DS 600 Data Mining
Winter 2021 Semester Midterm Exam

Honor pledge: I will abide by the rules which include the followings:

1. I will not receive any unauthorized assistance from students who are simultaneously taking/have taken the exam. I will use course materials for this exam.
2. I will not give any unauthorized assistance to students who are simultaneously taking and to **who have not taken the exam yet**.
3. I will not discuss exam questions or their variants on any social media until all students have participated.

Please write your name with the date:

Your printed name: ______________________________

Date: ______________________________

This exam contains 5 questions, 11 pages (including the cover) for the total of possible **108 points**. **All questions will be graded**. The exam will be graded out of **100 points**.

This exam is to be taken between 9:00 AM EST on January 16, 2022 and 9:30 PM EST on January 21, 2022 and will not be proctored. The submission of the final write-up (typed, no handwritten answers) must be uploaded by 9:30 PM EST on January 21, 2022. You have **3 hours and 30 minutes**. **Please note that longer answers do not imply you will get more credit for the answers**.

Problem 1 (20 points)

Short questions: For the (true/false) questions, **answer only** with **"True"** or **"False"** and **one/two sentences for explanation** (both parts necessary for any credit). For other questions, answer **as concisely as possible**.

(a) (4 points) Describe one of the assumptions used in clustering algorithm.

Solution:

(b) (4 points) **(True/False)** During data preprocessing stage, one can always drop features containing non-numeric values because they will not be useful in modeling.

Solution:

(c) (4 points) Why do we need the testing set and the validation set for building models?

Solution:

(d) (4 points) What is the end objective (goal) of data mining?

Solution:

(e) (4 points) **(True/False)** Hierarchical clustering requires raw data as the input.

Solution:

Problem 2 (32 points)

Exploratory Data Analysis

(a) (8 points) Describe two ways of handling missing values and when you would use them.

Solution:

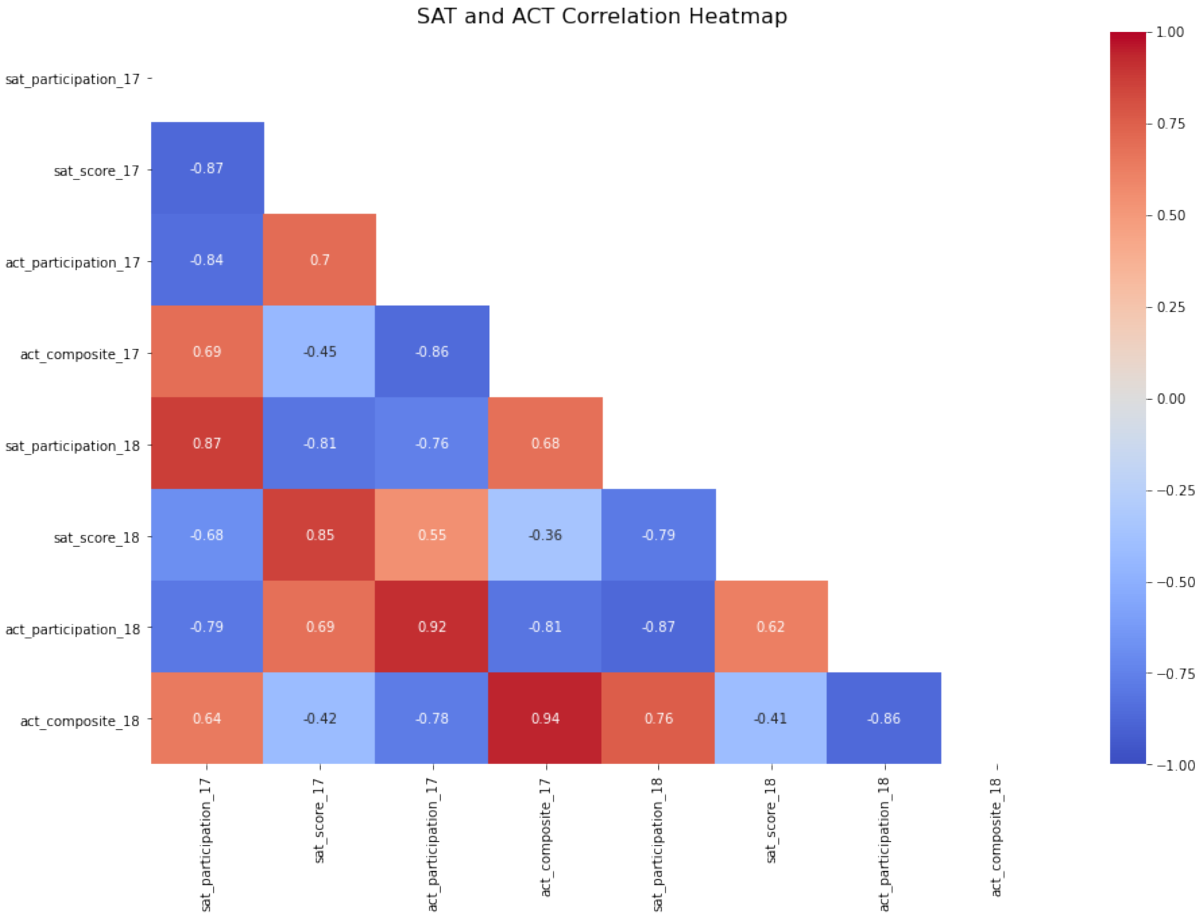
(b) (8 points) Describe what a box plot can show for the distribution of a given feature.

Solution:

(c) (8 points) Describe what a histogram plot can show for the distribution of a given feature. What additional information does it show compared to a box plot of the same feature? What information is less effective in the histogram representation compared to the box plot?

Solution:

(d) (8 points) For the correlation heapmap below, which pair of features most positive correlated and negatively correlated? What does the sign of the correlation mean?


SAT and ACT Correlation Heatmap
sat_participation_17
sat_score_17
-0.87
act_participation_17
-0.84
0.7
act_composite_17
0.69
-0.45
-0.86
sat_participation_18
0.87
-0.81
-0.76
0.68
sat_score_18
-0.68
0.85
0.55
-0.36
-0.79
act_participation_18
-0.79
0.69
0.92
-0.81
-0.87
0.62
act_composite_18
0.64
-0.42
-0.78
0.94
0.76
-0.41
-0.86
sat_participation_17
sat_score_17
act_participation_17
act_composite_17
sat_participation_18
sat_score_18
act_participation_18
act_composite_18
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00

Solution:

Problem 3 (24 points)

Clustering

(a) (8 points) From the class you know that measures of similarity is an important part of clustering algorithms. Compare and contrast two way of defining similarity.

Solution:

(b) (8 points) You are given a data set which does not fit in the main memory of the laptop you are currently working on. Your boss has asked you to produce a clustering of this data set. Which clustering algorithm can you try first? Explain your reasoning.

Solution:

(c) (8 points) You decide to compare spectral clustering and k-means clustering on a given data set. In the context of Part(a), explain how these clustering methods are different and the resulting qualitative differences in the clustering results you may get.

Solution:

Problem 4 (24 points)

Natural Language Processing:

(a) (8 points) Explain the difference between lemmatization and stemming.

Solution:

(b) (8 points) What is the purpose of removing stop words?

Solution:

(c) (8 points) Describe sentiment analysis and one of its application.

Solution:

Problem 5 (8 points)

Regular Expression: Explain in words what each regular expression pattern will match.

(a) (2 points) **j+**

Solution:

(b) (2 points) [ˆ**aeiou**]

Solution:

(c) (2 points) **.at**

Solution:

(d) (2 points) [**chp**]+**art**

Solution: