Labwork on **decision trees**

Nicolas Sutton-Charani (Euromov DHM)

IMT Mines-Alès

Solutions will be given on this online notebook.

Exercise 1

(a) Import the *drug200* dataset from this url.

(b) Evaluate the predictive power of the classification tree algorithm for the 'Drug' variable prediction.

(c) Plot a classification tree trained on the whole dataset.

(d) Preprocess the data.

(e) Redo questions 2 and 3.

Exercise 2

(a) Import the *gobelins*[1] dataset from the following link.

(b) Remove the 'id' variable from the dataset and explain why it would not be irrelevant to use it for the prediction of the monsters types.

(c) With the *rpart* package, learn of a classification tree from the entire imported dataset in order to predict the type of monster from the other attributes.

(d) Plot the resulting tree with the *rpart.plot* package.

(e) According to that tree, what would be the type of a green monster having a bone length of 0.3, a percentage of rotting flesh of 0.5, a hair length of 0.6 and a percentage of soul of 0.75? What about the confidence of this prediction?

(f) Compute and plot (with the *ggplot2* package) attributes importance weights from the tree.

(g) In order to justify the decision trees approach, evaluate decision tree algorithm on the *gobelins* dataset.

(h) Pre-prune the tree graphically by ploting the training and testing accuracies on the same plot according to different depths and choosing the right one.

(i) Run a cross-validation procedure on the gobelins dataset repeated many times comparing decision trees (tuned with the optimal depth from previous question) and several random forests (with different number of trees) and boxplot the results.

[1] https://www.kaggle.com/c/ghouls-goblins-and-ghosts-boo/overview

Exercise 3

(a) Import the *weather* dataset which is native in the *rattle* package.

(b) After having removed the irrelevant variables plot a decision tree which can predict if it will rain tomorrow or not.

(c) Evaluate the approach according to different metrics.

(d) Learn a tree that predict the fraction of sky obscured by cloud (*Cloud3pm*) from the temperature (degrees C) at 9am (*Temp9am*), the relative humidity (percent) at 9am (*Humidity9am*), the fraction of sky obscured by cloud at 9am (*Cloud9am*), the wind speed (km/hr) averaged over 10 minutes prior to 9am (*WindSpeed9am*) and the atmospheric pressure (hpa) reduced to mean sea level at 9am (*Pressure9am*).

(e) Evaluate the approach in terms of MSE (compare with reference values).

(f) Try to optimise the tree by tuning the *minsplit, minbucket, cp* and *maxdepth* hyperparameters and redo the previous question.

Exercise 4

(a) Import the premier league dataset from the following link.

(b) After having set the 'cp' parameter of *rpart* to 0.08, quantify (in terms of probability of winning) the difference between home team and away team winning at half time.

(c) Quantify the difference between Watford and Arsenal as home team when they play against Manchester City and there is a draw game at half time.