CS 6350 - ASSIGNMENT 4

Please read the instructions below before starting the assignment.

- There are 5 parts in this assignment. Please use a separate folder for each.
- You have to submit link to the dataset used for each question, your code and the analysis.
- You should use a cover sheet, which can be downloaded at: http://www.utdallas.edu/~axn112530/cs6350/CS6350_CoverPage.docx
- You are allowed to work in pairs i.e. a group of two students is allowed. Please write the names of the group members on the cover page. Only one submission per team is required.
- The deadline for this assignment is Wednesday October 26 at 11:59 PM. No extensions are allowed.
- You have a total of 4 free late days for the entire semester. You can use at most 2 days for any one assignment. After that, there will be a penalty of 10% for each late day. The submission for this assignment will be closed 2 days after the due date.
- Please ask all questions on Piazza, and not through email to the instructor or TA.

MACHINE LEARNING (ML) USING SPARK

In this assignment, you will perform machine learning tasks using Spark. You can find complete documentation and examples of code at the Spark ML Guide website: http://spark.apache.org/docs/latest/ml-guide.html

You should use the latest version (2.0.1) of Spark. Also, you should use DataFrame based API, since it is the new standard and has replaced RDD for machine learning applications. Try to use pipelines as far as possible for your code. You are free to write your code in either Scala or Python.

The assignment consists of the following machine learning tasks. Please submit a different folder for each task. You are given the flexibility to choose your own dataset, as specified in each section.

I. Classification & Dimensionality Reduction

For the classification part, you have to choose a dataset from the UCI Machine Learning repository, which is one of the most popular repositories and is available at: https://archive.ics.uci.edu/ml/datasets.html

For selecting appropriate datasets, choose "classification" from the "Default task" options at the top left of the page. You are free to choose any dataset that you want. After selecting the dataset, you would need to partition it into training and test parts – it is up to you to define the split. For example, you can choose an 80:20 split on training and testing.

You will then run **any two** of the classification algorithms available on Spark on the dataset: http://spark.apache.org/docs/latest/ml-classification-regression.html#classification
As explained in class, you will use the training portion of the dataset to train your model and the test portion to get the accuracy or other performance measures.

You have to run the classification algorithm twice as follows:

1. In the first attempt, use all of the attributes (features) while training and testing the model.

2. In the second attempt, perform **dimensionality reduction** and reduce the dimensions to some chosen value K. Then, perform the training and testing on these K dimensions and report your results.

3. You will report the accuracy and **two other evaluation metrics**. The evaluation metrics are described here: http://spark.apache.org/docs/latest/mllib-evaluation-metrics.html

The results should be reported in a tabular format like the one shown below:

Dataset name and URL: _____

Number of instances	Total number of attributes	Number of attributes used (in case of dimensionality reduction)	Classification Method	Train/Test Ratio	Test Accuracy/ Error	Other metric 1	Other metric 2

II. Clustering

In this part, you will perform clustering task. You will select another dataset by going to the UCI Machine Learning repository and filtering the default task to clustering. Select a dataset and perform k-means clustering on it. You are free to choose an appropriate value of K.
Report the output of your code. Be sure to mention the URL of your dataset.

III. Regression

In this part, you will perform regression task. You will select another dataset by going to the UCI Machine Learning repository and filtering the default task to regression. Select a dataset and perform **any one type of regression** that is detailed here:
http://spark.apache.org/docs/latest/ml-classification-regression.html#regression

You would need to create a regression model for the dataset and report the value of Root Mean Squared Error (RMSE), Coefficient of Determination (R^2), and Explained Variance on the data. Be sure to mention the URL of your dataset.

IV. Collaborative Filtering

Next, you will perform collaborative filtering, which is part of recommender systems machine learning approach. For this, you can find datasets at either of these two locations:
1. https://gist.github.com/entaroadun/1653794
2. Go to UCI ML repository and set the default task to "other" and then look for datasets that are suitable for Recommender Systems.

You will use the collaborative filtering (CF) interface available in Spark and create a recommender model. You can see details at: http://spark.apache.org/docs/latest/ml-collaborative-filtering.html

You should output the value of Mean Square Error for your dataset. Be sure to mention the source and URL of your dataset.

V. Frequent Pattern Mining

Frequent Pattern Mining searches for patterns that occur frequently in transactions or baskets of data. In this section, you will use the Spark interface for frequent patterns and search for association rules that are above a certain threshold of support and confidence. The details and a sample code is available at:
http://spark.apache.org/docs/latest/mllib-frequent-pattern-mining.html

You can choose a dataset from the Frequent Pattern Itemset repository available at:
http://fimi.ua.ac.be/data/

For the chosen dataset, you should output the association rules that are above a certain threshold of support and confidence. You are free to choose these values. Please be sure to report the source and URL of your dataset.