Tarea Machine learning - Arboles

Manuel Carmona Cabello de Alba

1-Carga de datos y análisis exploratorio

Primero cargamos los datos y renombramos las variables para entender bien la informacion:

```
setwd("C:/Users/Manuel/Desktop/CUNEF/MACHINE LEARNING/arboles")
abalone <- read.table("abalone.data.txt", sep=",")

names(abalone) <- c("Sex", "Length", "Diameter", "Height", "Whole
weight", "Shucked weight", "Viscera weight", "Shell weight", "Rings" )
head(abalone, 6)
```

```
##   Sex Length Diameter Height Whole weight Shucked weight Viscera
weight
## 1   M  0.455    0.365  0.095       0.5140         0.2245
0.1010
## 2   M  0.350    0.265  0.090       0.2255         0.0995
0.0485
## 3   F  0.530    0.420  0.135       0.6770         0.2565
0.1415
## 4   M  0.440    0.365  0.125       0.5160         0.2155
0.1140
## 5   I  0.330    0.255  0.080       0.2050         0.0895
0.0395
## 6   I  0.425    0.300  0.095       0.3515         0.1410
0.0775
##   Shell weight Rings
## 1        0.150    15
## 2        0.070     7
## 3        0.210     9
## 4        0.155    10
## 5        0.055     7
## 6        0.120     8
```

Comprobamos de qué tipo es la información de la que disponemos

```
str(abalone)
```

```
## 'data.frame':    4177 obs. of  9 variables:
##  $ Sex           : Factor w/ 3 levels "F","I","M": 3 3 1 3 2 2 1 1 3 1
...
##  $ Length        : num  0.455 0.35 0.53 0.44 0.33 0.425 0.53 0.545
0.475 0.55 ...
```

```
##  $ Diameter      : num  0.365 0.265 0.42 0.365 0.255 0.3 0.415 0.425
0.37 0.44 ...
##  $ Height        : num  0.095 0.09 0.135 0.125 0.08 0.095 0.15 0.125
0.125 0.15 ...
##  $ Whole weight  : num  0.514 0.226 0.677 0.516 0.205 ...
##  $ Shucked weight: num  0.2245 0.0995 0.2565 0.2155 0.0895 ...
##  $ Viscera weight: num  0.101 0.0485 0.1415 0.114 0.0395 ...
##  $ Shell weight  : num  0.15 0.07 0.21 0.155 0.055 0.12 0.33 0.26
0.165 0.32 ...
##  $ Rings         : int  15 7 9 10 7 8 20 16 9 19 ...
```

La variable Sex ya está en modo factor, pero vamos a renombrarla para que sea mas facil su lectura

```
abalone$Sex<- factor(abalone$Sex, levels=c("M","F", "I"),
                     labels=c("male", "female", "infant"))
```

Comprobamos si existen valores perdidos y la distribución de los valores:

```
summary(abalone)
```

```
##      Sex           Length         Diameter          Height
##  male  :1528   Min.   :0.075   Min.   :0.0550   Min.   :0.0000
##  female:1307   1st Qu.:0.450   1st Qu.:0.3500   1st Qu.:0.1150
##  infant:1342   Median :0.545   Median :0.4250   Median :0.1400
##                Mean   :0.524   Mean   :0.4079   Mean   :0.1395
##                3rd Qu.:0.615   3rd Qu.:0.4800   3rd Qu.:0.1650
##                Max.   :0.815   Max.   :0.6500   Max.   :1.1300
##   Whole weight    Shucked weight   Viscera weight    Shell weight
##  Min.   :0.0020   Min.   :0.0010   Min.   :0.0005   Min.   :0.0015
##  1st Qu.:0.4415   1st Qu.:0.1860   1st Qu.:0.0935   1st Qu.:0.1300
##  Median :0.7995   Median :0.3360   Median :0.1710   Median :0.2340
##  Mean   :0.8287   Mean   :0.3594   Mean   :0.1806   Mean   :0.2388
##  3rd Qu.:1.1530   3rd Qu.:0.5020   3rd Qu.:0.2530   3rd Qu.:0.3290
##  Max.   :2.8255   Max.   :1.4880   Max.   :0.7600   Max.   :1.0050
##      Rings
##  Min.   : 1.000
##  1st Qu.: 8.000
##  Median : 9.000
##  Mean   : 9.934
##  3rd Qu.:11.000
##  Max.   :29.000
```

Todo es correcto, procedemos a comprobar si existe o no inestabilidad.

2-Muestra de entrenamiento y muestra de test

Definimos una muestra aleatoria de aprendizaje del arbol

```
set.seed(1234)
train <- sample(nrow(abalone), 0.7*nrow(abalone))  #esto al azar el 70% de la muestra
```

La muestra de tes será el total de observaciones menos aquellas empleadas en la muestra de aprendizaje.

```
abalone.train <- abalone[train,]   #con los elementos de la muestra que acabo de crear

abalone.validate <- abalone[-train,]  #con los elementos restantes
```

Comprobamos valores

```
table(abalone.train$Sex)

##
##   male female infant
##   1078    913    932

table(abalone.validate$Sex)

##
##   male female infant
##    450    394    410
```

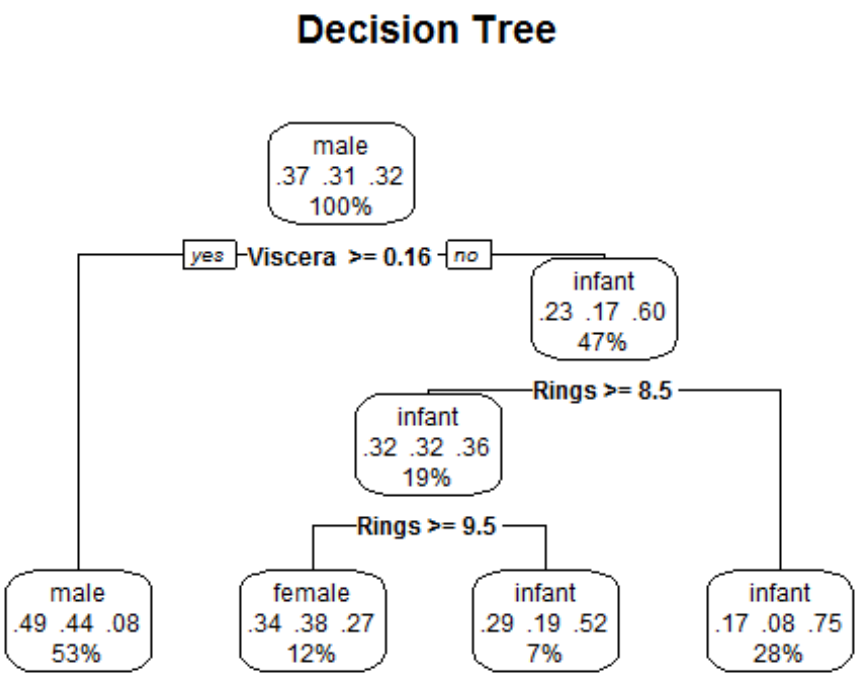
Esta balanceado en distribucion si comparamos train y validate.

3-Primer arbol

Estimamos un arbol con la función rpart y lo representamos para una mejor interpretacion:

```
library(rpart)
library(rpart.plot)

## Warning: package 'rpart.plot' was built under R version 3.4.2

# Estimamos el arbol

arbol <- rpart(Sex ~ ., data=abalone.train, method="class",
               parms=list(split="information"))

print(arbol) #esta info sera mas completa con la representacion gráfica

## n= 2923
##
```

```
## node), split, n, loss, yval, (yprob)
##       * denotes terminal node
##
##  1) root 2923 1845 male (0.36879918 0.31235033 0.31885050)
##    2) Viscera weight>=0.16025 1558  795 male (0.48973042 0.43517330
0.07509628) *
##    3) Viscera weight< 0.16025 1365  550 infant (0.23076923 0.17216117
0.59706960)
##      6) Rings>=8.5 542  346 infant (0.32287823 0.31549815 0.36162362)
##       12) Rings>=9.5 351  216 female (0.34188034 0.38461538
0.27350427) *
##       13) Rings< 9.5 191   91 infant (0.28795812 0.18848168
0.52356021) *
##      7) Rings< 8.5 823  204 infant (0.17010936 0.07776428 0.75212637)
*

prp(arbol, type = 2, extra = 104,
    fallen.leaves = TRUE, main="Decision Tree")
```




```
summary(arbol)

## Call:
## rpart(formula = Sex ~ ., data = abalone.train, method = "class",
##     parms = list(split = "information"))
##   n= 2923
##
##           CP nsplit rel error    xerror       xstd
```

```
## 1 0.27100271      0 1.0000000 1.0000000 0.01413829
## 2 0.01056911      1 0.7289973 0.7333333 0.01461127
## 3 0.01000000      3 0.7078591 0.7154472 0.01458287
##
## Variable importance
## Viscera weight   Whole weight   Shell weight       Diameter
Length
##             19             16             15             15
14
## Shucked weight          Rings         Height
##             14              5              1
##
## Node number 1: 2923 observations,    complexity param=0.2710027
##   predicted class=male    expected loss=0.6312008  P(node) =1
##     class counts:  1078   913   932
##    probabilities: 0.369 0.312 0.319
##   left son=2 (1558 obs) right son=3 (1365 obs)
##   Primary splits:
##       Viscera weight < 0.16025 to the right, improve=495.6382, (0
missing)
##       Whole weight   < 0.77275 to the right, improve=481.6755, (0
missing)
##       Shell weight   < 0.1925  to the right, improve=456.7567, (0
missing)
##       Shucked weight < 0.31175 to the right, improve=414.4825, (0
missing)
##       Diameter       < 0.4325  to the right, improve=402.3604, (0
missing)
##   Surrogate splits:
##       Whole weight   < 0.774   to the right, agree=0.945, adj=0.881,
(0 split)
##       Diameter       < 0.4075  to the right, agree=0.913, adj=0.814,
(0 split)
##       Length         < 0.5325  to the right, agree=0.913, adj=0.813,
(0 split)
##       Shucked weight < 0.30675 to the right, agree=0.911, adj=0.810,
(0 split)
##       Shell weight   < 0.19525 to the right, agree=0.908, adj=0.802,
(0 split)
##
## Node number 2: 1558 observations
##   predicted class=male    expected loss=0.5102696  P(node) =0.533014
##     class counts:   763   678   117
##    probabilities: 0.490 0.435 0.075
##
## Node number 3: 1365 observations,    complexity param=0.01056911
##   predicted class=infant  expected loss=0.4029304  P(node) =0.466986
##     class counts:   315   235   815
##    probabilities: 0.231 0.172 0.597
##   left son=6 (542 obs) right son=7 (823 obs)
```

```
##   Primary splits:
##       Rings          < 8.5     to the right, improve=113.42140, (0
missing)
##       Shell weight   < 0.11525 to the right, improve= 87.71521, (0
missing)
##       Viscera weight < 0.08675 to the right, improve= 86.21448, (0
missing)
##       Height         < 0.1025  to the right, improve= 80.72430, (0
missing)
##       Whole weight   < 0.34825 to the right, improve= 76.54570, (0
missing)
##   Surrogate splits:
##       Shell weight   < 0.15125 to the right, agree=0.749, adj=0.367,
(0 split)
##       Height         < 0.1225  to the right, agree=0.733, adj=0.328,
(0 split)
##       Viscera weight < 0.11225 to the right, agree=0.719, adj=0.293,
(0 split)
##       Diameter       < 0.3575  to the right, agree=0.705, adj=0.258,
(0 split)
##       Whole weight   < 0.55475 to the right, agree=0.705, adj=0.256,
(0 split)
##
## Node number 6: 542 observations,    complexity param=0.01056911
##   predicted class=infant  expected loss=0.6383764  P(node) =0.1854259
##     class counts:   175   171   196
##    probabilities: 0.323 0.315 0.362
##   left son=12 (351 obs) right son=13 (191 obs)
##   Primary splits:
##       Rings          < 9.5     to the right, improve=18.957050, (0
missing)
##       Viscera weight < 0.04575 to the left,  improve=10.775790, (0
missing)
##       Length         < 0.3625  to the left,  improve=10.193620, (0
missing)
##       Height         < 0.0975  to the left,  improve= 9.994489, (0
missing)
##       Diameter       < 0.2875  to the left,  improve= 9.728561, (0
missing)
##   Surrogate splits:
##       Whole weight   < 0.2015  to the right, agree=0.655, adj=0.021,
(0 split)
##       Shucked weight < 0.389   to the left,  agree=0.655, adj=0.021,
(0 split)
##       Length         < 0.3025  to the right, agree=0.653, adj=0.016,
(0 split)
##       Diameter       < 0.23    to the right, agree=0.653, adj=0.016,
(0 split)
##       Viscera weight < 0.15925 to the left,  agree=0.653, adj=0.016,
(0 split)
```

```
##
## Node number 7: 823 observations
##   predicted class=infant  expected loss=0.2478736  P(node) =0.28156
##     class counts:   140    64   619
##    probabilities: 0.170 0.078 0.752
##
## Node number 12: 351 observations
##   predicted class=female  expected loss=0.6153846  P(node) =0.1200821
##     class counts:   120   135    96
##    probabilities: 0.342 0.385 0.274
##
## Node number 13: 191 observations
##   predicted class=infant  expected loss=0.4764398  P(node) =0.06534382
##     class counts:    55    36   100
##    probabilities: 0.288 0.188 0.524
```

Observamos que hay 4 nodos terminales.

4-Segundo arbol

Repetimos el proceso pero cambiando la semilla para ver si existe inestabilidad a la hora de generar el arbol. En el caso de que no exista aparentemente, variaremos la proporcion de muestra de entrenamiento y test para comprobar si realmente es estable.

```
set.seed(888)
train <- sample(nrow(abalone), 0.7*nrow(abalone))  #esto me selecciona al azar el 70% de la muestra

abalone.train <- abalone[train,]   #con los elementos de la muestra que acabo de crear

abalone.validate <- abalone[-train,]  #con los elementos restantes

table(abalone.train$Sex)

##
##   male female infant
##   1073    899    951

table(abalone.validate$Sex)

##
##   male female infant
##    455    408    391
```

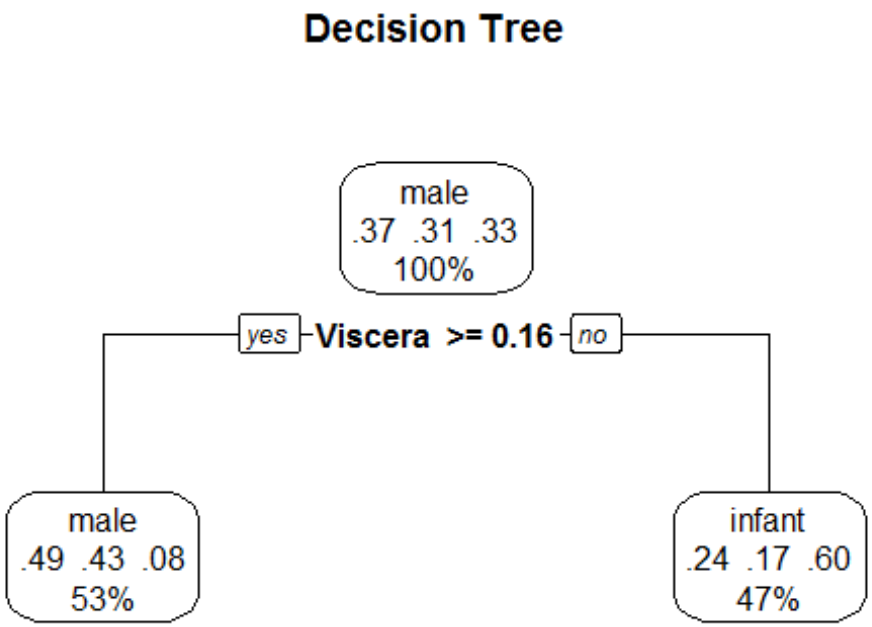
Esta balanceado en distribucion si comparamos train y validate

```
library(rpart)
library(rpart.plot)
```

```r
# Estimamos el arbol

arbol <- rpart(Sex ~ ., data=abalone.train, method="class",
               parms=list(split="information"))

print(arbol) #esta info sera mas completa con la representacion gráfica
```

```
## n= 2923
##
## node), split, n, loss, yval, (yprob)
##       * denotes terminal node
##
## 1) root 2923 1850 male (0.36708861 0.30756073 0.32535067)
##   2) Viscera weight>=0.16225 1535  790 male (0.48534202 0.43452769
0.08013029) *
##   3) Viscera weight< 0.16225 1388  560 infant (0.23631124 0.16714697
0.59654179) *
```

```r
prp(arbol, type = 2, extra = 104,
    fallen.leaves = TRUE, main="Decision Tree")
```




```r
summary(arbol)
```

```
## Call:
## rpart(formula = Sex ~ ., data = abalone.train, method = "class",
##     parms = list(split = "information"))
```

```
##   n= 2923
##
##          CP nsplit rel error   xerror       xstd
## 1 0.2702703      0 1.0000000 1.000000 0.01408639
## 2 0.0100000      1 0.7297297 0.732973 0.01457398
##
## Variable importance
## Viscera weight   Whole weight       Diameter         Length Shucked
weight
##             20             17             16             16
16
##   Shell weight
##             16
##
## Node number 1: 2923 observations,    complexity param=0.2702703
##   predicted class=male    expected loss=0.6329114  P(node) =1
##     class counts:  1073   899   951
##    probabilities: 0.367 0.308 0.325
##   left son=2 (1535 obs) right son=3 (1388 obs)
##   Primary splits:
##       Viscera weight < 0.16225 to the right, improve=482.2239, (0
missing)
##       Whole weight   < 0.633   to the right, improve=466.5766, (0
missing)
##       Shell weight   < 0.19375 to the right, improve=436.5648, (0
missing)
##       Height         < 0.1275  to the right, improve=400.0514, (0
missing)
##       Shucked weight < 0.38025 to the right, improve=396.5868, (0
missing)
##   Surrogate splits:
##       Whole weight   < 0.7495  to the right, agree=0.940, adj=0.875,
(0 split)
##       Diameter       < 0.4075  to the right, agree=0.912, adj=0.814,
(0 split)
##       Length         < 0.5325  to the right, agree=0.911, adj=0.813,
(0 split)
##       Shucked weight < 0.3215  to the right, agree=0.908, adj=0.806,
(0 split)
##       Shell weight   < 0.21575 to the right, agree=0.906, adj=0.802,
(0 split)
##
## Node number 2: 1535 observations
##   predicted class=male    expected loss=0.514658  P(node) =0.5251454
##     class counts:   745   667   123
##    probabilities: 0.485 0.435 0.080
##
## Node number 3: 1388 observations
##   predicted class=infant  expected loss=0.4034582  P(node) =0.4748546
```

```
##     class counts:   328   232   828
##    probabilities: 0.236 0.167 0.597
```

Como podemos observar, solo nos da 2 nodos terminales y cercena la posibilidad de que la observación sea mujer. Existe inestabilidad claramente.