KNN para el análisis de variables cualitativas (clasificación)

Contents

1 Predicción con un regresor **1**

2 Predicción con m regresores **2**

1 Predicción con un regresor

El objetivo es predecir los valores que tomará una **variable cualitativa** conocidos los valores de otra u otras variables, que pueden ser cuantitativas o cualitativas. Por ejemplo, se quiere predecir la especie de los pingüinos si se conoce la longitud de su aleta. La especie es la **variable respuesta** y la longitud de la aleta es el **regresor** (también se conoce como variable independiente, cofactor, predictor,...). Cuando la variable respuesta es cualitativa se suele denominar **problema de clasificación**.

```
## se leen los datos
d = read.csv("datos/pinguinos.csv", sep = ";")

# es conveniente convertir a factores las variables cualitativas
d$especie = factor(d$especie)
d$isla = factor(d$isla)
d$genero = factor(d$genero)
```

```
# por simplicidad se utilizan solo 10 datos seleccionados de manera aleatoria
set.seed(99)
pos = runif(10, min = 1, max = nrow(d))
# se seleccionan los 10 observaciones de la variable respuesta y el regresor
d1 = d[pos,c("especie","long_aleta")]
```

En este caso se quiere predecir la especie de un pingüino cuya aleta mide 210 mm. Para ello se va a utilizar el algoritmo **K-Nearest Neighbors (KNN)**. La idea es calcular la predicción como la especie más frecuente en los k-valores más cercanos a 210 mm. Primero se calcula la distancia al regresor que se va a predecir:

```
# se calcula la distancia y se guarda en d1
d1$dist = abs(d1$long_aleta - 210)
# se ordenan todas las observaciones de d1 en funcion de dist
(orden1 = sort(d1$dist, index.return = T))
```

```
## $x
##  [1]  3  5  5  8 13 14 20 24 24 26
##
## $ix
##  [1]  5  1  7  4  6  3  8  2  9 10
```

Ahora se ordenan los datos en función de la distancia al regresor predicho:

```
(d1s = d1[orden1$ix,])
```

```
##       especie long_aleta dist
## 178    Gentoo        207    3
```

```
## 195    Gentoo         215    5
## 223    Gentoo         215    5
## 330 Chinstrap         202    8
## 321 Chinstrap         197   13
## 228    Gentoo         224   14
## 98     Adelie         190   20
## 38     Adelie         186   24
## 119    Adelie         186   24
## 59     Adelie         184   26
```

Ya se puede calcular la predicción:

- Si K = 1 se utiliza el valor más cercano. Luego la predicción para la especie es Gentoo.
- Si K = 2 la predicción es Gentoo, ya que los dos puntos más cercanos son Gentoo. En caso de empate se elige al azar. Para evitar empates se suelen utilizar K impares.
- Si K = 3 la predicción es Gentoo, ya que los tres puntos más cercanos son Gentoo.
- Y así para otros valores de K. Por ejemplo, para K = 5, los 5 puntos más cercanos corresponden a 3 pingüinos Gentto y 2 Chinstrap, por lo que la predicción es Gentoo.

2 Predicción con m regresores

En el caso de que se tengan dos o más regresores el algoritmo sigue siendo el mismo: la predicción es la categoría más frecuente de los k valores más cercanos. Igual que ocurre en la predicción de variables cuantitativas, se utiliza la distancia euclídea para calcular la distancia entre cada observación y los datos a predecir. De igual manera, los regresores cuantitativos se tienen que normalizar y los regresores cualitativos se introducen en el algoritmo mediante la definición de variables auxiliares.

Para entender mejor el algoritmo se va a predecir la especie de un pingüino con long_pico = 40 mm, prof_pico = 18 mm, long_aleta = 210 mm, peso = 3500 g y genero = Hembra.

- Primero se normalizan los regresores cuantitativos:

```
# se lee la funcion que normaliza
source("funciones/knn_funciones.R")
```

```
# se seleccionan las 10 observaciones
d2 = d[pos,c("especie","long_pico","prof_pico","long_aleta","peso","genero")]
```

```
# se normalizan y se guardan
d2$long_pico1 = knn_normaliza(d2$long_pico)
d2$prof_pico1 = knn_normaliza(d2$prof_pico)
d2$long_aleta1 = knn_normaliza(d2$long_aleta)
d2$peso1 = knn_normaliza(d2$peso)
```

- Se definen las variables auxiliares:

```
#
d2$genero_H = ifelse(d2$genero == "hembra", 1, 0)
d2$genero_M = ifelse(d2$genero == "macho", 1, 0)
```

- Se preparan los valores que se quieren predecir:

```
# variables cuantitativas normalizadas
xp2 = data.frame(
  long_pico = knn_normaliza(40, min(d$long_pico), max(d$long_pico)),
  prof_pico = knn_normaliza(18, min(d$prof_pico), max(d$prof_pico)),
  long_aleta = knn_normaliza(210, min(d$long_aleta), max(d$long_aleta)),
  peso = knn_normaliza(3500, min(d$peso), max(d$peso)),
```

```
  genero_H = 0,
  genero_M = 1
)
```

- Y ahora se calculan las distancias y se ordena:

```
d2$dist = 0
for (ii in 1:10){
  xii = d2[ii,7:12]
  d2$dist[ii] = knn_dist(xii,xp2)
}
orden2 = sort(d2$dist, index.return = T)
(d2s = d2[orden2$ix,])
```

```
##       especie long_pico prof_pico long_aleta peso genero long_pico1 prof_pico1
## 98     Adelie      37.8      20.0        190 4250  macho 0.15294118  1.0000000
## 321 Chinstrap      52.2      18.8        197 3450  macho 1.00000000  0.8064516
## 195    Gentoo      45.2      15.8        215 5300  macho 0.58823529  0.3225806
## 228    Gentoo      50.0      15.9        224 5350  macho 0.87058824  0.3387097
## 330 Chinstrap      43.5      18.1        202 3400 hembra 0.48823529  0.6935484
## 38     Adelie      36.0      18.5        186 3100 hembra 0.04705882  0.7580645
## 59     Adelie      36.4      17.1        184 2850 hembra 0.07058824  0.5322581
## 119    Adelie      35.2      15.9        186 3050 hembra 0.00000000  0.3387097
## 223    Gentoo      45.2      13.8        215 4750 hembra 0.58823529  0.0000000
## 178    Gentoo      45.1      14.5        207 5050 hembra 0.58235294  0.1129032
##     long_aleta1 peso1 genero_H genero_M      dist
## 98        0.150  0.56        0        1 0.7415207
## 321       0.325  0.24        0        1 0.8123314
## 195       0.775  0.98        0        1 0.8659912
## 228       1.000  1.00        0        1 1.0638250
## 330       0.450  0.22        1        0 1.4457526
## 38        0.050  0.10        1        0 1.5671913
## 59        0.000  0.00        1        0 1.5854863
## 119       0.050  0.08        1        0 1.5860360
## 223       0.775  0.76        1        0 1.6544499
## 178       0.575  0.88        1        0 1.6570512
```

- Si K = 1 la predicción para el peso es Adelie.
- Si K = 2 la predicción es Adelie o Chinstrap. Se elige una al azar.
- Si K = 3 la predicción es Adelie o Chinstrap o Gentoo. Se elige una al azar.
- Si K = 4 la predicción es Gentoo.
- Y así sucesivamente.