Evaluación ML I

Ejercicio 1 (Análisis Discriminante, Naïve Bayes)

Inmaculada Perea Fernández

junio 2017

Sobre la base de datos *BreastCancer* de la librería *mlbench*, realice las siguientes actividades:

- Construya un clasificador *Naive-Bayes* usando una muestra aleatoria constituida por 2/3 de la totalidad del fichero de datos.
- Obtenga la matriz de confusión y el porcentaje de clasificación incorrecta a partir de las instancias no usadas en la construcción del clasificador.
- Determine el número de predicciones correspondientes a la clase *malignant*
- De las predicciones consideradas en el apartado anterior, determine cuántas de ellas se han obtenido con una probabilidad mayor que 0.75

Carga de las librerías necesarias

```
if (!require('mlbench')) install.packages('mlbench'); library('mlbench')
if (!require('e1071')) install.packages('e1071'); library('e1071')
```

1 Carga, inspección y preparación de los datos

1.1. Carga e inspección de los datos

El conjunto de datos *BreastCancer* consta de 699 observaciones y 11 variables:

- [,1] **Id**: Sample code number
- [,2] **Cl.thickness**: Clump Thickness
- [,3] **Cell.size**: Uniformity of Cell Size
- [,4] **Cell.shape**: Uniformity of Cell Shape
- [,5] **Marg.adhesion**: Marginal Adhesion
- [,6] **Epith.c.size**: Single Epithelial Cell Size
- [,7] **Bare.nuclei**: Bare Nuclei
- [,8] **Bl.cromatin**: Bland Chromatin
- [,9] **Normal.nucleoli**: Normal Nucleoli
- [,10] **Mitoses**: Mitoses
- [,11] **Class**: Class

```
# carga de los datos
data(BreastCancer)
# Dimensión de los datos
dim(BreastCancer)
```

```
## [1] 699  11
```

```
str(BreastCancer)
```

```
## 'data.frame':    699 obs. of  11 variables:
##  $ Id             : chr  "1000025" "1002945" "1015425" "1016277" ...
##  $ Cl.thickness   : Ord.factor w/ 10 levels "1"<"2"<"3"<"4"<..: 5 5 3 6 4 8 1 2 2 4 ...
```

```
##  $ Cell.size      : Ord.factor w/ 10 levels "1"<"2"<"3"<"4"<..: 1 4 1 8 1 10 1 1 1 2 ...
##  $ Cell.shape     : Ord.factor w/ 10 levels "1"<"2"<"3"<"4"<..: 1 4 1 8 1 10 1 2 1 1 ...
##  $ Marg.adhesion  : Ord.factor w/ 10 levels "1"<"2"<"3"<"4"<..: 1 5 1 1 3 8 1 1 1 1 ...
##  $ Epith.c.size   : Ord.factor w/ 10 levels "1"<"2"<"3"<"4"<..: 2 7 2 3 2 7 2 2 2 2 ...
##  $ Bare.nuclei    : Factor w/ 10 levels "1","2","3","4",..: 1 10 2 4 1 10 10 1 1 1 ...
##  $ Bl.cromatin    : Factor w/ 10 levels "1","2","3","4",..: 3 3 3 3 3 9 3 3 1 2 ...
##  $ Normal.nucleoli: Factor w/ 10 levels "1","2","3","4",..: 1 2 1 7 1 7 1 1 1 1 ...
##  $ Mitoses        : Factor w/ 9 levels "1","2","3","4",..: 1 1 1 1 1 1 1 1 5 1 ...
##  $ Class          : Factor w/ 2 levels "benign","malignant": 1 1 1 1 1 2 1 1 1 1 ...
```

```
head(BreastCancer)
```

```
##        Id Cl.thickness Cell.size Cell.shape Marg.adhesion Epith.c.size
## 1 1000025            5         1          1             1            2
## 2 1002945            5         4          4             5            7
## 3 1015425            3         1          1             1            2
## 4 1016277            6         8          8             1            3
## 5 1017023            4         1          1             3            2
## 6 1017122            8        10         10             8            7
##   Bare.nuclei Bl.cromatin Normal.nucleoli Mitoses     Class
## 1           1           3               1       1    benign
## 2          10           3               2       1    benign
## 3           2           3               1       1    benign
## 4           4           3               7       1    benign
## 5           1           3               1       1    benign
## 6          10           9               7       1 malignant
```

```
summary(BreastCancer)
```

```
##       Id             Cl.thickness   Cell.size     Cell.shape
##  Length:699         1      :145   1      :384   1      :353
##  Class :character   5      :130   10     : 67   2      : 59
##  Mode  :character   3      :108   3      : 52   10     : 58
##                     4      : 80   2      : 45   3      : 56
##                     10     : 69   4      : 40   4      : 44
##                     2      : 50   5      : 30   5      : 34
##                     (Other):117   (Other): 81   (Other): 95
##  Marg.adhesion  Epith.c.size  Bare.nuclei   Bl.cromatin  Normal.nucleoli
##  1      :407   2      :386   1      :402   2      :166   1      :443
##  2      : 58   3      : 72   10     :132   3      :165   10     : 61
##  3      : 58   4      : 48   2      : 30   1      :152   3      : 44
##  10     : 55   1      : 47   5      : 30   7      : 73   2      : 36
##  4      : 33   6      : 41   3      : 28   4      : 40   8      : 24
##  8      : 25   5      : 39   (Other): 61   5      : 34   6      : 22
##  (Other): 63   (Other): 66   NA's   : 16   (Other): 69   (Other): 69
##     Mitoses          Class
##  1      :579   benign   :458
##  2      : 35   malignant:241
##  3      : 33
##  10     : 14
##  4      : 12
##  7      :  9
##  (Other): 17
```

En la inspección de los datos realizadas en este apartado observamos que existe una variable que no aporta información al estudio (*Id*), y que existen 16 valores perdidos pertenecientes a la variable *Bare.nuclei*.

1.2 Preparación de los datos

A continuación realizaremos las transformaciones necesarias a los datos antes de usarlos en la construcción del modelo.

1.2.1 Eliminación de la variable *Id*

Eliminamos la variable *Id* del estudio, ya que se trata de un identificador tipo caracter de la muestra, y no aporta información relevante para la clasificación.

```
datos<-subset(BreastCancer, select=-Id)
```

1.2.2 Estudio y eliminación de los valores perdidos

```
table(is.na(datos))
```

```
##
## FALSE  TRUE
##  6974    16
```

A continuación eliminaremos los valores perdidos del estudio, porque aunque sea posible indicar que no se tengan en cuenta en la construcción del modelo con la función *naviveBayes* y la opción *na.action=na.omit*, no queremos que formen parte tampoco del conjunto test que construiremos a continuación.

```
datos<-na.omit(datos)
dim(datos)
```

```
## [1] 683  10
```

Si volvemos a consultar la existencia de valores perdidos observamos que se han eliminado correctamente, y que nuestro dataset ahora sólo contiene valores completos.

```
table(is.na(datos))
```

```
##
## FALSE
##  6830
```

1.2.3. División entrenamiento y test

A continuación dividiremos el conjunto de datos en entrenamiento y test. Destinaremos 2/3 de los datos a entrenamiento y 1/3 a test

```
set.seed(123456789)
n=nrow(datos)
train.index=sort(sample(1:n, ceiling((2/3)*n)))
train=datos[train.index,]
test=datos[-train.index,]
```

Conjunto de entrenamiento

```
dim(train)
```

```
## [1] 456  10
```

```
summary(train)
```

```
##   Cl.thickness   Cell.size     Cell.shape  Marg.adhesion  Epith.c.size
##  1      :99    1      :247   1      :228   1      :268   2      :248
##  5      :90    10     : 45   10     : 40   3      : 41   3      : 51
##  3      :65    3      : 36   3      : 39   2      : 36   1      : 30
##  4      :50    2      : 30   2      : 34   10     : 35   6      : 29
##  10     :49    4      : 23   4      : 27   8      : 18   4      : 27
##  2      :32    6      : 21   5      : 22   4      : 17   10     : 24
##  (Other):71    (Other): 54   (Other): 66   (Other): 41   (Other): 47
##   Bare.nuclei   Bl.cromatin  Normal.nucleoli    Mitoses          Class
##  1      :263   3      :111   1      :289     1      :372   benign   :294
##  10     : 94   2      :102   10     : 43     2      : 23   malignant:162
##  5      : 22   1      :100   3      : 30     3      : 23
##  2      : 20   7      : 51   2      : 21     10     : 11
##  3      : 19   4      : 27   5      : 17     4      :  8
##  4      : 14   5      : 22   6      : 15     8      :  7
##  (Other): 24   (Other): 43   (Other): 41     (Other): 12
```

```
str(train)
```

```
## 'data.frame':    456 obs. of  10 variables:
##  $ Cl.thickness   : Ord.factor w/ 10 levels "1"<"2"<"3"<"4"<..: 5 5 3 4 8 1 2 4 1 2 ...
##  $ Cell.size      : Ord.factor w/ 10 levels "1"<"2"<"3"<"4"<..: 1 4 1 1 10 1 1 2 1 1 ...
##  $ Cell.shape     : Ord.factor w/ 10 levels "1"<"2"<"3"<"4"<..: 1 4 1 1 10 1 1 1 1 1 ...
##  $ Marg.adhesion  : Ord.factor w/ 10 levels "1"<"2"<"3"<"4"<..: 1 5 1 3 8 1 1 1 1 1 ...
##  $ Epith.c.size   : Ord.factor w/ 10 levels "1"<"2"<"3"<"4"<..: 2 7 2 2 7 2 2 2 1 2 ...
##  $ Bare.nuclei    : Factor w/ 10 levels "1","2","3","4",..: 1 10 2 1 10 10 1 1 1 1 ...
##  $ Bl.cromatin    : Factor w/ 10 levels "1","2","3","4",..: 3 3 3 3 9 3 1 2 3 2 ...
##  $ Normal.nucleoli: Factor w/ 10 levels "1","2","3","4",..: 1 2 1 1 7 1 1 1 1 1 ...
##  $ Mitoses        : Factor w/ 9 levels "1","2","3","4",..: 1 1 1 1 1 1 5 1 1 1 ...
##  $ Class          : Factor w/ 2 levels "benign","malignant": 1 1 1 1 2 1 1 1 1 1 ...
##  - attr(*, "na.action")=Class 'omit'  Named int [1:16] 24 41 140 146 159 165 236 250 276 293 ...
##   .. ..- attr(*, "names")= chr [1:16] "24" "41" "140" "146" ...
```

```
table(train$Class)
```

```
##
##    benign malignant
##       294       162
```

Conjunto de test

```
dim(test)
```

```
## [1] 227  10
```

```
summary(test)
```

```
##   Cl.thickness   Cell.size     Cell.shape  Marg.adhesion  Epith.c.size
##  1      :40    1      :126   1      :118   1      :125   2      :128
##  3      :39    10     : 22   2      : 24   2      : 22   4      : 21
##  5      :38    3      : 16   10     : 18   10     : 20   3      : 20
##  4      :29    2      : 15   4      : 16   3      : 17   5      : 17
##  10     :20    4      : 15   3      : 14   4      : 16   1      : 14
##  2      :18    5      : 10   7      : 12   5      :  9   6      : 11
```

```
##  (Other):43    (Other): 23   (Other): 25   (Other): 18   (Other): 16
##   Bare.nuclei   Bl.cromatin Normal.nucleoli    Mitoses          Class
##  1      :139   2      :58   1      :143     1      :191   benign   :150
##  10     : 38   1      :50   10     : 17     2      : 12   malignant: 77
##  8      : 11   3      :50   2      : 15     3      : 10
##  2      : 10   7      :20   3      : 12     4      :  4
##  3      :  9   4      :12   8      : 12     7      :  4
##  5      :  8   5      :12   6      :  7     10     :  3
##  (Other): 12   (Other):25   (Other): 21     (Other):  3
```

```
str(test)
```

```
## 'data.frame':    227 obs. of  10 variables:
##  $ Cl.thickness   : Ord.factor w/ 10 levels "1"<"2"<"3"<"4"<..: 6 2 1 8 7 10 5 2 10 5 ...
##  $ Cell.size      : Ord.factor w/ 10 levels "1"<"2"<"3"<"4"<..: 8 1 1 7 4 7 1 1 10 4 ...
##  $ Cell.shape     : Ord.factor w/ 10 levels "1"<"2"<"3"<"4"<..: 8 2 1 5 6 7 1 1 10 4 ...
##  $ Marg.adhesion  : Ord.factor w/ 10 levels "1"<"2"<"3"<"4"<..: 1 1 1 10 4 6 1 1 8 9 ...
##  $ Epith.c.size   : Ord.factor w/ 10 levels "1"<"2"<"3"<"4"<..: 3 2 2 7 6 4 2 2 6 2 ...
##  $ Bare.nuclei    : Factor w/ 10 levels "1","2","3","4",..: 4 1 3 9 1 10 1 1 1 10 ...
##  $ Bl.cromatin    : Factor w/ 10 levels "1","2","3","4",..: 3 3 3 5 4 4 2 3 8 5 ...
##  $ Normal.nucleoli: Factor w/ 10 levels "1","2","3","4",..: 7 1 1 5 3 1 1 1 9 6 ...
##  $ Mitoses        : Factor w/ 9 levels "1","2","3","4",..: 1 1 1 4 1 2 1 1 1 1 ...
##  $ Class          : Factor w/ 2 levels "benign","malignant": 1 1 1 2 2 2 1 1 2 2 ...
##  - attr(*, "na.action")=Class 'omit'  Named int [1:16] 24 41 140 146 159 165 236 250 276 293 ...
##   .. ..- attr(*, "names")= chr [1:16] "24" "41" "140" "146" ...
```

```
table(test$Class)
```

```
##
##    benign malignant
##       150        77
```

En ambos conjuntos estan presentes las dos clases existentes *benign* y *malignant*, por lo que no será necesario realizar suavizado de Laplace ni tampoco seleccionar el conjunto test y aprendizaje con técnicas de estratificación de la librería *caret*.

2. Contrucción del clasificador *Naive-Bayes*

A continuación construiremos el modelo

```
clasificador=naiveBayes(x=subset(train, select=-Class), # variables predictoras
                        y=train$Class,                  # variable objetivo (clases)
                        laplace=0,
                        na.action=na.omit)
```

2.1 Proporción de muestras en cada clase estimadas a partir de la muestra

```
clasificador$apriori
```

```
## train$Class
##    benign malignant
##       294       162
```

2.3 Probabilidades de cada variable condicionadas a la clase

```
clasificador$tables
```

```
## $Cl.thickness
##            Cl.thickness
## train$Class           1           2           3           4           5
##   benign    0.329931973 0.095238095 0.193877551 0.149659864 0.193877551
##   malignant 0.012345679 0.024691358 0.049382716 0.037037037 0.203703704
##            Cl.thickness
## train$Class           6           7           8           9          10
##   benign    0.030612245 0.000000000 0.006802721 0.000000000 0.000000000
##   malignant 0.061728395 0.074074074 0.185185185 0.049382716 0.302469136
##
## $Cell.size
##            Cell.size
## train$Class           1           2           3           4           5
##   benign    0.836734694 0.081632653 0.057823129 0.020408163 0.000000000
##   malignant 0.006172840 0.037037037 0.117283951 0.104938272 0.123456790
##            Cell.size
## train$Class           6           7           8           9          10
##   benign    0.000000000 0.003401361 0.000000000 0.000000000 0.000000000
##   malignant 0.129629630 0.067901235 0.123456790 0.012345679 0.277777778
##
## $Cell.shape
##            Cell.shape
## train$Class           1           2           3           4           5
##   benign    0.775510204 0.105442177 0.078231293 0.023809524 0.006802721
##   malignant 0.000000000 0.018518519 0.098765432 0.123456790 0.123456790
##            Cell.shape
## train$Class           6           7           8           9          10
##   benign    0.006802721 0.003401361 0.000000000 0.000000000 0.000000000
##   malignant 0.117283951 0.104938272 0.129629630 0.037037037 0.246913580
##
## $Marg.adhesion
##            Marg.adhesion
## train$Class           1           2           3           4           5
##   benign    0.826530612 0.078231293 0.068027211 0.010204082 0.006802721
##   malignant 0.154320988 0.080246914 0.129629630 0.086419753 0.074074074
##            Marg.adhesion
## train$Class           6           7           8           9          10
##   benign    0.006802721 0.000000000 0.000000000 0.000000000 0.003401361
##   malignant 0.074074074 0.061728395 0.111111111 0.018518519 0.209876543
##
## $Epith.c.size
##            Epith.c.size
## train$Class           1           2           3           4           5
##   benign    0.102040816 0.802721088 0.054421769 0.010204082 0.010204082
##   malignant 0.000000000 0.074074074 0.216049383 0.148148148 0.117283951
##            Epith.c.size
## train$Class           6           7           8           9          10
##   benign    0.003401361 0.006802721 0.006802721 0.000000000 0.003401361
##   malignant 0.172839506 0.030864198 0.086419753 0.012345679 0.141975309
##
```

```
## $Bare.nuclei
##            Bare.nuclei
## train$Class           1           2           3           4           5
##   benign    0.867346939 0.057823129 0.023809524 0.010204082 0.027210884
##   malignant 0.049382716 0.018518519 0.074074074 0.067901235 0.086419753
##            Bare.nuclei
## train$Class           6           7           8           9          10
##   benign    0.000000000 0.000000000 0.003401361 0.000000000 0.010204082
##   malignant 0.012345679 0.037037037 0.055555556 0.037037037 0.561728395
##
## $Bl.cromatin
##            Bl.cromatin
## train$Class           1           2           3           4           5
##   benign    0.340136054 0.326530612 0.289115646 0.013605442 0.013605442
##   malignant 0.000000000 0.037037037 0.160493827 0.141975309 0.111111111
##            Bl.cromatin
## train$Class           6           7           8           9          10
##   benign    0.003401361 0.013605442 0.000000000 0.000000000 0.000000000
##   malignant 0.030864198 0.290123457 0.111111111 0.049382716 0.067901235
##
## $Normal.nucleoli
##            Normal.nucleoli
## train$Class           1           2           3           4           5
##   benign    0.891156463 0.061224490 0.023809524 0.003401361 0.006802721
##   malignant 0.166666667 0.018518519 0.141975309 0.074074074 0.092592593
##            Normal.nucleoli
## train$Class           6           7           8           9          10
##   benign    0.010204082 0.000000000 0.003401361 0.000000000 0.000000000
##   malignant 0.074074074 0.055555556 0.061728395 0.049382716 0.265432099
##
## $Mitoses
##            Mitoses
## train$Class           1           2           3           4           5
##   benign    0.962585034 0.020408163 0.006802721 0.000000000 0.003401361
##   malignant 0.549382716 0.104938272 0.129629630 0.049382716 0.030864198
##            Mitoses
## train$Class           6           7           8          10
##   benign    0.000000000 0.003401361 0.003401361 0.000000000
##   malignant 0.006172840 0.024691358 0.037037037 0.067901235
```

3. Evaluación del rendimiento

A continuación evaluaremos la bondad del ajuste del modelo *Naïve Bayes* a los datos.

3.1 Cálculo de predicciones sobre el conjunto test

Indicamos *type="clase"* porque en este caso nos interesa conocer la predicción de pertenencia o no a cada clase.

```
prediccion.class = predict(object=clasificador,
                           newdata=test,
                           type="class")
```

3.2 Matriz de confusión

```
(matconf = table(prediccion.class,
                 test$Class,
                 dnn=c("clase pronosticada","clase real")))
```

```
##                    clase real
## clase pronosticada benign malignant
##          benign       146         3
##          malignant      4        74
```

3.3 Porcentaje de clasificación incorrecta

```
round(100*(1-(sum(diag(matconf))/nrow(test))), 3)
```

```
## [1] 3.084
```

El porcentaje de clasificación incorrecta es bajo, con lo que podemos concluir que el modelo obtenido se ajusta bien a los datos.

3.4 Número de predicciones correspondientes a la clase *malignant*

A contiuación el número de predicciones clasificadas como *malignant*

```
length(which(prediccion.class=="malignant"))
```

```
## [1] 78
```

A continuación el porcentaje de predicciones del total calculado que corresponden a la clase *malignant*

```
round(100*length((which(prediccion.class=="malignant"))) / length(prediccion.class), 3)
```

```
## [1] 34.361
```

3.5 Predicciones correspondientes a la clase *malignant* con probabilidad mayor que *0.75*

Indicamos la opción *type="raw"* en el cálculo de las predicciones en el conjunto test para obtener el valor de la probabilidad de pertenencia a cada clase.

```
prediccion.raw = predict(object=clasificador,
                         newdata=test,
                         type="raw")

length(which(prediccion.raw[,2] > 0.75))
```

```
## [1] 77
```

Sólamente una de las 78 predicciones obtenidas de pertenencia a la clase *malignant* tiene una probabilidad inferior a 0.75. Las otras 77 tienen una probabilidad de pertenecer a *malignant* superior a 0.75.