Big Data Frameworks
(Spark)

PSB – Efrei Paris
MSc
Data Management

2019/2020

Machine Learning with Spark

Salim NAHLE

Organization:

- You can work on any Spark environment
- A **PDF report** is expected. It shall contain the code, explanations and necessary screenshots.
- Please work in **pairs**! Each group (composed of 2 persons at most) shall submit one report. Do not forget to indicate your names in the report.
- The report shall be sent by email before **Thursday 16/07/2020 at 23:55.**
- Late reports are penalized (2 points per day)

Abstract:

- The objective of this mini-project is to use the different Spark machine learning libraries to build a predictive model.
- An open data set is provided. The correct answers are given. Supervised learning algorithms are thus used.
- In the data set, the output is continuous, you shall build several regression models, tune them and compare them

Bike Rental Data Set from UCI Machine Learning Repository

1. Citations

Consider the Bike Rental data set

Fanaee-T, Hadi, and Gama, Joao, 'Event labeling combining ensemble detectors and background knowledge', Progress in Artificial Intelligence (2013): pp. 1-15, Springer Berlin Heidelber

2. Attributes on original data

- season : season (1:springer, 2:summer, 3:fall, 4:winter)
- yr : year (0: 2011, 1:2012)
- mnth : month (1 to 12)
- hr : hour (0 to 23)
- holiday : weather day is holiday or not (extracted from [Web Link])
- weekday : day of the week
- workingday : if day is neither weekend nor holiday is 1, otherwise is 0.
- weathersit :
    - 1: Clear, Few clouds, Partly cloudy, Partly cloudy
    - 2: Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist
    - 3: Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds
    - 4: Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog
- temp : Normalized temperature in Celsius. The values are derived via (t-t_min)/(t_max-t_min), t_min=-8, t_max=+39 (only in hourly scale)
- hum: Normalized humidity. The values are divided to 100 (max)
- windspeed: Normalized wind speed. The values are divided to 67 (max)

3. URL:

https://archive.ics.uci.edu/ml/datasets/Bike+Sharing+Dataset

4. Consulting Project

You have been contacted to build a predictive model to help Bike Rental companies in predicting the hourly and daily demand on bikes.

- Build a first linear model to predict the ‘demands’ and evaluate it (display meanAbsoluteError and r2)
- Improve your model by doing cross validation. You shall tune and cross-validate the model using:
  - pyspark.ml.Pipeline
  - pyspark.ml.tuning.ParamGridBuilder
  - pyspark.ml.tuning.CrossValidator
- Try to get some insights from the results you obtained:
  - Display, for instance, the average real demand versus the average predicted demand and the standard deviation of both by grouping your data by:
    - hour
    - season
    - other features that you think useful
- Add dummy variables to improve the accuracy of your model.
- Try other machine learning algorithms and compare.