Laboratory 7: Supervised classification

Felipe Bravo Oviedo
Universidad de Valladolid, Spain

In this Lab we will learn how to classify observations by:

- Logisitic regression
- Lineas discriminant analysis

We will built upon the previous lab and we'll introduce the library MASS (https://cran.r-project.org/web/packages/MASS/). In this lab we will two different datasets that you can find with the documentation of this lab. One of the data sets can be also downloaded from the Zenodo repository: https://zenodo.org/record/198522#.W0E-bcJ9jDc As usual, first we must define our working directory.

Logistic regression

As usual firstly we must establish our working directory and import the dataset by using this code:

```
# establishing the working directory

setwd('C:/your_desired_working_directoryR')

# Importing data

valdepoza<- read.csv2("valdepoza.csv")

names(valdepoza) # we'll see the names of the variables
head(valdepoza) # we'll see the first 6 observations in the dataset

#indicating if the variables are factors
valdepoza$SP <- factor(valdepoza$SP)
valdepoza$education <- factor(valdepoza$education)
valdepoza$forestry <- factor(valdepoza$forestry)
valdepoza$gender <- factor(valdepoza$gender)

# education (1=postsecondary studies, 0= secondary o lower studies)
# forestry (1=forestry experience by education or work, = =no forestry background)
# gender (1=woman, 0=man)
# age is the age of the marker (the person conducting the tree marking for harvest)
# DBH is the tree diameter at breast height in cm
# HT is the total tree height in m
```

Now we can fit the logistic model for different explanatory variables sets.

```
# Logistic model
#round 1
round1 <- glm(harvest ~  DBH + HT + SP, data = valdepoza, family = "binomial")
summary(round1)
```

we will obtain the following output:

```
Call:
glm(formula = harvest ~ DBH + HT + SP, family = "binomial", data = valdepo
za)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.1407  -0.6316  -0.5592  -0.4802   2.3438

Coefficients:
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)         0.200139   0.169889   1.178    0.239
DBH                 0.030450   0.006416   4.746 2.07e-06 ***
HT                 -0.132103   0.012419 -10.637  < 2e-16 ***
SPPinus sylvestris -0.026770   0.101668  -0.263    0.792
SPQuercus pyrenaica -1.361773   0.123254 -11.049  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 6894.9  on 7354  degrees of freedom
Residual deviance: 6635.9  on 7350  degrees of freedom
  (31 observations deleted due to missingness)
AIC: 6645.9

Number of Fisher Scoring iterations: 4
```

Where we can see that the intercept is not significant but far more important that there is no differences between *Pinus nigra* (the default species) and *Pinus sylvestris* so we can generate a new variable to unify both pine species in one variable.

```
# Defining if the tree is a pine or not
# these instructions add a variable indicating if the tree is a pine or not
str(valdepoza)
valdepoza$pine[valdepoza$SP == "Quercus pyrenaica"] <- 0
valdepoza$pine[valdepoza$SP  == "Pinus nigra"] <- 1
valdepoza$pine[valdepoza$SP  == "Pinus sylvestris"] <- 1

#indicating that pine is a factor
valdepoza$pine <- factor(valdepoza$pine)

# pine (1= Pinus nigra or Pinus sylvestris, 0 =Quercus pyrenaica)
```

Now we can test other structures for the logistic model as the following:

```
#round 2
round2 <- glm(harvest ~  DBH + HT + pine + age, data = valdepoza, family = "binomial")
summary(round2)
```

obtaining the following ouput:

```
Call:
glm(formula = harvest ~ DBH + HT + pine + age, family = "binomial",
    data = valdepoza)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.2288  -0.6312  -0.5581  -0.4761   2.3789

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.404234   0.100009 -14.041  < 2e-16 ***
DBH          0.030586   0.006409   4.773 1.82e-06 ***
HT          -0.131270   0.011735 -11.186  < 2e-16 ***
pine1        1.342373   0.090827  14.780  < 2e-16 ***
age          0.006327   0.001925   3.286  0.00102 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 6894.9  on 7354  degrees of freedom
Residual deviance: 6625.3  on 7350  degrees of freedom
  (31 observations deleted due to missingness)
AIC: 6635.3

Number of Fisher Scoring iterations: 4
```

Here we can see that all variables are significant, so we could fit ther alternative models and compare its on the base of the Akaike Information Criterion (AIC). Let's try other models including gender, education and/or forestry background and test if you improve the round2 model. Remember that all the variables must be significant to be valid for the AIC comparison.

Full script

You can copy and paste into your R environment (changing your working directory)

```
# establishing the working directory

setwd('C:/your_desired_working_directoryR')

# Importing data

valdepoza<- read.csv2("valdepoza.csv")

names(valdepoza) # we'll see the names of the variables
head(valdepoza) # we'll see the first 6 observations in the dataset

#indicating if the variables are factors
valdepoza$SP <- factor(valdepoza$SP)
valdepoza$education <- factor(valdepoza$education)
```

```
valdepoza$forestry <- factor(valdepoza$forestry)
valdepoza$gender <- factor(valdepoza$gender)

# education (1=postsecondary studies, 0= secondary o lower studies)
# forestry (1=forestry experience by education or work, = =no forestry background)
# gender (1=woman, 0=man)

# Importing data
# Logistic models
#round 1
round1 <- glm(harvest ~  DBH + HT + SP, data = valdepoza, family = "binomial")
summary(round1)

# Defining if the tree is a pine or not
# these instructions add a variable indicating if the tree is a pine or not
valdepoza$pine[valdepoza$SP == "Quercus pyrenaica"] <- 0
valdepoza$pine[valdepoza$SP  == "Pinus nigra"] <- 1
valdepoza$pine[valdepoza$SP  == "Pinus sylvestris"] <- 1

#indicating that pine is a factor
valdepoza$pine <- factor(valdepoza$pine)

# pine (1= Pinus nigra or Pinus sylvestris, 0 =Quercus pyrenaica)

#round 2
round2 <- glm(harvest ~  DBH + HT + pine + age, data = valdepoza, family = "binomial")
summary(round2)
```

Linear Discriminant analysis

Now we are going to develop a linear discrimination analysis (LDA) classifier for *Pinus halepensis* plantations in Northern Spain (see Bueis et al 2007 for details). To use this method we will load the library MASS (remember install.package(MASS) if you did not before). Details about the method can be obtained in the chapter 4 of James et al (2013).

To start we must set our working directory if you did not before and read the file containing the dataset

```
#setting working directory
setwd('C:/your_desired_working_directoryR')

#loading dataset
soilsite <- read.csv2("SI_phal_LDA_data.csv", header=TRUE)
names(soilsite)
head(soilsite)
```

The dataset contains information about 32 plots established in *Pinus halepensis* plantations in Northern Spain (Bueis et al, 2017) including geographical variables (SLOPE, ALTITUDE, LATITUDE, LONGITUDE), soil characteristics (CLAY, SILT and SAND under USDA and IS methodology, pH, CCC, Ca, K and Mg), climatic characteristics (mean annual temperature and rainfall and Martonne and Lang index), stand level variables (N, QMD, H0, G and AGE) and site productivity (SI and SI3)

Now me must load the library MASS

```
#loading library
library(MASS) # remember install.package ( ) if you did not before
```

At this moment we are ready to obtain our first LDA classifier

```
#MODEL 1
SI.lda1 <- lda(factor(SI3) ~ SANDIS + CLAY + Martonne + CCC , data = soilsite)
# summary the model 1
SI.lda1
```

To obtain this outcome:

```
SI.lda1
Call:
lda(factor(SI3) ~ SANDIS + CLAY + Martonne + CCC, data = soilsite)

Prior probabilities of groups:
      8      11      14
0.25000 0.46875 0.28125

Group means:
     SANDIS     CLAY Martonne      CCC
8  31.08875 21.81375 20.17500 19.90000
11 36.73000 22.72467 21.18000 21.62000
14 36.93889 22.50667 21.47778 20.84444

Coefficients of linear discriminants:
                  LD1         LD2
SANDIS   -0.01374631  0.01699722
CLAY     -0.03987327  0.01485128
Martonne -0.71801245 -0.30288892
CCC      -0.03830357  0.18853800

Proportion of trace:
   LD1    LD2
0.9472 0.0528
```

Now we can see the prediction with this model and the rate of correct classifications.

```
# predictions
SI.lda.values1 <- predict(SI.lda1)
SI.lda.values1
table(Predicted=SI.lda.values1$class, Observed=soilsite$SI3)
```

The table of observed vs predicted can help us to insight on how good our model is if we compare to get right classifications by chance. We will use the metric called Cohen's Kappa (see Cohen, 1960 for details)

$$K = \frac{A_G - P_f}{1 - P_f}$$

Where AG is the overall agreement ratio and Pf is the proportion of the most frequent class. K can take values between -1 and 1 (but values below zero are unlikely (Cohen, 1960) The K values obtained inform us about the agreement between the model and the most frequent class decision to classify new observations. Cohen (1960) suggested that when K is ≤ 0 there is no agreement (as we saw before this unlikely) values from 0,01 to 0,20 shows from none to slight improvement, from 0,21 to 0,40 fair improvement, from 0,41 to 0,60 moderate improvement, from 0,61 to 0,80 substantial improvement and finally from 0,81 to1,00 strong improvement. See now the output obtained with the previous code:

```
          Observed
Predicted  8 11 14
       8   3  1  1
       11  5 13  7
       14  0  1  1
```

The proportion of right classifications with this model can be obtaining the sum of the diagonal of the previous matrix (3 +13+1) divided by the total number of observations (32). The overall agreement ratio is 0,53125. On the other hand, if we assign every observation to the most frequent class (in our example the class 11 in the SI3 variable) we will correct classify with a rate of 0,46875 which is the proportion of the most frequent class. With these values the Cohen's kappa for our example is:

$K = (0{,}53125 - 0{,}46875)/(1 - 0{,}46875) = 0{,}11765$

So our method is 11,765% better that classify all the observations to the most frequent class.

Let's try now new models to see if you is able to improve the proposed model.

References

Bravo, F. Bravo-Núñez, A. 2017 Clasificación de la calidad de estación forestal mediante técnicas de aprendizaje automático (*machine learning*) 7Congreso Forestal Español (in Spanish) http://7cfe.congresoforestal.es/content/clasificacion-de-la-calidad-de-estacion-forestal-mediante-tecnicas-de-aprendizaje-0

Bueis, Teresa; Bravo, Felipe; Pando, Valentín & Turrion, Mª Belén (2017) Site factors as predictors for Pinus halepensis Mill. productivity in Spanish plantations. Annals of Forest Science 74: 6. doi:10.1007/s13595-016-0609-7

Cohen, J.A. 1960. A coefficient of agreement for nominal scales. Educational and Pshychological Measurements 20:37-46 doi:10.1177/001316446002000104

James, G., Witten, D. Hastie, T. and Tibshirani, R. 2013 An Introduction to Statistical Learning with Applications in R. Springer Freely available at http://www-bcf.usc.edu/~gareth/ISL

McHugh, M.L. 2012 Interrater reliability: the kappa statistic *Biochem Med (Zagreb)* 22(3):276-282 https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3900052/