IMT Atlantique
Bretagne-Pays de la Loire
École Mines-Télécom

Induction of decision trees

- exploratory data analysis including visual analysis
- learn decision trees
- compare their performance with baseline classifiers
- tune training parameters
- use of train and test framework and cross validation frameworl

Commentaires (version 1.0 – November 19, 2019) :

Philippe Lenca
philippe.lenca@imt-atlantique.fr

1 Introduction to decision trees with the Iris dataset

The Iris flower dataset was collected by Edgar Anderson to quantify the morphologic variation of Iris flowers of three related species. It was used by Ronald Aylmer Fisher in his classic 1936 paper on linear discriminant analysis, "The use of multiple measurements in taxonomic problems" (Annals of Eugenics, 7: 179-188 (1936)), and then became very famous. You may find it at http://archive.ics.uci.edu/ml/datasets/Iris but the Iris dataset is available in datasets libraries in R and Python[1].

So, this dataset consists of three types of irises' (Setosa, Versicolour, and Virginica) and some properties about each flower, stored in a 150x4 array: the rows being the samples and the columns being Sepal Length, Sepal Width, Petal Length, Petal Width and Species. There are 50 samples of each species. The target attribute is Species and the predictive ones are Sepal Length, Sepal Width, Petal Length, and Petal Width.

The goal of this lab is to introduce basic concepts in learning decision trees. There are many packages for modelling decision trees, for example in R: rpart, party, tree, RWeka, ipred and C50, and Scikit-learn in Python.

Exercice 1.[Exploratory data analysis]

Perform an exploratory data analysis (make a summary of the data, is there missing value?, outliers?).

Perform a visual exploratory data analysis (plot the data: planes with two attributes, parallel coordinates plots, scatter plot matrix, density plot and 3d scatter plots).

What can you conclude?

Exercice 2.[Decision trees]

1. calculate several decision trees with conventional impurity measures (gini, information, etc.) and default parameters

[1]This dataset differs from the data presented in Fishers article –on the 35^{th} and 38^{th} samples, see http://archive.ics.uci.edu/ml/datasets/Iris.

2. evaluate and compare their performance (with accuracy for example) and make also comparison with baseline classifiers (for example with the predict the most common class value algorithm and the One Rule algorithm) ; use train and test framework (is there a risk of an overestimated evaluation), and cross validation framework

3. analyse some training parameters like the minimum number of observations that must exist in a node in order for a split to be attempted, the minimum number of observations in any terminal leaf node and the cp complexity parameter in CART.

4. if you have time please add a comparison with other classifiers like knn and lda (as in Fisher's paper).

Please send your code and analysis by... to