COMP 6940: Big Data Analytics - Assignment 3

University of the West Indies, St. Augustine

Due Date: April 5th @ 11:59 PM

1 Part 1: Classification and Clustering

The data for this section is stored in a comma-separated file (csv) here.

1.1 Part A (10 marks)

You are required to load, explore and clean the provided dataset. Be sure to look out for values that imply the same across features. This section includes whether or not you choose to use scaling and PCA (if you use PCA, set the variance to 95%). Explain each of your steps and choices using markdown code.

1.2 Part B (60 marks)

1. You will perform binary classification on the dataset to determine if a patient had a stroke or not. You are required to use 3 classifiers and compare the results using appropriate graphs and performance metrics. Your classifiers should include the Random Forest Classifier and the KNN classifier. For KNN, use cross-validation to determine an appropriate value for the number of neighbours. Use markdown code to explain your steps, choices, and results.

2. Explain the purpose of performing hyperparameter tuning.

Feature Description:

- **id**: unique identifier
- **gender**: "Male", "Female" or "Other"
- **age**: age of the patient
- **hypertension**: 0 if the patient doesn't have hypertension, 1 if the patient has hypertension
- **heart disease**: 0 if the patient doesn't have any heart diseases, 1 if the patient has a heart disease
- **ever married**: "No" or "Yes"
- **work type**: "children", "Govt job", "Never worked", "Private" or "Self-employed"
- **Residence type**: "Rural" or "Urban"
- **avg glucose level**: average glucose level in blood
- **bmi**: body mass index
- **smoking status**: "formerly smoked", "never smoked", "smokes" or "Unknown"*
- **stroke**: 1 if the patient had a stroke or 0 if not

3. Evaluate the performance of the machine learning models selected. Make any recommendations (minimum 2) for improvement if necessary.

1.3 Part C (20 marks)

You are required to perform clustering on the dataset using the KMeans algorithm. Your solution should include steps to find a suitable value for "k", as well as graphs showing the results. Use markdown code to explain your steps and results. The last column of the dataset is not needed in this section. State how analysing the resulting clusters (regarding the optimal cluster number only) can aid in decision making.

2 Part 2: Spark Analysis of Amazon Review Dataset

The data for this section is stored in a comma-separated file (csv) here.

You are tasked to use Spark to investigate the following review dataset on appliances. The *asin* (Amazon Standard Identification Number) represents the identifier of the product and *unixReviewTime* represents the time of the review in unix time. Analysis here will solely be based on products, reviewers, and their vote on the product. Note that the digits for the asin go from 0-9 and then a-z (example: 1118461304, B0006GVNOA).

1. What are the implications of loading additional features (that might not be necessary for analysis) into PySpark? [1 mark]
2. Create a sql context from PySpark SQLContext. [1 mark]
3. Load the Amazon Review Dataset into a PySpark RDD. You must ensure that each row is properly separated and the headers are matched to their respective columns. [5 marks]
4. Convert the RDD to a PySpark DataFrame. [1 mark]
5. Using the DataFrame from Question 4, show the top 20 most reviewed products that were purchased in the year 2015. [5 marks]
6. Using the DataFrame from the previous question, show the top 20 reviewers and the products they reviewed in 2015. [4 marks]
7. Referring to Question 5, determine whether the top 20 most reviewed products in 2015 are the same as the top 20 products with the highest average rating in 2015. [5 marks]
8. Create an RDD of tuples from the DataFrame in Question 4 with only 2 columns: *asin* and *reviewerId*, in that order. [1 mark]
9. Using methods from PySpark's RDD object e.g., groupByKey, map, reduceByKey, derive the top 20 products. [5 marks]

   Sample: [('B000AST3AK', 6510), ('B004UB1O9Q', 5702), ...]
10. Create another RDD of tuples from the DataFrame from Question 4 with the columns *reviewerId* and *asin* in that order. [1 mark]
11. Using methods from PySpark's RDD object, produce the top 10 customers who reviewed the most items. The top 10 list must show the *reviewerId* and a list of all the items they reviewed with the number of that item they reviewed. [8 marks]

    Sample: [('A8WEXFRWX1ZHH', ({'B0014CN8Y8': 2, 'B0006GVNOA': 204, 'B00IYZP4AY': 1, 'B00J4V77DO': 1}, 208)), ...]

Submission Details:

- Each section should be in a separate notebook.
- Ensure that your notebooks are named according to the following format: *firstname_lastname_idnumber_sectionnumber.ipynb*.
- Export/download your file from Jupyter notebook.
- There would be a 5% penalty per day for late submissions, up to 5 days.