ΟΙΚΟΝΟΜΙΚΟ
ΠΑΝΕΠΙΣΤΗΜΙΟ
ΑΘΗΝΩΝ

ATHENS UNIVERSITY
OF ECONOMICS
AND BUSINESS

Athens University of Economics and Business

Department of Management Science and Technology

MSc in Business Analytics—Big Data Systems

Spark Assignment

Associate Professor P. Louridas

Instructions

Your submission will consist of a Scala file containing the code that you will write.

The Problem

You will write a Scala Spark script that creates and evaluates a Random Forest regressor on a given dataset. The dataset is available at https://archive.ics.uci.edu/ml/machine-learning-databases/housing/housing.data and a detailed description can be found at https://archive.ics.uci.edu/ml/datasets/Housing. The dataset concerns housing values in suburbs of Boston. Each line of the file contains values for the following:

- `crim`: per capita crime rate by town
- `zn`: proportion of residential land zoned for lots over 25,000 sq. ft.
- `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`: $1000(b_k - 0.63)^2$, where b_k is the proportion of blacks by town
- `lstat`: % lower status of the population
- `medv`: median value of owner-occupied homes in $1000s

Your goal is to train the regression and then gauge its performance, as given by the Mean Square Error (MSE). Check the notes in the next page for more information, if needed.

Notes

- Although you may write your solution in a standalone Spark program, you do not have to. It is enough to put in a Scala file *all* the code that you use in the Spark shell, so that by using `:load YourFile.scala` I will be able to run it.

- If you want to learn Scala, here are some links:

  - http://www.scala-lang.org/documentation/getting-started.html
  - http://www.scala-lang.org/docu/files/ScalaTutorial.pdf

- A good book for getting up to speed with Scala fast is *Scala for the Impatient* by Cay Horstmann, Addison-Wesley, 2012. Curiously, this seems to be available online in PDF form at places.

- Documentation about regression trees in Spark is at http://spark.apache.org/docs/1.6.3/mllib-ensembles.html#regression-1.

- Documentation about random forests in Spark is at http://spark.apache.org/docs/1.6.3/mllib-ensembles.html#random-forests.

- From the description of the features, you can see that the `chas` feature is categorical. It is worth checking the performance of your model when you treat it as numerical compared to what happens when you treat it as categorical. To do that, the `categoricalFeaturesInfo` variable of the example in the lecture slides must be defined as:

  ```
  val categoricalFeaturesInfo = Map(
    3 -> 2
  )
  ```

  That is, you indicate which columns are categorical and how many categorical values there are in each column. So, the above indicates that feature 3 (that is, the fourth column, as they are zero-indexed) is binary.

Go for it!