TP Clasificación Estadística

Camilo D'Aloisio, Kenet Chapetón y Agustín Arias

2022-12-08

Teórico

Ejercicio 1

Por definición tenemos

$$L(g) = P(g(\mathbf{X}) \neq Y) = E_{(\mathbf{X},Y)}\left[I_{\{g(\mathbf{X})\neq Y\}}\right]$$

Luego, por propiedades de la esperanza condicional, tenemos que

$$\begin{aligned} E_{(\mathbf{X},Y)}\left[I_{g(\mathbf{X}\neq Y)}\right] &= E_{\mathbf{X}}\left[E_{Y|\mathbf{X}}\left[I_{\{g(\mathbf{X}\neq Y)\}}\right]\right] \\ &= E_{\mathbf{X}}\left[I_{\{g(\mathbf{X})\neq 1\}}P(Y=1|\mathbf{X}=\mathbf{x}) + I_{\{g(\mathbf{X})\neq 0\}}P(Y=0|\mathbf{X}=\mathbf{x})\right] \end{aligned}$$

Entonces para minimizar $L(g)$ es suficiente con minimizar el argumento de la esperanza, puesto que es una función creciente

$$I_{\{g(\mathbf{X})\neq 1\}} \cdot P(Y=1|\mathbf{X}=\mathbf{x}) + I_{\{g(\mathbf{X})\neq 0\}} \cdot P(Y=0|\mathbf{X}=\mathbf{x}) \qquad (\forall \mathbf{x})$$

Como los argumentos de las indicadoras determinan conjuntos disjuntos, es decir, como

$$\{\omega: g\big(\mathbf{X}(\omega)\big) \neq 1\} \cap \{\omega: g\big(\mathbf{X}(\omega)\big) \neq 0\} = \emptyset$$

Tenemos que g^{op} queda determinada como

$$g^{op}(\mathbf{x}) = \begin{cases} 1 & si \quad P(Y=1|\mathbf{X}=\mathbf{x}) \geq P(Y=0|\mathbf{X}=\mathbf{x}) \\ 0 & si \quad P(Y=1|\mathbf{X}=\mathbf{x}) < P(Y=0|\mathbf{X}=\mathbf{x}) \end{cases}$$

Ejercicio 2:

Por Bayes y multiplicando y dividiendo por $P(Y=y)$ tenemos

$$P(Y=y|\mathbf{X}=\mathbf{x}) = \frac{P(\mathbf{X}=\mathbf{x}, Y=y)}{P(\mathbf{X}=\mathbf{x})} \cdot \frac{P(Y=y)}{P(Y=y)} = \frac{P(\mathbf{X}=\mathbf{x}|Y=y)\cdot P(Y=y)}{P(\mathbf{X}=\mathbf{x})}$$

luego reemplazando en

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

nos queda

$$\frac{P(\mathbf{X} = \mathbf{x}|Y = 1) \cdot P(Y = 1)}{P(\mathbf{X} = \mathbf{x})} \geq \frac{P(\mathbf{X} = \mathbf{x}|Y = 0) \cdot P(Y = 0)}{P(\mathbf{X} = \mathbf{x})}$$

$$P(\mathbf{X} = \mathbf{x}|Y = 1) \cdot P(Y = 1) \geq P(\mathbf{X} = \mathbf{x}|Y = 0) \cdot P(Y = 0)$$

Por supuesto todo tiene sentido siempre y cuando

$$P(\mathbf{X} = \mathbf{x}) \neq 0$$

u equivalentemente no sea

$$\omega: \mathbf{X}(\omega) = \mathbf{x}$$

un conjunto de medida nula.

Además, teniendo en cuenta que

$$\mathbf{X}|Y = 1 \sim f_1 \quad \mathbf{X}|Y = 0 \sim f_0 \quad \pi_1 = P(Y = 1) \quad \pi_0 = P(Y = 0)$$

Obtenemos

$$g^{op}(\mathbf{x}) = \begin{cases} 1 & si \quad f_1(\mathbf{x}) \cdot \pi_1 \geq f_0(\mathbf{x}) \cdot \pi_0 \\ 0 & si \quad f_1(\mathbf{x}) \cdot \pi_1 < f_0(\mathbf{x}) \cdot \pi_0 \end{cases}$$

Ejercicio 3:

Como tenemos que

$$\mathbf{X}|Y = 1 \sim \mathcal{N}(\mu_1, \Sigma_1) \quad \mathbf{X}|Y = 0 \sim \mathcal{N}(\mu_0, \Sigma_0)$$

Luego, reemplazando en

$$f_1(\mathbf{x}) \cdot \pi_1 \geq f_0(\mathbf{x}) \cdot \pi_0$$

queda

$$\frac{e^{-1/2 \cdot (\mathbf{x}-\mu_1)^t \Sigma_1^{-1} (\mathbf{x}-\mu_1)}}{(2\pi)^{p/2} det(\Sigma_1)^{1/2}} \cdot \pi_1 \geq \frac{e^{-1/2 \cdot (\mathbf{x}-\mu_0)^t \Sigma_0^{-1} (\mathbf{x}-\mu_0)}}{(2\pi)^{p/2} det(\Sigma_0)^{1/2}} \cdot \pi_0$$

donde $0 \leq \pi_i, det(\Sigma_i)$, puesto que π_i son probabilidades y Σ_i son definidas positivas $\forall i \in \{0,1\}$.

Tenemos entonces

$$e^{-1/2\cdot[(\mathbf{x}-\mu_1)^t\Sigma_1^{-1}(\mathbf{x}-\mu_1)-(\mathbf{x}-\mu_0)^t\Sigma_0^{-1}(\mathbf{x}-\mu_0)]} \geq \sqrt{\frac{det(\Sigma_1)}{det(\Sigma_0)}}\frac{\pi_0}{\pi_1}$$

Tomando ln a ambos miembros (es una función creciente), nos queda

$$-1/2\cdot[(\mathbf{x}-\mu_1)^t\Sigma_1^{-1}(\mathbf{x}-\mu_1)-(\mathbf{x}-\mu_0)^t\Sigma_0^{-1}(\mathbf{x}-\mu_0)] \geq (1/2)ln\left(\frac{det(\Sigma_1)}{det(\Sigma_0)}\right)+ln\left(\frac{\pi_0}{\pi_1}\right)$$

Finalmente, reemplazamos los

$$r_i(\mathbf{x}) = (\mathbf{x}-\mu_i)^t\Sigma_i^{-1}(\mathbf{x}-\mu_i)$$

y multiplicamos por 2 a ambos miembros $\forall i \in \{0,1\}$.

$$-r_1(\mathbf{x})+r_0(\mathbf{x}) \geq ln\left(\frac{det(\Sigma_1)}{det(\Sigma_0)}\right)+2ln\left(\frac{\pi_0}{\pi_1}\right)$$

$$r_0(\mathbf{x})-ln\left(\frac{det(\Sigma_1)}{det(\Sigma_0)}\right)-2ln\left(\frac{\pi_0}{\pi_1}\right) \geq r_1(\mathbf{x})$$

$$r_0(\mathbf{x})+ln\left(\frac{det(\Sigma_0)}{det(\Sigma_1)}\right)+2ln\left(\frac{\pi_1}{\pi_0}\right) \geq r_1(\mathbf{x})$$

Por lo tanto

$$g^{op}(\mathbf{x}) = \begin{cases} 1 & si \quad r_0(\mathbf{x})+ln(\frac{det(\Sigma_0)}{det(\Sigma_1)})+2ln(\frac{\pi_1}{\pi_0}) \geq r_1(\mathbf{x}) \\ 0 & en\ c.c \end{cases}$$

Práctico

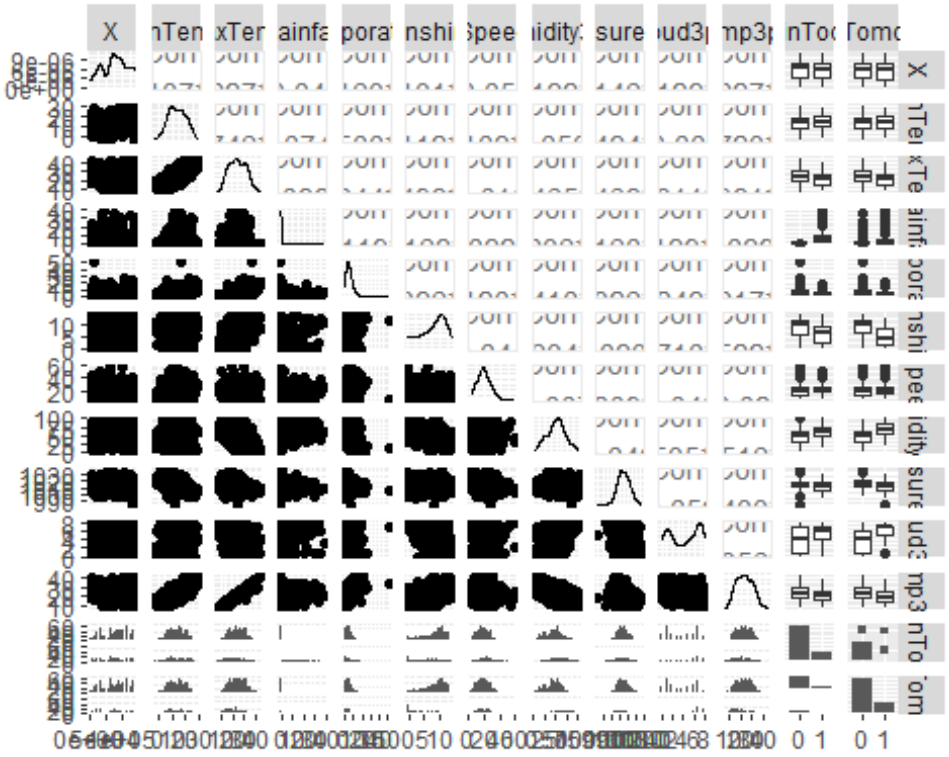
Cargamos los datos y modificamos las variables RainTomorrow y RainToday para que queden como categóricas y con las opciones 1 (si llovió) y 0 (si no llovió). Las otras variables son numéricas.

```
setwd("C:\\Users\\camil\\OneDrive\\Desktop\\2DO 2022\\ESTADÍSTICA")
datos <- read.csv("lluviaAus.csv")
datos$RainTomorrow <-  ifelse(datos$RainTomorrow=="Yes", 1, 0)
datos$RainTomorrow <- as.factor(datos$RainTomorrow)
datos$RainToday <-  ifelse(datos$RainToday=="Yes", 1, 0)
datos$RainToday <- as.factor(datos$RainToday)
```

Ejercicio 1:

```
library(ggplot2)
library(GGally)

ggpairs(datos)
```

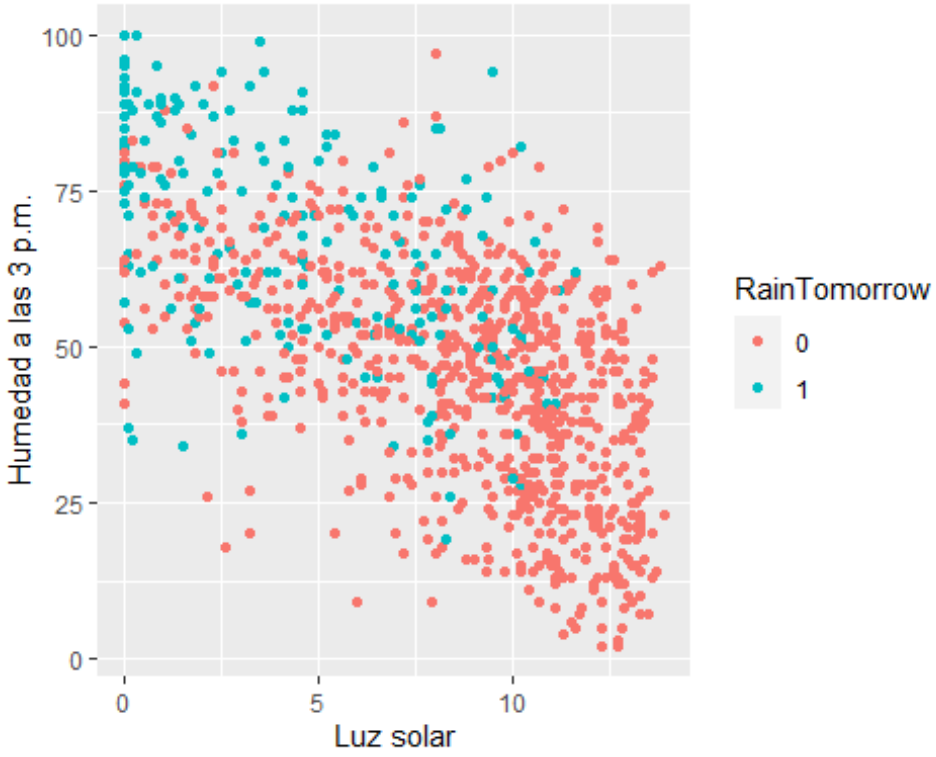


Las varaibles temperatura máxima, temperatura mínima y temperatura a las 3 de la tarde están bastante correlacionadas entre sí.

Ejercicio 2:

```
library(ggplot2)

datos %>% ggplot(mapping=aes(x=Sunshine, y=Humidity3pm,
color=RainTomorrow)) +
  geom_point() +
  labs(x="Luz solar", y="Humedad a las 3 p.m.")
```

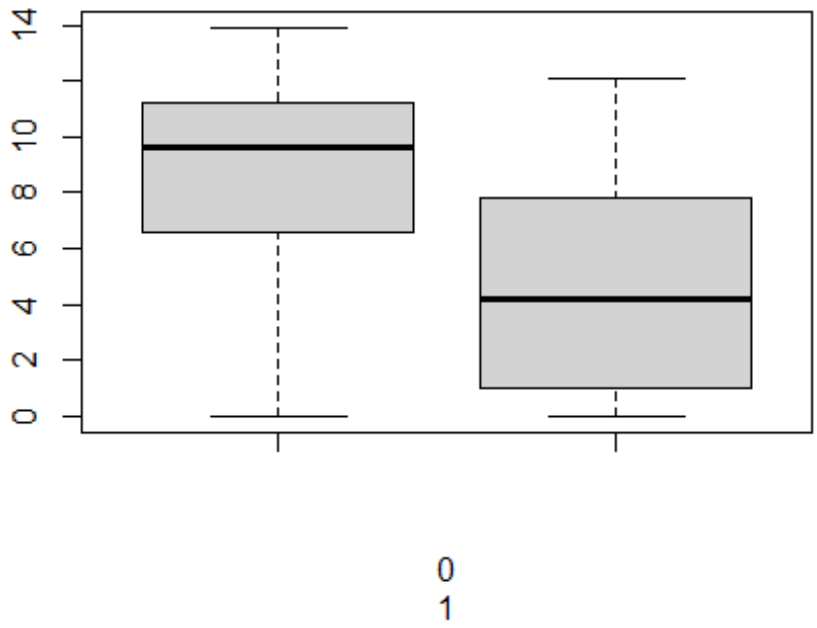


Cuando alguna de las dos variables toma valores altos ambas predicen razonablemente si va a llover o no. Si la humedad a las 3 p.m. es alta, es probable que al otro día llueva, mientras que si la luz solar es alta, es probable que al otro día no llueva.

Sin embargo, cuando la luz solar toma valores bajos no indica si al otro día va a llover o no, mientras que si la humedad a las 3pm es baja, es probable que no llueva al otro día.

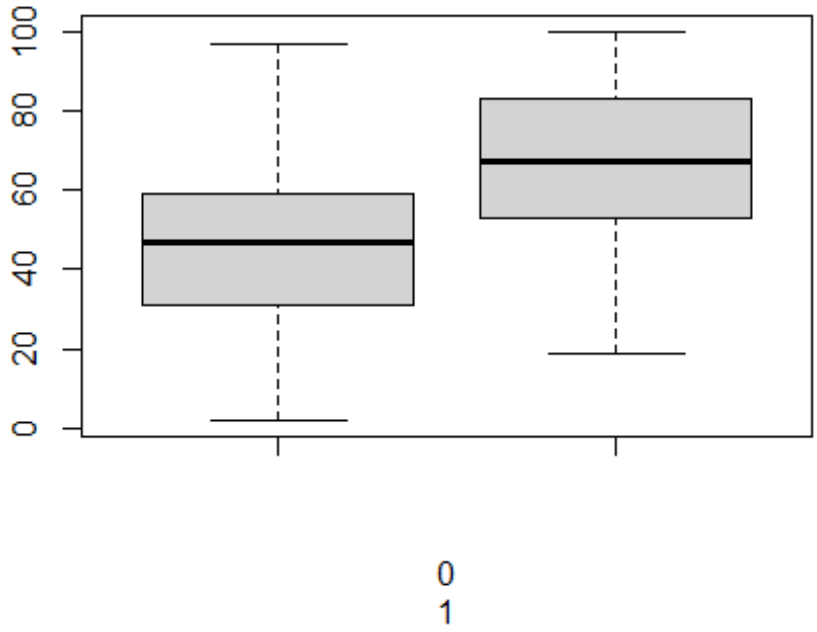
Ejercicio 3:

```
boxplot(datos$Sunshine[datos$RainTomorrow==0],
datos$Sunshine[datos$RainTomorrow==1], xlab=c(0, 1))
```



Observamos que, en lineas generales, cuando al otro día no lloverá hay más luz solar que cuando lloverá, ya que todos los cuantiles son mayores. Sin embargo, el bigote inferior se encuentra al mismo nivel (en 0) por lo que puede haber días en los que al siguiente no lloverá pero donde haya poca luz solar, tal como vimos también en el ejercicio anterior.

```
boxplot(datos$Humidity3pm[datos$RainTomorrow==0],
datos$Humidity3pm[datos$RainTomorrow==1], xlab=c(0, 1))
```



Acá, al contrario, observamos que los días donde al día siguiente lloverá suele haber más humedad a las 3 de la tarde que los días que al día siguiente no lloverá.

Ejercicio 4:

```
clasificador.movil <-  function(datos, etiquetas, h, x0) {
  etiquetas_en_ventana <-  etiquetas[abs(datos - x0) <= h]
  res <- ifelse(sum(etiquetas_en_ventana==1) /
length(etiquetas_en_ventana) >= 0.5, 1, 0)
  return(res)
}
```

Ejercicio 5:

```
ventana_optima <-  function(datos, etiquetas, ventanas) {
  ECM_ventanas <-  c()
  etiquetas <- as.numeric(etiquetas==1)
  for (h in 1:length(ventanas)) {
    clasificaciones <-  c()
    for (j in 1:length(etiquetas)) {
      clasificaciones[j] <-  clasificador.movil(datos[-j], etiquetas[-j],
ventanas[h], datos[j])
    }
    ECM_ventanas[h] <-  mean((etiquetas - clasificaciones)^2)
  }
  index_res <- order(ECM_ventanas)[1]
  res <-  ventanas[index_res]
  return(res)
}
```

Ejercicio 6:

Primero defimos la función

```
error_de_clasificacion <-  function(datos, etiquetas, ventana) {
  errores <-  c()
  etiquetas <- as.numeric(etiquetas==1)
  for (i in 1:length(etiquetas)) {
    clasificacion_i <-  clasificador.movil(datos[-i], etiquetas[-i],
ventana, datos[i])
    error <-  abs(etiquetas[i] - clasificacion_i)
    errores[i] <-  error
  }
  return(mean(errores))
}
```

Ahora vamos a evaluarla

```
ventana <- ventana_optima(datos$Sunshine, datos$RainTomorrow, seq(0.01,
2, length.out=200))

error_de_clasificacion(datos$Sunshine, datos$RainTomorrow, ventana)

## [1] 0.178
```

Luego el error es 0.178