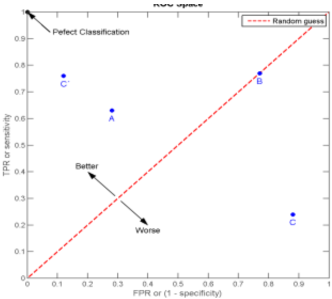
Prediction \ Reality	Positive 1	Negative 0
Positive 1	P	0
Negative 0	0	N

ROC Curve

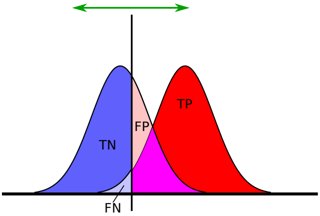
- Accuracy: $\frac{TP+TN}{TP+FN+FP+TN} = \frac{TP+TN}{P+N}$
- Sensibility (True Positive Rate); $\frac{TP}{TP+FN} = \frac{TP}{P}$ (OY axis)
- Specificity (True Negative Rate): $\frac{TN}{FP+TN} = \frac{TN}{N}$
- 1-Specificity (False Positive Rate): $\frac{FP}{FP+TN} = \frac{FP}{N}$ (OX axis)



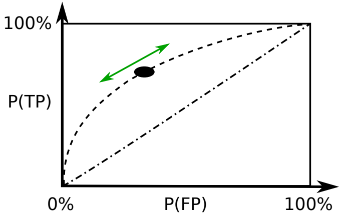
(A): TPR = 0.63, FPR = 0.28, ACC = 0.68		
$TP = 63$	$FN = 37$	100
$FP = 28$	$TN = 37$	100
91	109	200
(B): TPR = 0.77, FPR = 0.77, ACC = 0.5		
$TP = 77$	$FN = 23$	100
$FP = 77$	$TN = 23$	100
154	46	200
(C): TPR = 0.24, FPR = 0.88, ACC = 0.18		
$TP = 24$	$FN = 76$	100
$FP = 88$	$TN = 12$	100
112	88	200
(C'): TPR = 0.76, FPR = 0.12, ACC = 0.82		
$TP = 76$	$FN = 24$	100
$FP = 12$	$TN = 88$	100
88	112	200



ROC Curve



TP	FP
FN	TN



We choose the optimal t such that ROC point at s is the nearest of $(0,1)$, i.e minimizing $(1-Se(t))^2+(1-Sp(t))^2$. However:

- it ignores the predicted probability values and the goodness-of-fit of the model
- it summarizes the test performance over regions of the ROC space in which one would rarely operate
- it weights omission and commission errors equally

Referencias

- D. Peña, *Analisis de Datos Multivariantes*, Mac Graw Hill, 2002.
- G. James, D. Witten, T. Hastie, R. Tibshirani, *An Introduction to Statistical Learning with Applications in R*, Springer, 2013.
- T. Hastie, R. Tibshirani, Friedman. The Elements of Statistical Learning, Springer, 2003.