BBM469 - Data Intensive Applications Laboratory

Assignment 3	**:** Machine Learning with Spark
Date Issued	**:** 22.05.2020
Date Due	**:** 05.06.2020

Aim of the Experiment

This assignment aims to try to find out how we can diagnose breast cancer, using the machine learning methods using Spark Environment from the features created by digitizing the images of breast cancer. These features define the properties of the cell nuclei in the image. Your primary purpose here is to cluster and classify the data according to the diagnosis (M = malignant, B = benign). You should be able to predict the disease most accurately by using clustering and classification methods using Spark Environment. If there are missing values in the data set, you should explain how to deal with them. Also, you don't have to use all the features in the data set. After analyzing the dataset, you can choose which features you will use for clustering and classification.

At the end of this exercise, you will be familiar with basics of Apache Spark and machine learning methods using Spark Environment.

Background information

Apache Spark is a fast and general-purpose cluster computing system. It provides high-level APIs in Java, Scala, Python and R, 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.
Furthermore information: https://spark.apache.org/docs/latest/

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

(You need to submit your homework over the Colab using pyspark in Jupyter Notebook format. The necessary installation procedures on the Colab are described in the resources below.)

Spark notebooks and tutorials.

- https://github.com/aucan/DataScienceTutorials
- 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/
- https://towardsdatascience.com/pyspark-in-google-colab-6821c2faf41c

Experiment

1. Download the dataset. The dataset will be shared on the Piazza group (also, you can find explanations about the columns in the dataset next to the description file.).
2. Choose a clustering method (Kmeans, Kmedoids, etc.), and a classification method (naïve Bayes, SVM, Random Forest, etc.).
3. Import and organize the original dataset (OD) for clustering/classification methods.
4. Normalize the dataset using min-max standardization and create the normalized dataset (ND). Don't change the original dataset.

   (In the classification phase, you should apply the normalization process after split the data set as train and test.)
5. Cluster the OD dataset according to the class size of the original dataset from step 2 (set k to class size).
6. Cluster the ND dataset according to the class size of the original dataset from step 2 (set k to class size).
7. Present the clustering results.
8. Split the datasets into training and test sets. Split the OD, ND datasets with the same proportion and samples.
9. Classify the test dataset with a model trained with the training dataset.
10. 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)
11. 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.
12. You should submit your codes and report as a single Jupyter notebook.

    (The necessary Jupyter Notebook template that you will use to deliver your report will be shared on piazza.
13. While grading your assignments, we will evaluate your codes with you.

Grading

You will present your projects during the laboratory hours.

- Import dataset, split the data as training and test sets (%10)
- Clustering (%20)
- Visualization of clustering (%20)
- Classification (%20)
- Normalization (%10)
- Report (%20): You will submit your report and code before the presentation.

REMARKS:

- Submission format:
    - studentID_name_surname_hw3.ipynb

- Your submission should be matched with the format above**. 10 point** penalty will be applied on mismatched submissions.
- You will use online submission system to submit your experiments.
- https://submit.cs.hacettepe.edu.tr/ 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 with 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 course's Piazza group and you are supposed to be aware of everything discussed in the group.