Unidad 2 parametrizado

Amelia Martinez Sequera

Índice de contenidos:

1. Tabla fortalezas/debilidades.

Fortalezas	Debilidades
Simple y efectivo	No produce un modelo, limitando la capacidad para comprender cómo las características están relacionados con la clase
No hace suposiciones sobre la distribución de datos subyacente	Requiere la selección de un k apropiado
Fase de entrenamiento rápida	Características nominales y datos faltantes requieren procesamiento adicional.
	Fase de clasificación lenta.

2. Importación de los datos.

```
library(readr)
wbcd <- read.csv("C:/Users/Meli/Downloads/wisc_bc_data.csv", stringsAsFactors = FALSE)
View(wbcd)
str(wbcd)
```

```
## 'data.frame':    569 obs. of  32 variables:
```

```
##  $ id               : int  87139402 8910251 905520 868871 9012568 906539 925291 87880 862989 89827 .
##  $ diagnosis        : chr  "B" "B" "B" "B" ...
##  $ radius_mean      : num  12.3 10.6 11 11.3 15.2 ...
##  $ texture_mean     : num  12.4 18.9 16.8 13.4 13.2 ...
##  $ perimeter_mean   : num  78.8 69.3 70.9 73 97.7 ...
##  $ area_mean        : num  464 346 373 385 712 ...
##  $ smoothness_mean  : num  0.1028 0.0969 0.1077 0.1164 0.0796 ...
##  $ compactness_mean : num  0.0698 0.1147 0.078 0.1136 0.0693 ...
##  $ concavity_mean   : num  0.0399 0.0639 0.0305 0.0464 0.0339 ...
##  $ points_mean      : num  0.037 0.0264 0.0248 0.048 0.0266 ...
##  $ symmetry_mean    : num  0.196 0.192 0.171 0.177 0.172 ...
##  $ dimension_mean   : num  0.0595 0.0649 0.0634 0.0607 0.0554 ...
##  $ radius_se        : num  0.236 0.451 0.197 0.338 0.178 ...
##  $ texture_se       : num  0.666 1.197 1.387 1.343 0.412 ...
##  $ perimeter_se     : num  1.67 3.43 1.34 1.85 1.34 ...
##  $ area_se          : num  17.4 27.1 13.5 26.3 17.7 ...
##  $ smoothness_se    : num  0.00805 0.00747 0.00516 0.01127 0.00501 ...
##  $ compactness_se   : num  0.0118 0.03581 0.00936 0.03498 0.01485 ...
##  $ concavity_se     : num  0.0168 0.0335 0.0106 0.0219 0.0155 ...
##  $ points_se        : num  0.01241 0.01365 0.00748 0.01965 0.00915 ...
##  $ symmetry_se      : num  0.0192 0.035 0.0172 0.0158 0.0165 ...
##  $ dimension_se     : num  0.00225 0.00332 0.0022 0.00344 0.00177 ...
##  $ radius_worst     : num  13.5 11.9 12.4 11.9 16.2 ...
##  $ texture_worst    : num  15.6 22.9 26.4 15.8 15.7 ...
##  $ perimeter_worst  : num  87 78.3 79.9 76.5 104.5 ...
##  $ area_worst       : num  549 425 471 434 819 ...
##  $ smoothness_worst : num  0.139 0.121 0.137 0.137 0.113 ...
##  $ compactness_worst: num  0.127 0.252 0.148 0.182 0.174 ...
##  $ concavity_worst  : num  0.1242 0.1916 0.1067 0.0867 0.1362 ...
##  $ points_worst     : num  0.0939 0.0793 0.0743 0.0861 0.0818 ...
##  $ symmetry_worst   : num  0.283 0.294 0.3 0.21 0.249 ...
##  $ dimension_worst  : num  0.0677 0.0759 0.0788 0.0678 0.0677 ...
```

3. Exploración de los datos.

```r
#Eliminamos la columna id
wbcd <- wbcd[-1]
#tabla de frecuencias absolutas de diagnósticos
table(wbcd$diagnosis)
```

```
##
##   B   M
## 357 212
```

```r
#Renombramos los factores
wbcd$diagnosis<- factor(wbcd$diagnosis, levels = c("B", "M"),
labels = c("Benign", "Malignant"))
#Tabla de frecuencias relativas de diagnósticos
round(prop.table(table(wbcd$diagnosis)) * 100, digits = 1)
```

```
##
```

```
##    Benign Malignant
##      62.7      37.3
```

4. Normalización de los datos.

```r
#Observamos que las variables numéricas tienen rangos muy diferentes
summary(wbcd[c("radius_mean", "area_mean", "smoothness_mean")])
```

```
##   radius_mean       area_mean      smoothness_mean
##  Min.   : 6.981   Min.   : 143.5   Min.   :0.05263
##  1st Qu.:11.700   1st Qu.: 420.3   1st Qu.:0.08637
##  Median :13.370   Median : 551.1   Median :0.09587
##  Mean   :14.127   Mean   : 654.9   Mean   :0.09636
##  3rd Qu.:15.780   3rd Qu.: 782.7   3rd Qu.:0.10530
##  Max.   :28.110   Max.   :2501.0   Max.   :0.16340
```

```r
#Creamos una función para reescalar los datos numéricos.
normalize <- function(x) {
return ((x-min(x)) / (max(x)-min(x)))
}
```

```r
#Creamos un data frame con los datos numéricos normalizados
wbcd_n <- as.data.frame(lapply(wbcd[2:31], normalize))
```

```r
#Comprobamos que los datos ahora están normalizados, con valores que van del 0 al 1
summary(wbcd_n$area_mean)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.1174  0.1729  0.2169  0.2711  1.0000
```

5. Training y Test datasets.

```r
#Creamos los Training and test datasets
wbcd_train <- wbcd_n[1:469, ]
wbcd_test <- wbcd_n[470:569, ]

#Labels
wbcd_train_labels <- wbcd[1:469, 1]
wbcd_test_labels <- wbcd[470:569, 1]
```

6. Entrenamiento del modelo.

```r
library(class)
wbcd_test_pred<- knn(train = wbcd_train, test = wbcd_test, cl= wbcd_train_labels, k=21)
```

7. Evaluación del modelo.

```
library(gmodels)
CrossTable(x = wbcd_test_labels, y = wbcd_test_pred,
prop.chisq=FALSE)
```

```
##
##
##    Cell Contents
## |-------------------------|
## |                       N |
## |           N / Row Total |
## |           N / Col Total |
## |         N / Table Total |
## |-------------------------|
##
##
## Total Observations in Table:  100
##
##
##                  | wbcd_test_pred
## wbcd_test_labels |    Benign | Malignant | Row Total |
## -----------------|-----------|-----------|-----------|
##           Benign |        61 |         0 |        61 |
##                  |     1.000 |     0.000 |     0.610 |
##                  |     0.968 |     0.000 |           |
##                  |     0.610 |     0.000 |           |
## -----------------|-----------|-----------|-----------|
##        Malignant |         2 |        37 |        39 |
##                  |     0.051 |     0.949 |     0.390 |
##                  |     0.032 |     1.000 |           |
##                  |     0.020 |     0.370 |           |
## -----------------|-----------|-----------|-----------|
##     Column Total |        63 |        37 |       100 |
##                  |     0.630 |     0.370 |           |
## -----------------|-----------|-----------|-----------|
##
##
```

8. Mejorando el modelo.

Con el modelo anterior se obtenían un 2% de falsos negativos. Esto no interesa, así que intentaremos mejorar el modelo: - Reescalando las variables numéricas. - Utilizando diferentes valores de k.

z-score standardization

```
wbcd_z <- as.data.frame(scale(wbcd[-1]))

#Comprobamos que los valores han sido estandarizados
summary(wbcd_z$area_mean)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## -1.4532 -0.6666 -0.2949  0.0000  0.3632  5.2459
```

Observamos que la media de los valores es 0. Procedemos de nuevo a crear los Test y Training datasets y a evaluar el modelo:

```
wbcd_train <- wbcd_z[1:469, ]
wbcd_test <- wbcd_z[470:569, ]
wbcd_train_labels <- wbcd[1:469, 1]
wbcd_test_labels <- wbcd[470:569, 1]
wbcd_test_pred <- knn(train = wbcd_train, test = wbcd_test,
cl = wbcd_train_labels, k = 21)
CrossTable(x = wbcd_test_labels, y = wbcd_test_pred,
prop.chisq = FALSE)
```

```
##
##
##    Cell Contents
## |-------------------------|
## |                       N |
## |           N / Row Total |
## |           N / Col Total |
## |         N / Table Total |
## |-------------------------|
##
##
## Total Observations in Table:  100
##
##
##                  | wbcd_test_pred
## wbcd_test_labels |    Benign | Malignant | Row Total |
## -----------------|-----------|-----------|-----------|
##           Benign |        61 |         0 |        61 |
##                  |     1.000 |     0.000 |     0.610 |
##                  |     0.924 |     0.000 |           |
##                  |     0.610 |     0.000 |           |
## -----------------|-----------|-----------|-----------|
##        Malignant |         5 |        34 |        39 |
##                  |     0.128 |     0.872 |     0.390 |
##                  |     0.076 |     1.000 |           |
##                  |     0.050 |     0.340 |           |
## -----------------|-----------|-----------|-----------|
##     Column Total |        66 |        34 |       100 |
##                  |     0.660 |     0.340 |           |
## -----------------|-----------|-----------|-----------|
##
##
```

Se observa que los falsos negativos han aumentado a un 5%.

Valores alternativos de k

Utilizamos diferentes valores de k con los datos normalizados:

```
wbcd_test_pred1<- knn(train = wbcd_train, test = wbcd_test, cl= wbcd_train_labels, k=1)

CrossTable(x = wbcd_test_labels, y = wbcd_test_pred1,
prop.chisq=FALSE)
```

```
##
##
##    Cell Contents
## |-------------------------|
## |                       N |
## |           N / Row Total |
## |           N / Col Total |
## |         N / Table Total |
## |-------------------------|
##
##
## Total Observations in Table:  100
##
##
##                  | wbcd_test_pred1
## wbcd_test_labels |    Benign | Malignant | Row Total |
## -----------------|-----------|-----------|-----------|
##           Benign |        59 |         2 |        61 |
##                  |     0.967 |     0.033 |     0.610 |
##                  |     0.952 |     0.053 |           |
##                  |     0.590 |     0.020 |           |
## -----------------|-----------|-----------|-----------|
##        Malignant |         3 |        36 |        39 |
##                  |     0.077 |     0.923 |     0.390 |
##                  |     0.048 |     0.947 |           |
##                  |     0.030 |     0.360 |           |
## -----------------|-----------|-----------|-----------|
##     Column Total |        62 |        38 |       100 |
##                  |     0.620 |     0.380 |           |
## -----------------|-----------|-----------|-----------|
##
##
```

```
wbcd_test_pred5<- knn(train = wbcd_train, test = wbcd_test, cl= wbcd_train_labels, k=5)

CrossTable(x = wbcd_test_labels, y = wbcd_test_pred5,
prop.chisq=FALSE)
```

```
##
##
##    Cell Contents
## |-------------------------|
## |                       N |
## |           N / Row Total |
## |           N / Col Total |
## |         N / Table Total |
## |-------------------------|
```

```
##
##
## Total Observations in Table:  100
##
##
##                  | wbcd_test_pred5
## wbcd_test_labels |    Benign | Malignant | Row Total |
## -----------------|-----------|-----------|-----------|
##           Benign |        60 |         1 |        61 |
##                  |     0.984 |     0.016 |     0.610 |
##                  |     0.968 |     0.026 |           |
##                  |     0.600 |     0.010 |           |
## -----------------|-----------|-----------|-----------|
##        Malignant |         2 |        37 |        39 |
##                  |     0.051 |     0.949 |     0.390 |
##                  |     0.032 |     0.974 |           |
##                  |     0.020 |     0.370 |           |
## -----------------|-----------|-----------|-----------|
##     Column Total |        62 |        38 |       100 |
##                  |     0.620 |     0.380 |           |
## -----------------|-----------|-----------|-----------|
##
##
```

```
wbcd_test_pred11<- knn(train = wbcd_train, test = wbcd_test, cl= wbcd_train_labels, k=11)

CrossTable(x = wbcd_test_labels, y = wbcd_test_pred11,
prop.chisq=FALSE)
```

```
##
##
##    Cell Contents
## |-------------------------|
## |                       N |
## |           N / Row Total |
## |           N / Col Total |
## |         N / Table Total |
## |-------------------------|
##
##
## Total Observations in Table:  100
##
##
##                  | wbcd_test_pred11
## wbcd_test_labels |    Benign | Malignant | Row Total |
## -----------------|-----------|-----------|-----------|
##           Benign |        60 |         1 |        61 |
##                  |     0.984 |     0.016 |     0.610 |
##                  |     0.952 |     0.027 |           |
##                  |     0.600 |     0.010 |           |
## -----------------|-----------|-----------|-----------|
##        Malignant |         3 |        36 |        39 |
##                  |     0.077 |     0.923 |     0.390 |
##                  |     0.048 |     0.973 |           |
```

```
##                  |     0.030 |     0.360 |           |
## -----------------|-----------|-----------|-----------|
##     Column Total |        63 |        37 |       100 |
##                  |     0.630 |     0.370 |           |
## -----------------|-----------|-----------|-----------|
##
##
```

```
wbcd_test_pred15<- knn(train = wbcd_train, test = wbcd_test, cl= wbcd_train_labels, k=15)

CrossTable(x = wbcd_test_labels, y = wbcd_test_pred15,
prop.chisq=FALSE)
```

```
##
##
##    Cell Contents
## |-------------------------|
## |                       N |
## |           N / Row Total |
## |           N / Col Total |
## |         N / Table Total |
## |-------------------------|
##
##
## Total Observations in Table:  100
##
##
##                  | wbcd_test_pred15
## wbcd_test_labels |    Benign | Malignant | Row Total |
## -----------------|-----------|-----------|-----------|
##           Benign |        61 |         0 |        61 |
##                  |     1.000 |     0.000 |     0.610 |
##                  |     0.953 |     0.000 |           |
##                  |     0.610 |     0.000 |           |
## -----------------|-----------|-----------|-----------|
##        Malignant |         3 |        36 |        39 |
##                  |     0.077 |     0.923 |     0.390 |
##                  |     0.047 |     1.000 |           |
##                  |     0.030 |     0.360 |           |
## -----------------|-----------|-----------|-----------|
##     Column Total |        64 |        36 |       100 |
##                  |     0.640 |     0.360 |           |
## -----------------|-----------|-----------|-----------|
##
##
```

```
wbcd_test_pred27<- knn(train = wbcd_train, test = wbcd_test, cl= wbcd_train_labels, k=27)

CrossTable(x = wbcd_test_labels, y = wbcd_test_pred27,
prop.chisq=FALSE)
```

```
##
##
```

```
##    Cell Contents
## |-------------------------|
## |                       N |
## |           N / Row Total |
## |           N / Col Total |
## |         N / Table Total |
## |-------------------------|
##
##
## Total Observations in Table:  100
##
##
##                  | wbcd_test_pred27
## wbcd_test_labels |    Benign | Malignant | Row Total |
## -----------------|-----------|-----------|-----------|
##           Benign |        61 |         0 |        61 |
##                  |     1.000 |     0.000 |     0.610 |
##                  |     0.924 |     0.000 |           |
##                  |     0.610 |     0.000 |           |
## -----------------|-----------|-----------|-----------|
##        Malignant |         5 |        34 |        39 |
##                  |     0.128 |     0.872 |     0.390 |
##                  |     0.076 |     1.000 |           |
##                  |     0.050 |     0.340 |           |
## -----------------|-----------|-----------|-----------|
##     Column Total |        66 |        34 |       100 |
##                  |     0.660 |     0.340 |           |
## -----------------|-----------|-----------|-----------|
##
##
```

Observamos que con k=5 se obtiene un menor número de falsos positivos.