TÉCNICO LISBOA

MEIC 2018/19 Ciência de Dados

Lab 4: Decision Trees and Random Forests

1. Load the *breast_cancer* data. Train a decision tree using the C4.5 algorithm and k-fold cross validation.
   a. What is the mean accuracy and standard deviation achieved? What confidence do you have in that estimation for the accuracy?
   b. Change the minimum number of instances to split a node and study the accuracy.
   c. Compare the trees discovered with and without pruning.
   d. Study the effect of using just binary splits.
   e. Repeat the procedures with CART.
   f. Compare the performance of C4.5, CART and DecisionStumps.

2. Splitting the same dataset into train and test datasets, train a random forest.
   a. What is the accuracy achieved with default parameters?
   b. Change the minimum number of trees to use and study the accuracy achieved.
   c. Change the maximum tree size to use and study the accuracy achieved.
   d. Is there any model in overfitting?

3. Load the *credit* data. Evaluate the accuracy on the training and testing datasets separately, study the overfitting of the different models trained.
   a. Run C4.5, changing the number of instances to split a node.
   b. Run RandomForests, changing the maximum tree size to combine in the ensembler.

4. Load the *diabetes* data. Study the role of each pre-processing technique when learning decision trees and random forests.
   a. **discretization**.
   b. **normalization**.
   c. **feature selection**.