MUIInf
Máster Oficial Universitario en
Ingeniería Informática
muiinf.webs.upv.es

UNIVERSITAT POLITÈCNICA DE VALÈNCIA
etsinf

Data Science (CDA)

Practical 5:
Prediction Models

José Hernández Orallo (jorallo@dsic.upv.es)
(Material adapted by M. José Ramírez Quintana)

In this sesion we are going to illustrate prediction models by using one of the classification techniques seen in Unit 3: Decision Trees.

1. A classification case study

To illustrate how to deal with classification problems, we are going to work with the wine dataset (https://archive.ics.uci.edu/ml/datasets/Wine), also available on poliformat. In this practical we are going to focus on decision trees.
There are different R packages for working with decision trees. Some of them are specific for decision trees such as the C50 library (the original C5.0 implementation) and others are more general packages that support not only different kinds of trees (for instance, the party and rpart packages) but also other machine learning techniques, such as the RWeka package or the caret package (http://topepo.github.io/caret/).
The wine dataset describes 13 constituents found in a chemical analysis of three types of wines grown in the same region in Italy. The attributes are

1) Alcohol
2) Malic acid
3) Ash
4) Alcalinity of ash
5) Magnesium
6) Total phenols
7) Flavanoids
8) Nonflavanoid phenols
9) Proanthocyanins
10) Color intensity
11) Hue
12) OD280/OD315 of diluted wines
13) Proline

LOADING DATA

We load the dataset into R:

```
setwd("…the path where your file is located…")
wine<-read.table("wine.data.txt",header=F,sep=',')
```

ANALYSING DATA

and inspect it:

```
str(wine)
```

```
## 'data.frame':    178 obs. of  14 variables:
##  $ V1 : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ V2 : num  14.2 13.2 13.2 14.4 13.2 ...
##  $ V3 : num  1.71 1.78 2.36 1.95 2.59 1.76 1.87 2.15 1.64 1.35 ...
##  $ V4 : num  2.43 2.14 2.67 2.5 2.87 2.45 2.45 2.61 2.17 2.27 ...
##  $ V5 : num  15.6 11.2 18.6 16.8 21 15.2 14.6 17.6 14 16 ...
##  $ V6 : int  127 100 101 113 118 112 96 121 97 98 ...
##  $ V7 : num  2.8 2.65 2.8 3.85 2.8 3.27 2.5 2.6 2.8 2.98 ...
##  $ V8 : num  3.06 2.76 3.24 3.49 2.69 3.39 2.52 2.51 2.98 3.15 ...
##  $ V9 : num  0.28 0.26 0.3 0.24 0.39 0.34 0.3 0.31 0.29 0.22 ...
##  $ V10: num  2.29 1.28 2.81 2.18 1.82 1.97 1.98 1.25 1.98 1.85 ...
##  $ V11: num  5.64 4.38 5.68 7.8 4.32 6.75 5.25 5.05 5.2 7.22 ...
##  $ V12: num  1.04 1.05 1.03 0.86 1.04 1.05 1.02 1.06 1.08 1.01 ...
##  $ V13: num  3.92 3.4 3.17 3.45 2.93 2.85 3.58 3.58 2.85 3.55 ...
##  $ V14: int  1065 1050 1185 1480 735 1450 1290 1295 1045 1045 ...
```

The first column is the class which has been considered as numeric. Hence, our next step is to change the names of the columns (to make them more informative) and the type of the target variable.

```
t<-
c("class","alco","ma","ash","alc","mg","tp","flav","noflav","proa","co
l","hue","od","prol")
names(wine)<-t
wine$class=as.factor(wine$class)
head(wine)
```

```
##   class  alco   ma  ash  alc  mg   tp flav noflav proa  col  hue
od prol
## 1     1 14.23 1.71 2.43 15.6 127 2.80 3.06   0.28 2.29 5.64 1.04
3.92 1065
## 2     1 13.20 1.78 2.14 11.2 100 2.65 2.76   0.26 1.28 4.38 1.05
3.40 1050
## 3     1 13.16 2.36 2.67 18.6 101 2.80 3.24   0.30 2.81 5.68 1.03
3.17 1185
## 4     1 14.37 1.95 2.50 16.8 113 3.85 3.49   0.24 2.18 7.80 0.86
3.45 1480
## 5     1 13.24 2.59 2.87 21.0 118 2.80 2.69   0.39 1.82 4.32 1.04
2.93  735
## 6     1 14.20 1.76 2.45 15.2 112 3.27 3.39   0.34 1.97 6.75 1.05
2.85 1450
```

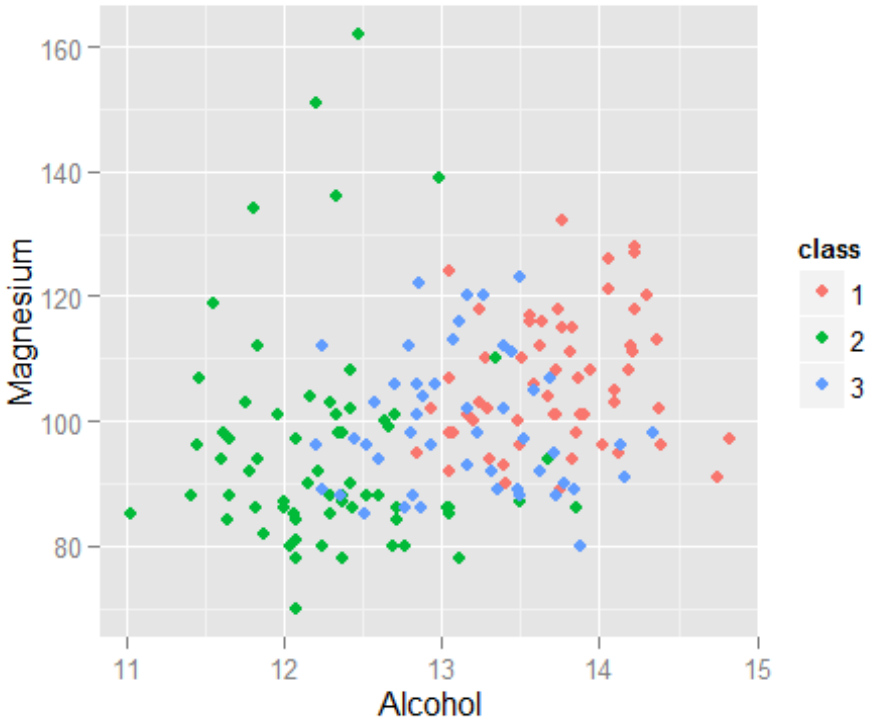
We can also show some plots:

```
if (!require("ggplot2"))
{install.packages("ggplot2",dependencies=TRUE); library("ggplot2")}
```

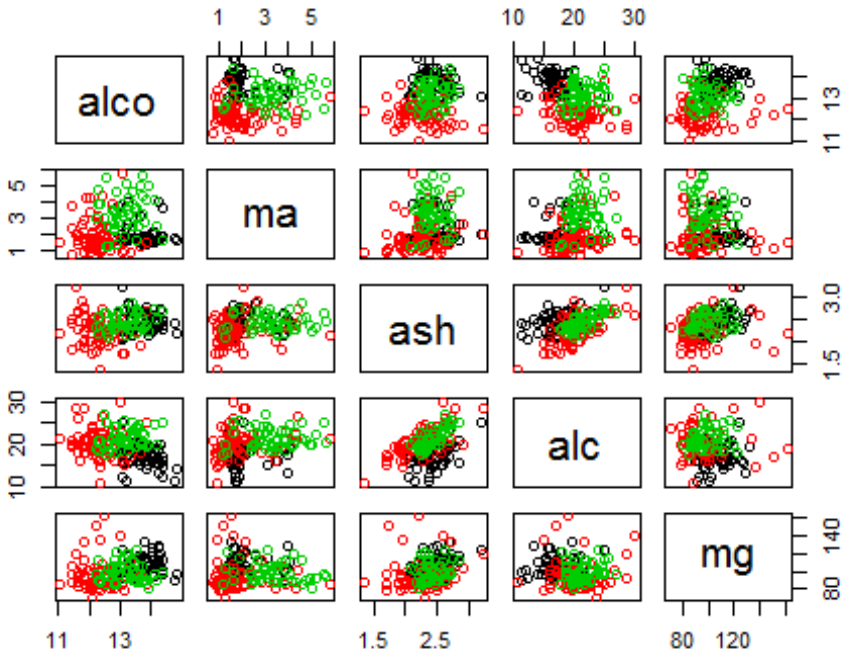
```
## Loading required package: ggplot2
```

```
## Warning: package 'ggplot2' was built under R version 3.2.2
```

```
qplot(alco,mg,data=wine,col=class, xlab="Alcohol",ylab="Magnesium")
```




```
pairs(wine[,2:6],col=wine$class)
```



LEARNING A MODEL

We learn a Decision Tree using the rpart package:

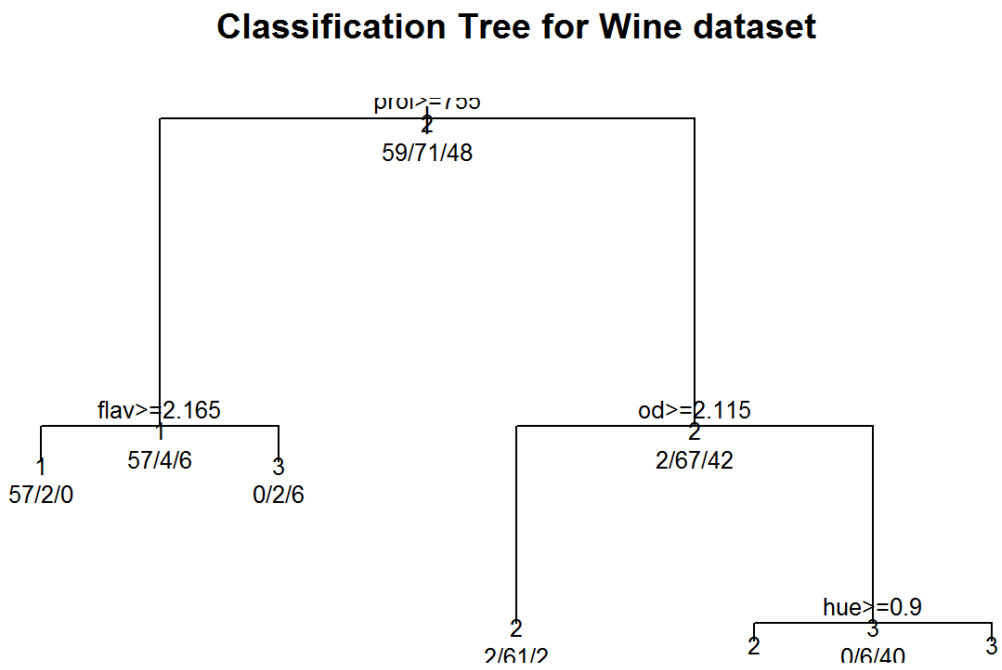
```
if (!require("rpart")) {install.packages("rpart",dependencies=TRUE);
library("rpart")}

## Loading required package: rpart

model<-rpart(class~.,data=wine,method="class")
```

The tree can be graphically visualised using plot and text:

```
plot(model, main="Classification Tree for Wine dataset")
text(model,use.n=TRUE, all=TRUE, cex=.8)
```



We can also display the model

```
printcp(model)

##
## Classification tree:
## rpart(formula = class ~ ., data = wine, method = "class")
##
## Variables actually used in tree construction:
## [1] flav hue  od   prol
```

```
##
## Root node error: 107/178 = 0.60112
##
## n= 178
##
##         CP nsplit rel error  xerror     xstd
## 1 0.495327      0   1.00000 1.00000 0.061056
## 2 0.317757      1   0.50467 0.45794 0.055693
## 3 0.056075      2   0.18692 0.29907 0.047880
## 4 0.028037      3   0.13084 0.20561 0.041037
## 5 0.010000      4   0.10280 0.18692 0.039378
```

2. Exercises

1.- Randomly split the dataset into 75% train and 25% test. Note: It can be done by using the sample function to generate integers belonging to the interval [1..size(data set)]. Use these numbers to identify the instances.
2.- Learn a decision tree using the training set.
3.- Visualise the tree and display the results. Is there any difference with respect to the model trained with the whole dataset?
4.- Use the model to predict the class label for the test set by using the "predict" function. Repeat the predictions but now using the parameter type="class" (use a different variable to keep the new results). What are the differences?
5.- Calculate the performance of the model when it is applied to the test set by displaying a table that shows the predicted classes versus the real classes.
6.- Try some other methods or parameters of the rpart package to see whether you can still improve the results further. You can also compare with other packages.