Guia 26

Agustin Muñoz Gonzalez

25/7/2020

Preparamos el entorno.

```
rm(list=ls())
library(ggplot2)
library(tidyr)
library(gganimate)
```

El objetivo de esta práctica es la implementación de reglas de clasificación teniendo en cuenta la selección de los parámetros de suavizado.

En un bosque de Bariloche hay dos variedades de hongos, que identificaremos como la variedad I y variedad II. En el archivo hongos clasificados.txt encontrará n = 500 registros correspondientes a la altura y variedad de cada uno los hongos examinados. A fin de clasificar un nuevo hongo de este bosque, implementaremos la regla de Bayes, pero sin suponer que las densidades condicionales involucradas en su cálculo pertenecen a una familia determinada.

Preliminar

Separamos un 10% de los datos elegidos al azar.

```
hongos=read.delim('hongos_clasificados.txt', header = TRUE, sep = " ", dec = ".")
indices=sample(length(hongos$Height),size=0.2*length(hongos$Height))
test_set=hongos[indices,]
training_set=hongos[-indices,]
```

En lo que sigue, llamaremos f_1 a la densidad de la altura de un hongo de la variedad I y f_0 a la densidad de la altura de un hongo de la variedad II.

1. A partir de los alturas medidas en los hongos de variedad I estime la función de densidad f_1 . Indique cómo determió la ventana y que núcleo usó. Llamemos $\widehat{f}_{1,h_1}$ a la estimación resultante de la función de densidad f_1.

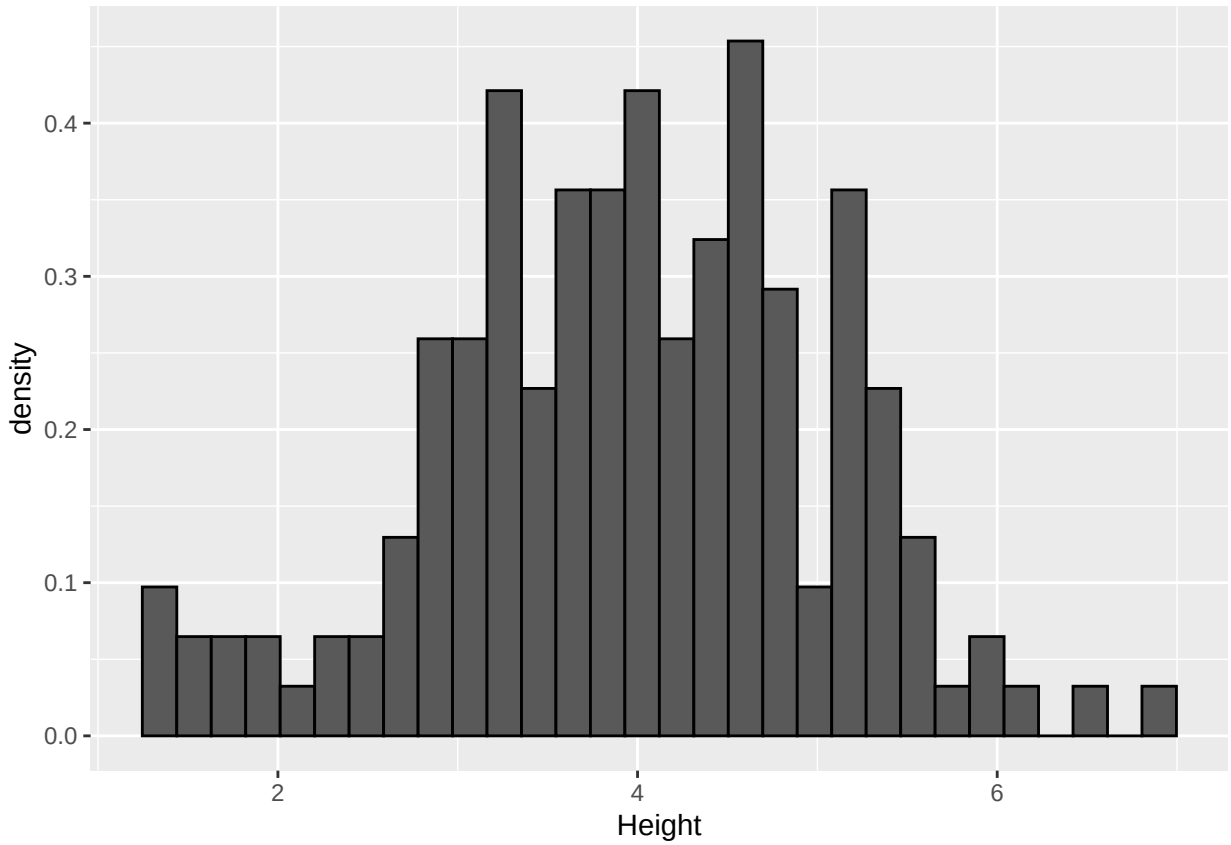
Resolución:

Como estamos suponiendo que f_0 y f_1 no pertenecen a ninguna familia determinada, vamos a hacer una estimación no paramétrica.

Veamos en primer lugar un histograma de cada variedad de los hongos para ver si a simple vista se observa alguna densidad conocida.

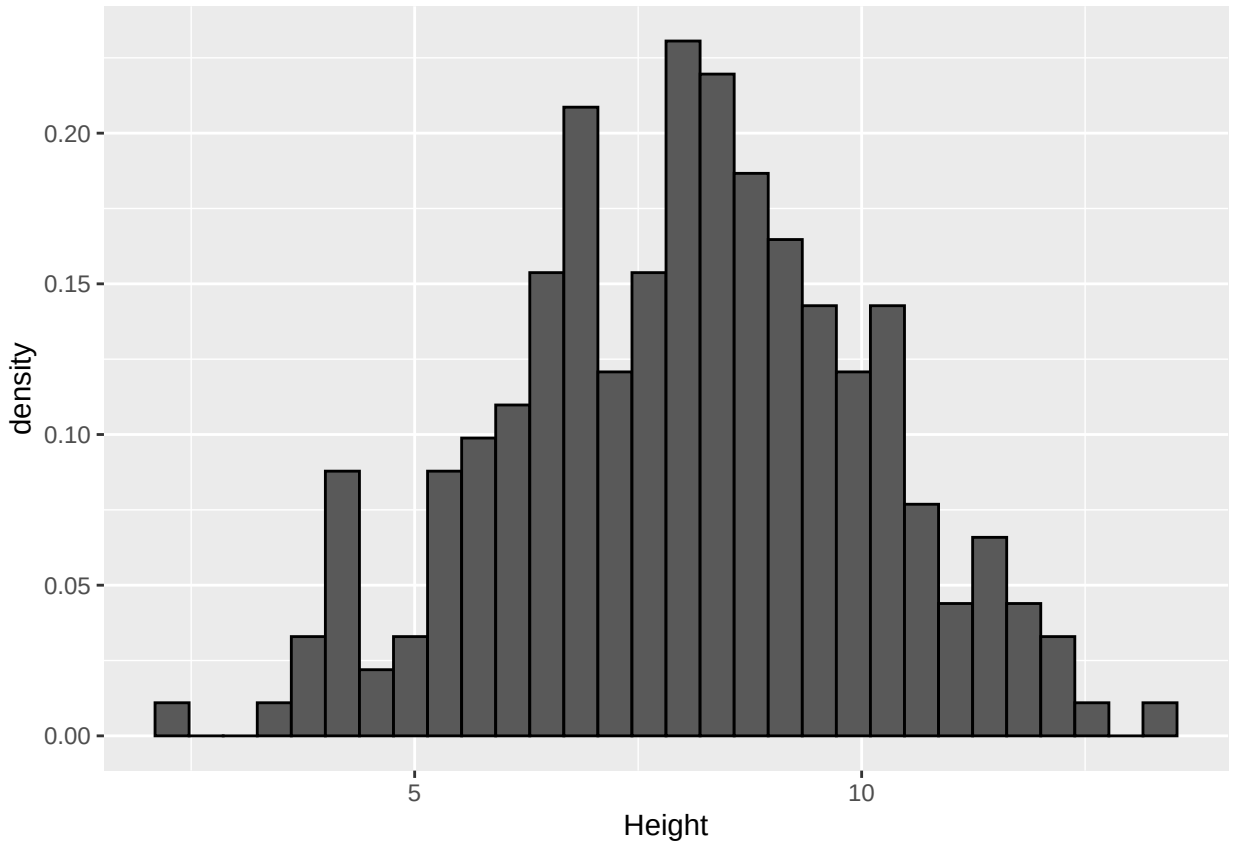
```
hongos_1=training_set[training_set$Variety==1,]
hongos_2=training_set[training_set$Variety==2,]
hongos_1 %>%
  ggplot()+
  geom_histogram(aes(x=Height,y=..density..),col='black')
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
hongos_2 %>%
  ggplot()+
  geom_histogram(aes(x=Height,y=..density..),col='black')
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Observamos que se parece mucho a la densidad normal.

Pero como por ahora no tenemos que suponer nada y como tenemos muchos datos procedemos a usar la estimación no parámetrica de la densidad por

- el estimador de Nadaraya-Watson con núcleo gaussiano;
- regresión polinomial de grado 10;
- k-nn;

Donde la ventana la vamos a elegir por el método de Convalidación Cruzada y elegiremos el estimador que tenga el menor error (error cuadratico medio ECM para los estimadores polinomiales, error cuadratico de prediccion promediado ECPP para el metodo NW y k-nn).

ELEGIR LA H A PARA MINIMIZAR LOG LIKELIHOOD USANDO CV

2. A partir de los alturas medidas en los hongos de variedad II estime la función de densidad f_0 . Indique cómo determinó la ventana y que núcleo usó. Llamemos $\widehat{f}_{0,h_0}$ a la estimación resultante de la función de densidad f_0.

Resolución:

3. Implemente una función **class.est.variedad** que determine la variedad de un hongo mediante la regla plug–in de Bayes $\widehat{g}$ basada en las estimaciones $\widehat{f}_{1,h_1}$ y $\widehat{f}_{0,h_0}$ ya obtenidas en los dos ítems anteriores y las proporciones de cada variedad en los datos registrados en el archivo hongos_clasificados.txt.

Resolución:

Usar la idea de la funcion siguiente de la entrega 4. DEFINIR f_0_h_0 y f_1_h_1

```
proporcion_uno=mean(training_set$Variety==1)
proporcion_dos=mean(training_set$Variety==2)
```

```
class.est.variedad=function(hongo){
  if(f_1_h_1(hongo)*proporcion_uno>f_0_h_0(hongo)*proporcion_2)
{1}else{2}
}
```

4. Calcule el Error de Clasificación Empírico de $\widehat{g}$ utilizando los datos del archivo hongos_clasificados.txt.

Resolución:

5. ¿Le parece que las ventanas halladas en a) y b) con las que implementó la regla de clasificación son las más adecuadas a los fines de la clasificación?

Implemente una función **class.optim.est.variedad** que determine la variedad de un hongo mediante la regla plug–in de Bayes $\widehat{g}$ basada en las estimaciones $\widehat{f}_{1,h_1}$ y $\widehat{f}_{0,h_0}$ y las proporciones de cada variedad en los datos registrados en el archivo hongos_clasificados.txt en la que las ventanas se determinan simultáneamente por Convalización Cruzada.

Resolución:

ES COMO LO QUE HICIMOS EN 1 Y 2 SOLO QUE AHORA BUSCO LAS H_0 H_1 SIMULT USANDO CV PERO AHORA LA FUNCION OBJETIVO A MINIMIZAR NO ES LA LOG LIKELIHOOD SINO EL ERROR DE CLASIFICACIÓN (QUIERO MEJORAR EL CLASIFICADOR)

6. Estime el Error de Clasificación de la regla de plug–in Bayes $\widehat{g}$ mediante el Error de Clasificación Empírico utilizando los datos del archivo hongos clasificados.txt, pero ahora implemente la regla asumiendo que las densidades f_1 y f_0 son normales y que desconoce sus parámetros. Compare con los resultados anteriores.

Resolución:

USAR LA VENTANA DE SILVERMAN PARA BUSCAR LA CONVALIDACION CRUZADA ALREDEDOR DE ESE VALOR

7. **Para entregar**: Implemente una función **class.nopar** que dado un punto x_{new} determine la clase a la que pertenece el nuevo individuo que tiene este valor en la covariable mediante la regla plug–in de Bayes $\widehat{g}$ basada en las estimaciones no paramétricas de las densidades f_1 y f_0 usando núcleo gaussiano. Para ello entrar como input de la función implementada el punto x_{new}, los vectores de datos X_{datos} e Y_{datos} y las ventanas h_1 y h_0 : class.nopar(x_{new}, X_{datos}, Y_{datos}, h_1, h_0).

Resolución: