Fundamentos de Data Science

LAB N. 6

Note that the instructions below not very "prescriptive" on the exact steps that you should carry out in this lab. It is important that you think by yourself about what you should do in order to perform a meaningful analysis of the data.

In this lab, you will implement Decision Trees, for both classification and regression. Depending on the problem, your trees will be using nominal or quantitative features.

DATASET DESCRIPTION

You will be using 3 different datasets, that you will find on the Eclass page for this lab:

1) Car.csv: this dataset is constituted by *1728* points in 6 dimensions. Each point refers to a car, which is described by 6 attributes, all of them (first 6 columns). Each car belongs to one of possible 4 classes (last column).

   The description of the attributes is as follows

   buying: buying price
   maint : price of the maintenance
   doors : number of doors
   persons: capacity in terms of persons to carry
   lug_boot: the size of luggage boot
   safety : estimated safety of the car

   The values of the attributes are:

   Buying: v-high, high, med, low
   Maint: v-high, high, med, low
   Doors: 2, 3, 4, 5-more
   Persons: 2, 4, more
   lug_boot: small, med, big
   safety: low, med, high

   The classes and their proportions are:

CLASS	N	%
unacc	1210	70.02
acc	384	22.22
good	69	4.00
v-good	65	3.76

2) GlassClassification.csv . This dataset is constituted by 198 points in 4 dimensions (first 4 columns), belonging to 2 classes, represented by the integers 1, 2 in the last column.

3) Housing.txt: this dataset is constituted by *506* points in 14 dimensions. Each point represents a house in the Boston area, and the 14 attributes that you find orderly in each column are the following:

   CRIM - per capita crime rate by town
   ZN - proportion of residential land zoned for lots over 25,000 sq.ft.
   INDUS - proportion of non-retail business acres per town.
   CHAS - Charles River dummy variable (1 if tract bounds river; 0 otherwise)
   NOX - nitric oxides concentration (parts per 10 million)
   RM - average number of rooms per dwelling
   AGE - proportion of owner-occupied units built prior to 1940
   DIS - weighted distances to five Boston employment centres
   RAD - index of accessibility to radial highways
   TAX - full-value property-tax rate per $10,000
   PTRATIO - pupil-teacher ratio by town
   B - 1000(Bk - 0.63)^2 where Bk is the proportion of blacks by town
   LSTAT - % lower status of the population
   MEDV - Median value of owner-occupied homes in $1000's

This dataset is normally associated with 2 regression tasks: predicting NOX (in which the nitrous oxide level is to be predicted); and predicting price MED (in which the median value of a home is to be predicted).

YOUR TASKS

You will build three decision trees for the car and glass classification problems and for the housing regression problem (here you should be predicting the MED attribute).

You will need to keep an eye on a few important details:

- for regression and classification problems the functions that you need to minimize are different.
- your decision trees should all be binary, even when your data contain nominal attributes that can assume more than 2 values. (for any non-binary tree there is a binary one which is equivalent)
- Don't forget that you will need to prune the trees once you have built them. That is, once you have the tree T, you will need to consider for elimination all pairs of neighbouring leaf nodes (i.e., leaves linked to a common antecedent node, one level above).
  For each pair: If its deletion yields an increase in performance on the cross-validation set, then delete it, and the common antecedent node becomes a leaf.
  You will need to repeat this for each possible pair, recursively).

Today you should aim at finishing a few scripts that implement the above descriptions.

During next week you should try to:

- divide your dataset into training and testing;
- measure the error on the training and testing dataset (which error measure should you use?);
- make your code more modular, factoring it into functions;
- generate some plots (with title, variable names on the axes, etc).

Have fun ! ☺