BBM469 - Data Intensive Applications Laboratory

Assignment 3 : Machine Learning with Spark
Date Issued **:** 04/04/2022
Date Due **:** 18/04/2022
Instructor : Dr. Tugba Gurgen Erdogan

Aim of the Assignment

The aim of this assignment is to make you familiar with the basics of Apache Spark and machine learning methods using Spark Environment.

Background information

Apache Spark is a fast and general-purpose cluster computing framework. It provides high-level APIs in Java, Scala, Python, and R programming languages, and an optimized engine that supports general execution graphs. It also supports a rich set of higher-level tools including Spark SQL for SQL and structured data processing, MLlib for machine learning, GraphX for graph processing, and Spark Streaming (*https://spark.apache.org/docs/latest/*).

We provide you with some basic tutorials for the installation and usage of the Spark system.

You may use ready-made virtual environments (Sandboxes).

- http://hortonworks.com/products/hortonworks-sandbox/#install

Or, install Spark system on your operating system.

- https://medium.com/@GalarnykMichael/install-spark-on-ubuntu-pyspark-231c45677de0

Useful Links

Spark notebooks and tutorials.

- https://github.com/tirthajyoti/Spark-with-Python
- https://github.com/jadianes/spark-py-notebooks
- https://www.datacamp.com/community/tutorials/apache-spark-tutorial-machine-learning
- https://www.tutorialspoint.com/apache_spark/

Your Assignment

In this assignment, you will implement the experiment design provided in the second assignment (Clustering and Classification with Python) with Apache Spark.

1. Download the dataset. The dataset will be shared through Piazza.
2. Use the template file of this assignment which will be shared through Piazza.
3. Choose a clustering method (Kmeans, Kmedoids, etc), and a classification method (naïve Bayes, SVM, Random Forest, etc.).
4. Import and organize the dataset for clustering/classification methods (OD).
5. Normalize the dataset using min-max standardization, and create the normalized dataset (ND). Do not change the original dataset.
6. Cluster the OD dataset according to the class size of the original dataset from step 2 (set k to class size) and create the COD dataset.
7. Cluster the ND dataset according to the class size of the original dataset from step 2 (set k to class size) and create the CND dataset.
8. Present the clustering results
9. Split the datasets into training and test sets (make these steps parametric). Split the ND, OD, COD, and CND datasets with the same proportion and samples.
10. Classify the test dataset with a model trained with the training dataset.
11. Use scatter plots to show the relation between features and clusters/classes (You may use two features on two axes, and values for clusters/classes).
12. Present the classification results for each dataset (classification accuracy, and confusion matrix). You will discuss the results for each sub-experiment in your experiment report with graphs and comments.

Tasks and Grading

You will present your projects online during laboratory hours.

Task-1: In the first task, you will install Apache Spark. Then you will implement (if you wish), examine, and run the “Word Count” application (%20).

Task-2: In the second task, you will implement the design of the second assignment and provide the necessary documents.

- Import dataset, split training, and test dataset (%10)
- Clustering (%20)
- Classification (%20)
- Normalization (%10)
- Visualization (%5)
- Report: You will submit your report and source files before the presentation. Report details will be provided from the Piazza page (%15)

REMARKS:

- Submission format:
    - Assignment3_groupnumber <folder>
        - src.zip
        - report.pdf

- Your submission should match the format above**. 10 point** penalty will be applied to mismatched submissions.
- You will use the online submission system (https://submit.cs.hacettepe.edu.tr/) to submit your experiments. The deadline is 23:59. No other submission method (such as CD or email) will be accepted.
- Do not submit any file via e-mail related to this assignment.
- The assignment must be original, INDIVIDUAL work. Duplicate or very similar assignments are both going to be punished. General discussion of the problem is allowed, but DO NOT SHARE answers, algorithms, or source codes.
- You can ask your questions through the course's Piazza group and you are supposed to be aware of everything discussed in the group piazza.com/hacettepe.edu.tr/spring2022/bbm469/home