New York University

Computer Science Department

Predictive Analytics (GA)

Summer 2022

Midterm Exam

Instructions:

Please read these instructions carefully, there will be no excuses if you do not follow these instructions:

1. The exam will start at 11am and the exam will end at 1:30pm.
2. You are not allowed to download the exam; you should keep the exam open in the browser on Brightspace (make sure your laptop is charged, excuses of laptop battery died will not be accepted **as you should remain connected to Zoom throughout the exam, your camera does not have to be on**)
3. The exam will be encrypted/traced and the password will be communicated to you at the beginning of the exam.
4. You will need to provide most your answers on paper, and you can attach the answers back to Brightspace in one PDF document (it is fine if you want to type the answers on a computer, but it might be easier if you write them by hand)
5. You have about 2.5hours to answer all the questions. Regardless of if you complete on time or not, you MUST return your exam on Bright Space by 1:30pm. Any submission after may not be accepted and major penalty will apply including failing the test and zero grade.
6. Cheating is not allowed, and you are NOT allowed to use Google or the web during the exam, googling a question and/or copying an answer will be traced and you will assign a zero in the exam and an F in the course.
7. Notes and books required for this class are allowed to be accessed and used during the exam. Copying an answer exactly as is from the slides will results in zero grade and a report to the administration.
8. You are NOT allowed to share this exam on web, social media or by email, during the exam time, before or after the exam. The exam is encrypted and will be traced.
9. You must hand-write SOME OF the answers *(if you decide to write them on paper)* CLEARLY on paper and attached your answers as images into pdf. (Your final answers document should have a .pdf extension.)
10. Write clearly and organize your answers clearly (specify the question's number and letter)

11. If your answers are NOT clear and not organized, you will receive a penalty.
12. Cheating and communicating about the exam during the exam time, after or before is strictly prohibited and not allowed. You will be reported to the administration.
13. Your zoom session must be ON during the whole session, disconnecting from the session will results in a zero on the exam.
14. If you have any question, **write on the chat to the instructor**, he/she will get back to you.
15. Any violation of the instructions above will result in a penalty and a null grade.

A. True or False

1. K-means algorithm ***you used in HW1*** forces all the points to join clusters and ignores outliers **(T/F)**

2. K-means algorithm ***you used in HW1*** tend to have a polynomial Big-O running time in the worst-case O(n^2) (T/F)

3. **In feature selection**, filters have the tendency to select large subsets because the filter objective functions are generally **not monotonic (T/F)**

4. **Filters** generally achieve better recognition rates than most *wrappers* because they are tuned to the specific interactions between the classifier and the dataset**. (T/F)**

5. From your reading of chapter 5 and 6 of the PA book, Data classification and regression are examples of unsupervised learning **(T/F)**

6. A classification model must have a **90% accuracy** on the training dataset to have a good accuracy on the testing dataset **(T/F)**

7. About 80% of the time in the data analytics project is spent in developing and deploying the analytic and the predictive models **(T/F)**

8. In feature engineering: **In a Filter**, the objective function is a classifier, which evaluates feature subsets by their predictive accuracy (recognition rate on test data) by statistical resampling or cross-validation. **(T/F)**

9. In feature engineering: **A Wrapper** relies a correlation coefficient to measure the strength of association between two variables **(T/F)**

10. SVD can be used a data reduction technique (**T/F**)

11. Data reduction techniques always preserve the original units of the attributes in the original matrix **(T/F)**

12. A feature that has high correlation with the class label will/must eventually end up in the set of features generated by a forward selection algorithm and NOT by a backward selection algorithm **(T/F)**

13. *Supervised learning* is learning (extracting insights) from labeled data**(T/F)**
14. In *unsupervised learning,* unseen data is labeled with the class of the observation **(T/F)**
15. In *unsupervised machine learning*, the learning is done from unlabeled data **(T/F)**

B. Data Preparation

Principal Component Analysis (PCA)

1. Consider a Matrix M of n rows and m columns. Briefly describe in **pseudo code** the algorithm of reducing the dimensions of matrix M using *PCA (double [][] M, int n, int m)*
   The input to your algorithm is a **n * m matrix.**

2. Consider the following matrix of users' review of products:
   Show all steps and calculations of applying PCA to reduce the matrix to n = 4 by m=1 column.

   Circle the final output reduced matrix.

	Review of Product One	Review of Product Two
User 1	1	2
User 2	2	1
User 3	3	5
User 4	4	3

Singular Value Decomposition

Consider the following the following matrix of Users vs. Movies Reviews' matrix:

	Matrix	Alien	Star Wars	Casablanca	Titanic
Joe	1	1	1	0	0
Jim	3	3	3	0	0
John	4	4	4	0	0
Jack	5	5	5	0	0
Jill	0	2	0	4	4
Jenny	0	0	0	5	5
Jane	0	1	0	2	2

The SVD decomposition of the Matrix M' is as follows:

$$\begin{bmatrix} .13 & .02 & -.01 \\ .41 & .07 & -.03 \\ .55 & .09 & -.04 \\ .68 & .11 & -.05 \\ .15 & -.59 & .65 \\ .07 & -.73 & -.67 \\ .07 & -.29 & .32 \end{bmatrix} \begin{bmatrix} 12.4 & 0 & 0 \\ 0 & 9.5 & 0 \\ 0 & 0 & 1.3 \end{bmatrix} \begin{bmatrix} .56 & .59 & .56 & .09 & .09 \\ .12 & -.02 & .12 & -.69 & -.69 \\ .40 & -.80 & .40 & .09 & .09 \end{bmatrix}$$

U Σ V^T

3. Explain in ***(to the point in bullet points***) ALL insights that you can learn from the decomposition showed above.

4. Explain in **bullet points** how will use this decomposition **to reduce** M' and represent movies and users in reduced representation (you do not have to perform the calculations)

5. Suppose a user X assigns rating ***3.5 to Alien and Rating 6.5 to Titanic***. Find the representation of user X in a hiden concept space. What can you deduce in terms of which hidden concept user X will belong to?

6. If you had the ratings of a specific movie n by m number of users. Provide the pseudo code of **an algorithm inspired from SVD** that can ***recommend*** the people in your matrix that would like a given movie n. Be as clear as detailed as possible and **write the algorithm as a method in clear pseudo code, not steps in English.**

The recommender system in this scenario will be purely collaborative filtering based.

C. Into to Text Mining and the Vector Space Model

Consider the following short texts:

D1: The sky is blue and very clear.
D2: The sun in the sky is bright.

1. Convert these documents into a Vector Space Model (VSM) and specify the word representation metric you are using for the document representation. Briefly explain if any preprocessing will be performed on the text.

2. Consider the following document D3: **"You can see the sun in the sky"**. Use one the cosine similarity to discover the most similar document to D3. Explain your approach in detail.

3. Write a generalized algorithm that takes a raw document as an input and returns a predicted label.
Make your own assumptions and your own design.
We will grade the assumptions, the design and the
algorithm.

4. Briefly explain the following concepts adopted in your HW1 and provide brief examples:

- Named Entity Recognition
- Stanford NLP
- Stemming
- Sliding Window
- Lemmatization
- N-grams
- Term-document matrix
- Cluster Visualization
- Precision
- Recall

D. Feature Selection

Towards Predicting Life Insurance' Discount based on our footsteps, heart rate or body movements (35pts) ***use log base 2 for your calculations***

Per an article published in CNNMoney (New York) on April 8, 2015, "for the first time ever in the United States, a life insurance company is offering a discount -- if you're willing to let it track your health, location and body" The life insurance company is offering a discount if you'll wear one a fitness tracker that measures your lifestyle.
A snapshot of a synthetic training historical dataset (imaginary dataset for this exercise of past customers) is shown below:
The predictive target (label) attribute is Discount

Customer Name	Zip Code	Age	Lifestyle (Health)	Discount
John	20037	30-40	Good	6%-15%
Zach	20052	20-30	Average	1%-5%
Srini	10011	40-50	Average	1%-5%
Mariah	10018	30-40	Good	6%-15%
Ana	10020	20-30	Good	1%-5%
Sarah	10036	30-40	Average	1%-5%

1. Could the attribute ***Zip Code*** be a predictive feature to label ***Discount*** per the samples shown in the dataset above? Provide a ***mathematical*** explanation to your answer.
2. Using the **Information Gain (IG)** and **Entropy** concepts, evaluate the predictive power of **Lifestyle and Age**. Which feature is more predictive than the other? Provide the **IG** of both features
3. Provide a rule-based algorithm in pseudo code (or clear steps) that will use the insights from *Information Gain* calculations made in (2) of Age and Lifestyle to develop ***decision rules*** on Discount based on Age and Lifestyle?

4. Apply the decision rules you developed in (3) to predict the Discount that would be given to

   {Customer Name = David, Age 20-30, Lifestyle=Average, **Discount=?**}
   {Customer Name = Jane, Age=40-50, Lifestyle= Good, **Discount**=?}

5. We covered in class *Forward Selection Algorithm.* **List one main drawback of the algorithm.** Design an algorithm (show detailed ***pseudo-code*** of the algorithm) that performs feature selection in the following way: it starts with fitting a model with all features then the least significant feature is being dropped following a greedy criterion (objective function to be minimized) that would you define. The algorithm would output an optimal set of features of high predictive value.

   List one drawback of the algorithm you create.

Provide the Algorithm in *pseudo-code*

E. Finding Similar Customers

1. List and explain **two methods** for computing the similarity in **Categorical Data**

2. Suppose we have four customers and each has **rated** both products A and B as shown in the table below.

Customer ID	Ratings of Product#A	Ratings of Product#B
1	1	1
2	4	3
3	5	4
4	2	1

Consider building a recommender system that will recommend products to customers.

Consider using the following algorithm that groups similar data objects for the phase of neighborhood creation (similar customers) in which similar customers are being discovered:

Choose two data points from your dataset as initial points (referred as centroids)
Iterated through the data points and assign each data points to the group that has the closest centroid (using a similarity measure) **(2)**

When all data points have been assigned, recalculate the positions of two centroids (by averaging the ratings for Product#A and Product#B of the data points that belongs to the same group) **(3)**

Repeat Steps 2 and 3 until the centroids no longer change.
This algorithm should produce a separation of the objects into groups

Question (2): Apply the algorithm of specified above in 2 to the dataset shown above in the table, **you must choose customer 1 and 2 as initial centroid.**

Show ALL steps and please highlight the final results of your clusters.
Use *Cosine similarity*

3. The choice of the initial centroids affects the results. Briefly propose an algorithmic approach in pseudo-code to choose initial centroids in such a way that the number of iterations could be ***minimize.***

F. Conceptual Questions

1. Consider a matrix of n rows and m columns that represent Patients (rows) and Genes' expressions (columns), the matrix has several missing values, the values are not in a specific range, there was no variance analysis or features correlation analysis done on the data. Propose a set of steps (in bullet point format) you would take to examine the data and prepare it for analytics (right before you apply a predictive model (assume the data comes with a predictive column – tumor (0) vs cancer (1))
2. Imagine you are in a situation where you are in an interview for a data scientist job: the interviewer is asking you the following questions (answer as briefly as possible – 2 to 3 sentences are perfectly fine):
    2.1. How is PA being applied in the field of politics? What is an example of a successful use case? Are there any specific predictive algorithms being used in politics analytics?
    2.2. Can you give an example where new datasets are being used to make data-driven investment decision in finance...?
    2.3. How would you apply the concept of "Alternative Data Sources" in the world of health care analytics? What data sources could bring relatively new value in the health analytics and for which data science problem(s) in healthcare?
3. Correlation does not always refer to causation: explain the concept from Granger Causality we discussed from the research paper "Twitter mood predicts the stock market"
4. In latest Pandemic Book, Bill Gates discusses the need for a more sophisticated disease surveillance. 4.1 Explain the difference between active disease surveillance and passive disease surveillance. 4.2 Briefly explain in your own words how alternative data and predictive analytics could help in setting a new way of disease surveillance.