FGV EMAp
ESCOLA DE MATEMÁTICA APLICADA

Machine Learning 1

LAB EXERCISE 5

This lab exercise must be submitted by February 16nd, 2024 at 10:00 am.
Late submissions will not be accepted and will be marked as zero.

This lab exercise is mandatory and could be assessed.
It is worth 1% of your total final grade for this course.

Submission Instructions

Submit your files through EClass. The files you submit cannot be read by any other students. You can replace your submission as many times as you like by resubmitting it, although only the last version sent is kept.

If you have last-minute problems with your EClass submission, email your assignment as an attachment to alberto.paccanaro@fgv.br with the subject "URGENT – LAB 5 SUBMISSION". In the body of the message, explain the reason for not submitting it through EClass.

All the work you submit should be solely your own work.
Coursework submissions will be checked for this.

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

DATASET DESCRIPTION

You will be using the Boston Housing dataset, that you will find on the Eclass page for this lab:

This dataset (in file Housing.txt) 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 TASK

You will build a decision tree 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:

- 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 ! ☺