

**Data Science with Spark**

**Master Data Science X**

**ING5 ECE Big Data & Analytics**

**2018/2019**

**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 uploaded on the campus page before **Sunday 09/12/2018 midnight.**
- ❖ Late reports will be penalized (3 points/day)

## Abstract:

- ❖ The objective of this mini-project is to use the different Spark machine learning libraries to build predictive models.
- ❖ Two open data sets are provided. In both cases, the correct answers are given. Supervised learning algorithms are thus used.
- ❖ In the first data set, the output is continuous, you shall build several regression models, tune them and compare them
- ❖ In the second, the output is discrete. You shall build several classifiers, tune and compare them

# I. Bike Rental Data Set from UCI Machine Learning Repository

## 1. Citations

Reconsider the Bike Rental data set and the provided notebook.

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

## 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

The first built linear model performs very poorly and the following results were obtained.

meanAbsoluteError	106.06265434811826
r2	0.38740890414713036

After analyzing the results, we added dummy variables on the feature hrs, the model's accuracy was significantly improved but it is still far from being acceptable.

meanAbsoluteError	75.3746
r2	0.670212

Your role is to improve the performance of the linear model to an acceptable stage. You may also decide to change the model and build your proper model. In all cases, you shall tune and cross-validate the model using:

- `pyspark.ml.Pipeline`
- `pyspark.ml.tuning.ParamGridBuilder`
- `pyspark.ml.tuning.CrossValidator`

Provide the necessary steps and methodology (with screenshot) you took for improving the linear and other models. Results shall be commented and analyzed.

## II. Part 2: SMS Spam Collection Data Set

### 1. Understanding the data set

#### a) Direct Link

Direct link on UCI's machine learning repository:

<https://archive.ics.uci.edu/ml/machine-learning-databases/00228/>

#### b) Data Set Information :

This corpus has been collected from free or free for research sources at the Internet:

- A collection of 425 SMS spam messages was manually extracted from the Grumbletext Web site. This is a UK forum in which cell phone users make public claims about SMS spam messages, most of them without reporting the very spam message received. The identification of the text of spam messages in the claims is a very hard and time-consuming task, and it involved carefully scanning hundreds of web pages.
- A subset of 3,375 SMS randomly chosen ham messages of the NUS SMS Corpus (NSC), which is a dataset of about 10,000 legitimate messages collected for research at the Department of Computer Science at the National University of Singapore. The messages largely originate from Singaporeans and mostly from students attending the University. These messages were collected from volunteers who were made aware that their contributions were going to be made publicly available.
- A list of 450 SMS ham messages collected from Caroline Tag's PhD Thesis
- Finally, we have incorporated the SMS Spam Corpus v.0.1 Big. It has 1,002 SMS ham messages and 322 spam messages and it is public. This corpus has been used in the following academic researches:

[1] GÃ³mez Hidalgo, J.M., Cajigas Bringas, G., Puertas Sanz, E., Carrero GarcÃa, F. Content Based SMS Spam Filtering. Proceedings of the 2006 ACM Symposium on Document Engineering (ACM DOCENG'06), Amsterdam, The Netherlands, 10-13, 2006.

[2] Cormack, G. V., GÃ³mez Hidalgo, J. M., and Puertas SÃ¡nchez, E. Feature engineering for mobile (SMS) spam filtering. Proceedings of the 30th Annual international ACM Conference on Research and Development in information Retrieval (ACM SIGIR'07), New York, NY, 871-872, 2007.

[3] Cormack, G. V., GÃ³mez Hidalgo, J. M., and Puertas SÃ¡nchez, E. Spam filtering for short messages. Proceedings of the 16th ACM Conference on Information and Knowledge Management (ACM CIKM'07). Lisbon, Portugal, 313-320, 2007.

### c) Attribute Information:

The collection is composed by just one text file, where each line has the correct class followed by the raw message. We offer some examples bellow:

ham What you doing?how are you?

ham Ok lar... Joking wif u oni...

ham dun say so early hor... U c already then say...

ham MY NO. IN LUTON 0125698789 RING ME IF UR AROUND! H\*

spam FreeMsg: Txt: CALL to No: 86888 & claim your reward of 3 hours talk time to use from your phone now! unsubscribe6GBP/ mnth inc 3hrs 16 stop?txtStop

spam Sunshine Quiz! Win a super Sony DVD recorder if you canname the capital of Australia? Text MQUIZ to 82277. B

spam URGENT! Your Mobile No 07808726822 was awarded a L2,000 Bonus Caller Prize on 02/09/03! This is our 2nd attempt to contact YOU! Call 0871-872-9758 BOX95QU

Note: the messages are not chronologically sorted.

### d) Relevant Papers:

We offer a comprehensive study of this corpus in the following paper. This work presents a number of statistics, studies and baseline results for several machine learning methods.

Almeida, T.A., GÃ³mez Hidalgo, J.M., Yamakami, A. Contributions to the Study of SMS Spam Filtering: New Collection and Results. Proceedings of the 2011 ACM Symposium on Document Engineering (DOCENG'11), Mountain View, CA, USA, 2011.

### e) Citation

<http://www.dt.fee.unicamp.br/~tiago/smsspamcollection/>

## 2. Consulting Project

You have been contacted to build a predictive model to classify an incoming sms as Spam or Safe. You are provided the dataset described above. For each observation in the dataset, the row sms as well as the correct class is given.

- a) The first step is to change the text features into numeric using the suitable classes (StringIndexer, Tokenizer, StopWordsRemover, CountVectorizer, IDF, VectorAssembler).
- b) Then You shall train 4 classifiers and compare them. These are:
  1. LogisticRegression,
  2. DecisionTreeClassifier
  3. RandomForestClassifier
  4. NaiveBayes
- c) For one of these classifiers, you shall tune at least one important hyper parameter using ParamGridBuilder and CrossValidator
- d) Conclusions: Compare and comment the obtained results (you may use a comparison table).

### **Required libraries:**

- pyspark.ml.feature.StringIndexer
- pyspark.ml.feature.Tokenizer
- pyspark.ml.feature.StopWordsRemover
- pyspark.ml.feature.CountVectorizer
- pyspark.ml.feature.IDF
- pyspark.ml.feature.VectorAssembler
- pyspark.ml.Pipeline
- pyspark.ml.classification.LogisticRegression
- pyspark.ml.classification.RandomForestClassifier
- pyspark.ml.classification.DecisionTreeClassifier
- pyspark.ml.classification.NaiveBayes
- pyspark.ml.tuning.ParamGridBuilder
- pyspark.ml.tuning.CrossValidator
- pyspark.ml.evaluation.MulticlassClassificationEvaluator