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 simulta- neously 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.

Mir Ahmed 1/18/2022

Please write your name with the date:

Your printed name: Mir Ahmed

Date: 1/18/2022

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**.

1

**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**.

1. (4 points) Describe one of the assumptions used in clustering algorithm.

**Solution:**

K-Means clustering method considers two assumptions regarding the clusters – first that the clusters are spherical and second that the clusters are of similar size. Spherical assumption helps in separating the clusters when the algorithm works on the data and forms clusters. If this assumption is violated, the clusters formed may not be what one expects.

K-means assume the variance of the distribution of each attribute (variable) is spherical.

1. (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: False**

1. (4 points) Why do we need the testing set and the validation set for building mod- els?

**Solution:**

**Validation Dataset**: The sample of data used to provide an unbiased evaluation of a model fit on the training dataset while tuning model hyperparameters. The evaluation becomes more biased as skill on the validation dataset is incorporated into the model configuration.

**Test Dataset**: The sample of data used to provide an unbiased evaluation of a final model fit on the training dataset.

Validation set is used for tuning the parameters of a model. Test set is used for performance evaluation.

DS-600 Midterm, Page 2 of 11 January 21, 2022

1. (4 points) What is the end objective (goal) of data mining?

**Solution:**

The ultimate goal of data mining is prediction and discovery. The process searches for consistent patterns and systematic relationships between variables, then validates the findings by applying the patterns to new subsets of data.

1. (4 points) **(True/False)** Hierarchical clustering requires raw data as the input.

**Solution: True**

DS-600 Midterm, Page 3 of 11 January 21, 2022

**Problem 2** (32 points)

# Exploratory Data Analysis

1. (8 points) Describe two ways of handling missing values and when you would use them.

**Solution:**

1. **Deleting the row(s):**  If the dataset is huge and the features having null values do not have high importance then I will just delete the rows.

Also,if data is missing for more than 60% of the observations, it may be wise to discard it if the variable is insignificant.

1. **Imputation of mean/ median / mode:** This is one of the most common methods of imputing values when dealing with missing data. In cases where there are a small number of missing observations, data scientists can [calculate the mean or median or mode of the existing observations](https://towardsdatascience.com/how-to-handle-missing-data-8646b18db0d4). However, when there are many missing variables, mean or median results can result [in a loss of variation in the data.](https://blogs.oracle.com/datascience/3-methods-to-handle-missing-data) This method does not use time-series characteristics or depend on the relationship between the variables.

I would use this method if the dataset is too small and the variable that is missing data is significant. If the dataset has too many outliers, then I will use median to replace the missing values. Otherwise, I would use mean.

DS-600 Midterm, Page 4 of 11 January 21, 2022

1. (8 points) Describe what a box plot can show for the distribution of a given feature.

**Solution:**

A boxplot is a standardized way of displaying the distribution of data based on a five number summary (“minimum”, first quartile (Q1), median, third quartile (Q3), and “maximum”). It can tell you about your outliers and what their values are. It can also tell you if your data is symmetrical, how tightly your data is grouped, and if and how your data is skewed.

1. (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:**

A histogram is a graphical representation that organizes a group of data points into user-specified ranges. Similar in appearance to a bar graph, the histogram condenses a data series into an easily interpreted visual by taking many data points and grouping them into logical ranges or bins.

Histogram shows the frequencies of each bin.

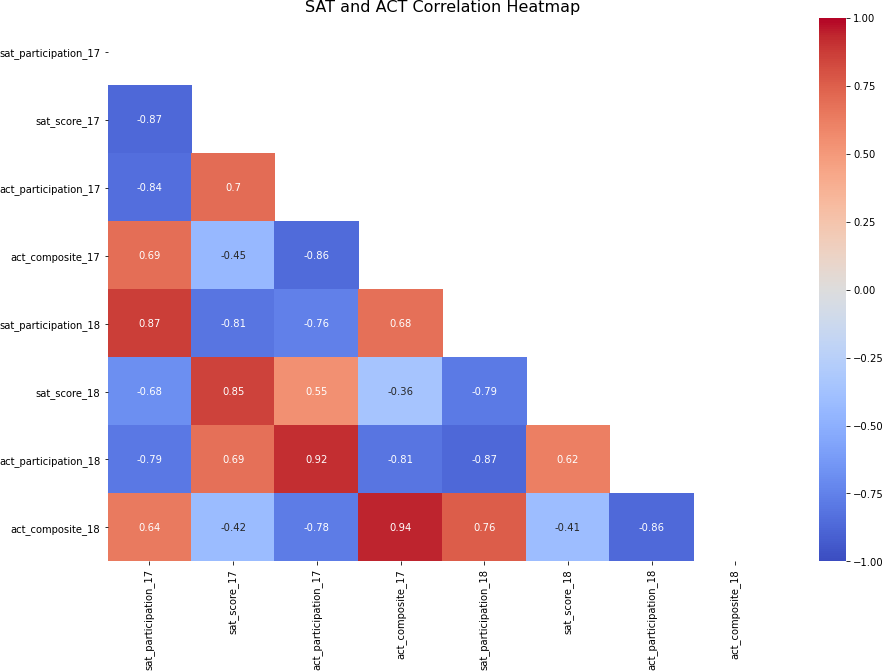
Histograms are preferred to determine the underlying [probability distribution](https://citoolkit.com/articles/probability-distributions/) of a data. Box plots on the other hand are more useful when comparing between several data sets. They are less detailed than histograms and take up less space.

Although histograms are better in displaying the [distribution](https://citoolkit.com/articles/probability-distributions/) of data, we can use a box plot to tell if the distribution is symmetric or skewed. In a symmetric distribution, the mean and median are nearly the same, and the two whiskers has almost the same length.

Histograms give a good sense of the distribution of a variable. Box plots attempt to do the same thing, however, don't give as good of a picture of the distribution of this variable.

1. (8 points) For the correlation heatmap below, which pair of features most positive correlated and negatively correlated? What does the sign of the correlation mean?

DS-600 Midterm, Page 5 of 11 January 21, 2022



**Solution:**

The pair of features with the most positive correlation – act\_composite\_18 & act\_composite\_17 (.94)

The pair of features with the most negative correlation –

1. sat\_score\_17 & sat\_participation\_17 (-.87)
2. act\_participation\_18 & sat\_participation\_18 (-.87)

The sign of the correlation coefficient indicates whether the direction of the relationship is positive (direct) or negative (inverse). Variables which have a direct relationship (a positive correlation) increases together and decrease together.

DS-600 Midterm, Page 6 of 11 January 21, 2022

**Problem 3** (24 points)

# Clustering

1. (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:**

Euclidean distance is considered the traditional metric for problems with geometry. It can be simply explained as the **ordinary distance** between two points. Mathematically it computes the**root of squared differences** between the coordinates between two objects.

This determines the absolute difference among the pair of the coordinates.

Manhattan distance we have two points P and Q to determine the distance between these points we simply have to calculate the perpendicular distance of the points from X-Axis and Y-Axis.  
In a plane with P at coordinate (x1, y1) and Q at (x2, y2).

Manhattan distance between P and Q = |x1 – x2| + |y1 – y2|

Typically, we use Euclidean metric, but Manhattan may be appropriate if different dimensions are not comparable.

1. (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:**

I will use Mini Batch K-means.

Main idea of Mini Batch K-means is to use small random batches of data of a fixed size, so they can be stored in memory. In each iteration a new random sample from the dataset is obtained and used to update the clusters and this is repeated until convergence. Each mini batch updates the clusters using a convex combination of the values of the prototypes and the data, applying a learning rate that decreases with the number of iterations. This learning rate is the inverse of the number of data assigned to a cluster during the process. As the number of iterations increases, the effect of new data is reduced, so convergence can be detected when no changes in the clusters occur in several consecutive iterations.

So, this way I do not have to worry about exceeding main memory size and have the clustering model created in multiple iterations.

DS-600 Midterm, Page 7 of 11 January 21, 2022

1. (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:**

In K-Means Clustering, points that lie close to each other fall in the same cluster and are compact around the cluster center. The closeness can be measured by the distance between the observations.

In Spectral Clustering, points that are connected or immediately next to each other are put in the same cluster. Even if the distance between 2 points is less, if they are not connected, they are not clustered together.

Visually speaking, K-means cares about distance (Euclidean/ Manhattan) while spectral is more about connectivity since it is semi-convex.

Spectral clustering does not cluster data points directly in their native data space but instead form a similarity matrix where the (i,j)−th entry is some similarity distance you define between the i−th and j−th data points in the dataset.

DS-600 Midterm, Page 8 of 11 January 21, 2022

**Problem 4** (24 points)

# Natural Language Processing:

1. (8 points) Explain the difference between lemmatization and stemming.

**Solution:**

**Stemming** algorithms work by cutting off the end or the beginning of the word, considering a list of common prefixes and suffixes that can be found in an inflected word. This indiscriminate cutting can be successful on some occasions, but not always, and that is why we affirm that this approach presents some limitations. So, ‘Studying’ would become ‘Study’ but ‘Studies’ would become ‘Studi’.

Lemmatization, on the other hand, takes into consideration the morphological analysis of the words. To do so, it is necessary to have detailed dictionaries which the algorithm can look through to link the form back to its lemma. Here, both ‘Studying’ and ‘Studies’ would become ‘Study’.

1. (8 points) What is the purpose of removing stop words?

**Solution:**

Here are a few key benefits of removing stopwords:

* On removing stopwords, dataset size decreases and the time to train the model also decreases
* Removing stopwords can potentially help improve the performance as there are fewer and only meaningful tokens left. Thus, it could increase classification accuracy.

DS-600 Midterm, Page 9 of 11 January 21, 2022

1. (8 points) Describe sentiment analysis and one of its applications.

**Solution:**

Sentiment analysis is contextual mining of text which identifies and extracts subjective information in source material and helping a business to understand the social sentiment of their brand, product or service while monitoring online conversations.

One of the applications of sentiment analysis is product analysis.

We can find out what the public is saying about a new product right after launch or analyze years of feedback. We can search keywords for a particular product feature (interface, UX, functionality) and train sentiment analysis models to find only the information we need.

We can discover how a product is perceived by our target audience, which elements of our product(s) need to be improved. Sentiment analysis provides better results than humans because AI doesn’t alter its results and it’s not subjective.

DS-600 Midterm, Page 10 of 11 January 21, 2022

**Problem 5** (8 points)

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

1. (2 points) **j+**

**Solution:**

It will look whether the string has ‘j’ one or more times.

1. (2 points) **[**ˆ**aeiou]**

**Solution:**

It will look for anything but lower-case vowels at the start of the string. It will not match if the string starts with a lower-case vowel.

1. (2 points) **.at**

**Solution:**

It will look for any character followed by ‘at’ in the string.

1. (2 points) **[chp]+art**

**Solution:**

It will look for either one of the characters - ‘c’, ‘h’, ‘p’ followed by ‘art’ in the string.

DS-600 Midterm, Page 11 of 11 January 21, 2022