CS 696 Spring 2019 Exam

Due March 28
Part One 20 points

1. What is the difference between SparkContext and SparkSession?
2. Explain the shuffle operation in Spark operations. Give an example.
3. What is the difference between a Spark transformation and a Spark action.
4. Cleaning Data
    a. What operations do we have on Panda DataFrames to deal with missing values?
    b. What are some of the problems that occur in dealing with missing values in Panda DataFrames?
    c. What operations do we have on Spark DataFrames to deal with missing values.
5. What is a categorical variable? Give an example.
6. What are hyperparameters? Give an example.

Part Two 40 points

This part uses the athlete data set in the file ALL-ATHLETES.csv. This dataset contains information about athletes that participated in the London 2012 Olympics.

1. The data requires some cleaning. Make sure that you clean the data before working on the rest of the problems. All solutions should start with the original dataset.
2. Produce a scatter plot of the athletes weight verses height.
3. Produce box plots of the following. How do the weights compare?
    a. Female weights
    b. Male weights
    c. Weight of the male metal winners
    d. Weight of the female metal winners
4. Produce separate swarm plots of the ages of male metal winners and the female metal winners. How do the weights compare?
5. Produce separate violin plots for the ages of the athletes in Archery, Sailing, and Swimming. How do the ages compare?
6. Produce a histogram of the metals won per country, sorted by the number of metals won.

Part Three 60 points

This part uses the movie dataset in the movies.csv file. The data set classifies the movies as either a romance or an action movie. The columns other than "Title", "Genre", "Year", "Rating", "# Votes" and "# Words" in the file are all words that appear in some of the movies in our dataset. The words are stemmed. The column for a word shows the percentage of the total words in the movie were that particular word. So for example the word “the” was 0.043807463 percent of the words that were spoken in the movie “The Terminator”.

1. Using scikit-learn split the movie data into a training and test set. Create three different models from the training set to classify movies as action or romance using K-means, DBSCAN and GaussianNB.

2. Compare the performance of the three classifiers using accuracy score and confusion matrix. Which classifier is better? Why?

3. Compare two cluster results using Adjusted Rand index, Homogeneity, completeness and V-measure and Silhouette Coefficient.

4. The movie dataset has a lot of features. Use PCA to reduce the dataset to 30 independent variables. Repeat 1-3 on the reduced dataset.
   a. How do the result compare using the reduced dataset
   b. How much of the variation of the original dataset is in the new 30 dimensions.

What to turn in

You are to turn in a Jupyter Python notebook containing the code and answers to the questions above. Since Jupyter notebooks can contain text and code, before each problem indicate which problem it is in text, not in code comment.

Late Penalty

An assignment turned in 1-4 days late, will lose **10**% of the total value of the assignment per day late. After day 4 the late penalty will be **20**% per day.

No Team work

You are to work alone on this exam. Software will be used to examine all students work to look for possible copying and/or people working together.