CS 649 Intro to Big Data
Spring Semester, 2022
Exam

San Diego State University -- This page was last updated 3/16/22

Due Mar 25 23:59

Part One 20 points

In part one, you are to answer the question in your own words. If you use references, you need to provide them. Violations will incur a penalty.

1. (2 pts) What is the difference between SparkContext and SparkSession?

2. (2 pts) What is the difference between a Spark transformation and a Spark action.

3. (3 pts) Cleaning Data

    a. What operations do we have on Panda DataFrames to deal with missing values?

    b. What are some problems in dealing with missing values in Panda DataFrames?

    c. What operations do we have on Spark DataFrames to deal with missing values.

4. (2 pts) What is a categorical variable? Give an example.

5. (2 pts) What are hyperparameters? Give an example.

6. (9 pts) What are two ways to determine a good value for the number of clusters when using K-means?

Part Two 20 points

Problem 1. (18 points) The dataset GPA-GRE (gpa-gre.zip on assignment page) contains 16 years of GPA and the Verbal and Quantitative GRE scores for graduate students.

The file contains four columns. Each row represents a single student. The columns are:

> Year - The year the student graduated. Values from 1 to 16.
> GPA - The GPA of the student when they graduated.
> Verbal - The verbal GRE score obtained by the student. As this is historical data, the scores are in the old scale from 200-700.
> Quant - The quantitative GRE score obtained by the student. Again in the old scale.

a. Use linear regression with the combined GRE score (Verbal + Quantitive) as the independent variable and the GPA as the dependent variable to create a model that predicts a student's GPA given their combined GRE score. Show the regression equation.

b. Plot GRE versus GPA data with the regression line.

c. Are GRE scores a good indicator of students' grades in graduate school? Use the adjusted R^2 to support your answer.

Problem 2. (20 points**)** Comparing TfidfVectorizer and CountVectorizer in classifying text. Using sklearn.datasets.fetch_20newsgroups create training set and test set on the newsgroups: comp.os.ms-windows.misc, sci.electronics, comp.sys.ibm.pc.hardware, comp.sys.mac.hardware, comp.graphics. Using Multinomial Naive Bayes, create a model to classify newsgroup postings as belonging to one of the groups. Create a heat map of the confusion matrix when using TfidfVectorizer and when using CountVectorizer. Using the two heatmaps, compare the effectiveness of the TfidfVectorizer and CountVectorizer.

Problem 3. (32 points) Some people claim widespread election fraud in the last USA presidential election. Often such widespread fraud can be detected. In this problem, you will look at the 2012 Russian presidential election to see evidence of fraud.

a. Data Cleaning. The first task is to read the data into dataframes and perform any needed data cleaning. Did you need to perform any data cleaning to be able to answer the simple queries given in #2?

b. For each district, compute the turnout rate. That is the number of votes divided by the number of voters in each district. Produce a histogram of the turnout rate in each country per district.

c. Produce a scatterplot of votes obtained by the winners in each district and the turnout rate. If there was election fraud or pressure on voters to vote for a given party, one might see a high percentage of votes for the winners in some districts. In particular, some claim that many districts with nearly 100% voter turn and almost 100% votes for the winner indicate election fraud.

d. Given the number of districts in Russia, the plot in c will likely contain many points that can hide information. It is not clear if the density of points varies in the plot. One way to handle this is to set a low alpha for the color of each point so you will see dark areas where there are more points clustered. Another way to handle this is to select a random sample of the data. Use one of these methods to produce a scatterplot to see if there is a hotspot of districts with early 100% voter turn and nearly 100% votes for the winner.

The Data - (Russia2012.zip on the assignment page)

Russia2012_1of2.csv
Russia2012_2of2.csv

These files contain data for the 2012 presidential election in Russia. Five people ran for office. Vladimir Putin won the election. The second file is a continuation of the first file. The columns in the files are:

Name of the district, region, or Republic
Number of the polling district (unique to the district, not overall)
Number of voters included in the voter's list
The number of ballots received by the precinct election commission
The number of ballots issued to voters who voted early
The number of ballots issued to voters at the polling
The number of ballots issued to voters outside the polling station

The number of canceled ballots
The number of ballots in mobile ballot boxes
The number of ballots in the stationary ballot boxes
Number of invalid ballots
Number of valid ballots
The number of absentee ballots received
The number of absentee ballots issued to voters at a polling station
The number of voters who voted with absentee ballots
The number of the unused absentee ballots
The number of absentee ballots issued to voters of the territorial election commission
Number of lost absentee ballots
The number of lost ballots
The number of ballots not counted after being obtained
Blank Column
Zhirinovsky Vladimir Zhirinovsky - number of votes received in the polling district
Zhirinovsky Vladimir Zhirinovsky - percent of votes received in the polling district
Gennady Andreyevich Zyuganov - number of votes received in the polling district
Gennady Andreyevich Zyuganov - percent of votes received in the polling district
Sergei M. Mironov - number of votes received in the polling district
Sergei M. Mironov - percent of votes received in the polling district
Mikhail Prokhorov - number of votes received in the polling district
Mikhail Prokhorov - percent of votes received in the polling district
Vladimir Putin - number of votes received in the polling district
Vladimir Putin - percent of votes received in the polling district

What to turn in

All your answers are to be in a Jupyter notebook. If needed, you can use multiple notebooks. Do not include the datafiles when you turn in your exam. At the top of your notebook(s), have a variable that is the path to the directory containing the data files for the exam. I will edit that path when I run your exam. Changing that one path should be enough to run the notebook on my machine. **That path is worth 5 points**.

When you turn in your exam, make sure that all the code results, including graphs and plots, are in the notebook. **This is worth 5 points**.

I will run your exam notebooks. I have a stock conda install. If that is not enough to run your notebook, you need to include the instructions to run your notebooks. You can lose many points if I can not run your notebook.

You need to turn in a zipped version of your notebook file. There are several ways to do this. One is to zip up your notebook file(s). Another is downloading your Jupyter notebook as an IPython Notebook (.ipynb). Note that when you download your exam, it will create a file with the extension .ipynb.json. I will remove the .json extension. Once you have downloaded the assignment, zip it up and then upload the zip file to the course portal.

An exam turned in 1-7 days late will lose 5% of the total value of the exam per day late. The eight-day late penalty will be 40% of the exam, the ninth day late; the penalty will be 60%; after the ninth day late, the penalty will be 90%. Late penalties are always rounded up to the next integer value.