Laboratorio 3, Tópicos en análisis datos 1

Joshua Isaac Cervantes Artavia

2023-09-06

1

Discretización de variables cuantitativas

a- Se forman tres clases

```r
# Se fija la semilla con la que se generan los número aleatorios
set.seed(654)

X <- rnorm(15) # Vector normal

# Se crean los intervalos abiertos por la izquierda y cerrados por la derecha
X_intervalos <- c(X, breaks = c(min(X) - 1e-10, -0.2, 0.2, max(X)))

# Se muestran los intervalos
X_intervalos
```

```
-0.76031762 -0.38970450  1.68962523 -0.09423560  0.09530146  0.81727228

 1.06576755  0.93984563  0.74121222 -0.43531214 -0.10726012 -0.83816833
                                         breaks1     breaks2     breaks3
-0.98260589 -0.82037099 -0.87143256 -0.98260589 -0.20000000  0.20000000
    breaks4
 1.68962523
```

b- Se forman clases a partir de cuantiles

```r
# Cuantiles
corte <- quantile(X, probs = seq(0, 1, by = 0.25))
```

```r
X_intervalos_2 <- cut(X, breaks = corte, include.lowest = TRUE)

table(X_intervalos_2)
```

```
X_intervalos_2
[-0.983,-0.79] (-0.79,-0.107] (-0.107,0.779]   (0.779,1.69]
             4              4              3              4
```

Con lowest se toma en el primer intervalo el menor cerrado, de tal forma que es el mínimo.

2

```r
# Se crea la tabla de datos
df_notas_escolares <- (data.frame(
    Estudiante =
        c(
            "Lucía",
            "Pedro",
            "Inés",
            "Luis",
            "Andrés",
            "Ana",
            "Carlos",
            "José",
            "Sonia",
            "María"
        ),
    Mate = c(7, 7.5, 7.6, 5.0, 6.0, 7.8, 6.3, 7.9, 6.0, 6.8),
    Cien = c(6.5, 9.4, 9.2, 6.5, 6, 9.6, 6.4, 9.7, 6, 7.2),
    Espa = c(9.2, 7.3, 8, 6.5, 7.8, 7.7, 8.2, 7.5, 6.5, 8.7),
    Hist = c(8.6, 7, 8, 7, 8.9, 8, 9, 8, 5.5, 9),
    EdFi = c(8, 7, 7.5, 9, 7.3, 6.5, 7.2, 6, 8.7, 7),
    Peso_lbs = c(126, 140, 130, 150, 142, 128, 144, 134, 135, 128),
    Estatura_cm = c(162, 168, 169, 172, 165, 165, 170, 165, 170, 166)
))

df_notas_escolares_solo_notas <- df_notas_escolares[, -c(7, 8)]
```

a- Se estima el centro de gravedad

```r
# Centro de gravedad
(g <- apply(df_notas_escolares_solo_notas[, -1], 2, mean))
```

```
Mate Cien Espa Hist EdFi
6.79 7.65 7.74 7.90 7.42
```

b-

```r
# Funcion para estimar la inercia I(N)
fn_inercia <- function(df_datos, a, M, pesos = 0) {
    n <- nrow(df_datos)
    pesos <- pesos * (pesos != 0) + rep(1 / n, n) * (pesos == 0)
    # Donde se suma la inercia
    I_N <- 0
    for (i in 1:n) {
        difference <- as.matrix((df_datos[i, ] - a))

        I_N <- I_N + pesos[i] * (difference) %*% M %*% t(difference)
    }
    return(I_N)
}



# Valores de estudiantes
a <- df_notas_escolares_solo_notas[df_notas_escolares_solo_notas$Estudiante %in% c(
    "Lucía",
    "Andrés",
    "Sonia"
), ]

# Metrica
M <- diag(1, 5)


# Lucia
fn_inercia(df_notas_escolares_solo_notas[, -1], a[1, -1], M = M)
```

```
       1
1 10.046
```

```r
# Andres
fn_inercia(df_notas_escolares_solo_notas[, -1], a[2, -1], M = M)
```

```
       1
1 10.086
```

```r
# Sonia
fn_inercia(df_notas_escolares_solo_notas[, -1], a[3, -1], M = M)
```

```
       1
1 18.004
```

c-

```r
# Inercia
fn_inercia(df_notas_escolares_solo_notas[, -1], g, M = M)
```

```
       1
1 5.7214
```

d-

```r
df_notas_escolares_solo_notas_centradas <- df_notas_escolares_solo_notas
(df_notas_escolares_solo_notas_centradas[, -1] <- df_notas_escolares_solo_notas[, -1]
    - matrix(rep(g, 10), nrow = 10, byrow = TRUE))
```

```
    Mate  Cien  Espa Hist  EdFi
1   0.21 -1.15  1.46  0.7  0.58
2   0.71  1.75 -0.44 -0.9 -0.42
3   0.81  1.55  0.26  0.1  0.08
4  -1.79 -1.15 -1.24 -0.9  1.58
5  -0.79 -1.65  0.06  1.0 -0.12
6   1.01  1.95 -0.04  0.1 -0.92
7  -0.49 -1.25  0.46  1.1 -0.22
8   1.11  2.05 -0.24  0.1 -1.42
9  -0.79 -1.65 -1.24 -2.4  1.28
10  0.01 -0.45  0.96  1.1 -0.42
```

e-

```
# Se centran las variables


#Centro de gravedad de las variables centradas
(g_centradas <- apply(df_notas_escolares_solo_notas_centradas[,-1], 2, mean))
```

```
         Mate          Cien          Espa          Hist          EdFi 
-8.883732e-17 -4.440892e-16 -4.440892e-16 -3.552605e-16  0.000000e+00 
```

```
#Se estima la inercia de estas variable centradas
fn_inercia(df_notas_escolares_solo_notas_centradas[, -1], g_centradas, M = M)
```

```
       1
1 5.7214
```

Se puede observar que se sigue manteniendo la misma inercia. Es decir el centras las variables no afecta a la misma.

f-

Se emplea la métrica como la diagonal de la división de las varianzas

```
# Metrica de inversa de varianzas
D_s <- diag(apply(df_notas_escolares_solo_notas[, -1], 2, function(x) {
    1 /
        (var(x) * (length(x) - 1) / (length(x)))
}))

fn_inercia(df_notas_escolares_solo_notas[, -1], g, M = D_s)
```

```
  1
1 5
```

g-

Se emplea la métrica de Mahalanobis

```
M_halanobis <- solve((nrow(df_notas_escolares) - 1) /
    nrow(df_notas_escolares) *
    cov(df_notas_escolares_solo_notas[, -1]))
```

```r
fn_inercia(df_notas_escolares_solo_notas[, -1], g, M = M_halanobis)
```

```
  1
1 5
```

3

a-

```r
# Se estandarizan las variables
df_notas_escolares_solo_notas_centradas_est <- df_notas_escolares_solo_notas_centradas
(df_notas_escolares_solo_notas_centradas_est[, -1] <-
    df_notas_escolares_solo_notas_centradas_est[, -1] / (
        (sqrt((nrow(df_notas_escolares_solo_notas_centradas_est) - 1) /
            nrow(df_notas_escolares_solo_notas_centradas_est))) * matrix(rep(
            apply(df_notas_escolares_solo_notas[, -1], 2, sd),
            10
        ), nrow = 10, byrow = TRUE)))
```

```
          Mate       Cien        Espa        Hist        EdFi
1   0.23263076 -0.7529862  1.78848525  0.65792263  0.65858084
2   0.78651352  1.1458486 -0.53899555 -0.84590053 -0.47690337
3   0.89729007  1.0148944  0.31849737  0.09398895  0.09083874
4  -1.98290027 -0.7529862 -1.51898747 -0.84590053  1.79406505
5  -0.87513476 -1.0803715  0.07349939  0.93988948 -0.13625811
6   1.11884317  1.2768027 -0.04899960  0.09398895 -1.04464547
7  -0.54280510 -0.8184633  0.56349535  1.03387842 -0.24980653
8   1.22961972  1.3422797 -0.29399757  0.09398895 -1.61238758
9  -0.87513476 -1.0803715 -1.51898747 -2.25573474  1.45341979
10  0.01107766 -0.2946468  1.17599030  1.03387842 -0.47690337
```

b-

Se calcula la inercia con la métrica identidad

```r
g_cent_std <- apply(df_notas_escolares_solo_notas_centradas_est[, -1], 2, mean)

fn_inercia(df_notas_escolares_solo_notas_centradas_est[, -1], g_cent_std, M = M)
```

```
  1
1 5
```

Se puede observar que el resultado es el mismo que con Mahalanobis y la métrica invera de las varianzas.

c-

```
correlaciones_materias <- cor(df_notas_escolares_solo_notas_centradas_est[, -1])

materias <- colnames(df_notas_escolares_solo_notas_centradas_est[, -1])
materia_mas_corr <- materias[1]

for (i in materias[-1]) {
    if (sum(abs(correlaciones_materias[, i]) >
        abs(correlaciones_materias[, materia_mas_corr])) >= 3) {
        materia_mas_corr <- i
    }
}

materia_mas_corr
```

```
[1] "Mate"
```

La variable que está más correlacionada con todas las demás es matemática.

4

```
ponderacion <- c(4, 4, 3, 3, 1)

M_ponderacion <- diag(ponderacion / sum(ponderacion))


(inercia_ponderacion <- fn_inercia(df_notas_escolares_solo_notas_centradas_est[, -1], g_cent_st
```

```
  1
1 1
```

Se obtiene que la inercia es de 1.