FAST- National University of Computer & Emerging Sciences, Karachi.

FAST School of Computing,

Final Exam, Spring 2021

10th June, 2021, 12:30 pm – 03:30 pm

Course Code: CS 481	**Course Name:** Data Science
Instructor Names: Dr. Muhammad Nouman Durrani and Muhammad Sohail Afzal	
Student Roll No:	**Section No:**

Instructions:

- Read each question completely before answering it. There are 9 questions on 5 pages. Page 6 is the reference materials sheet.
- In case of any ambiguity, you may make an assumption. But your assumption should not contradict any statement in the question paper
- Show all steps clearly

Time Allowed: 180 minutes **Maximum Points**: 53

==

Short Questions 25-35 Minutes

Q . No. 1 Briefly answer the following short questions: [1 x 13 = 13 Points]

i. What is the difference between underfitting and overfitting?

ii. In k-NN, which distance measure do we use in the case of categorical variables?

iii. The k-NN algorithm does more computation on test time rather than train time. True or False**.** Why?

iv. What does it mean if the training and testing accuracies of the machine learning model are closer to each other?

v. What do you think when to use precision over recall as an evaluation metric in any machine learning problem?

vi. Write at least two possible termination conditions for the K-Means algorithm?

vii. Do we use k-fold cross-validation to improve the performance of our model? If yes then how, if no then why?

viii. Computationally we can argue that ensembles can be used to build good models. Why?

ix. When do we use weighted average and maximum voting in ensemble learning?

x. Discuss the difference between stacking and blending.

xi. In AdaBoost, the weighted training error ϵ_t of the t^{th} weak classifier on training data with weights D_t tends to increase as a function of t. True or False? Why?

xii. In boosting, the individual base learners can be used in parallel: True or False. Why?

xiii. What are the assumptions in Kendall's Tau correlation?

xv.
xvi.

Numerical Questions related to Machine Learning Algorithms:

Time Distribution	Q. No. 2	Q. No. 3	Q. No. 4	Q. No. 5	Q. No. 6	Q. No. 7
	15-20 Minutes	20 Minutes	15-20 Minutes	20 Minutes	25 Minutes	5 Minutes

Q . No. 2 Apply k-means algorithm. Consider the following <x1,x2> pairs. [2+2 = 4 Points]

x1	x2
1.76	0.84
2.31	2.09
5.02	3.02
2.25	3.47
3.17	4.96

a) Consider that you are given {cluster1: (1.76, 0.84)}, {cluster2: (3.17, 4.96)} as the initial assignment for the first and second cluster center. What are the cluster assignments after ONE iteration for the k-means (k=2) algorithm? Assume k-means uses Euclidean distance.

b) Suppose you are given {cluster1: (1.76, 0.84)}, {cluster2: (3.17, 4.96)}, and {cluster 3: (5.02, 3.02)} as the initial assignments for the three cluster. Use your best understanding to calculate the Within-Cluster-Sum-of-Squares (WCSS) using the formula:

$$\text{WCSS} = \sum_{C_k}^{C_n} \left(\sum_{d_i \, in \, C_i}^{d_m} distance(d_i, C_k)^2 \right)$$

Where,

C is the cluster centroids and d is the data point in each Cluster.

Q. No. 3 a) In this problem, the following dataset has been used to learn a decision tree which predicts that if students pass Introduction to Data Science (Yes or No), based on their previous CGPA (High, Medium, or Low) and their study (Yes, No). [1+ 1 + 1 + 2 = 5 Points]

CGPA	Study	Pass
L	Y	Y
M	N	N
L	N	N
M	Y	Y
H	N	Y
H	Y	Y

i. What is the entropy H(Pass)?

ii. What is the entropy H(Pass | CGPA)?

iii. What is the entropy H(Pass | Study)?

iv. Which attribute would you consider as the root node? What was the information gain of the attribute you chose as the root node?

b) According to the naive Bayes classifier, what is the probability P (Pass= Y | CGPA= H ∧ Study = N)? [2 Points]

Q. No. 4 Let A be an *m* x *n* matrix of data points: $A = \begin{bmatrix} 1 & 0.45 \\ 0.23 & 0.87 \end{bmatrix}$ [1+2+1+1 = 5 Points]

Also, consider the following initial code:

i. Calculate $\boldsymbol{AA^T}$.

ii. Calculate the eigenvalues $\boldsymbol{\lambda_i}$ for $\mathbf{A} \bullet \mathbf{A^T}$. Also find the eigenvectors $\boldsymbol{V_i}$ of $\boldsymbol{A} \bullet \boldsymbol{A^T}$, using the eigenvalues $\boldsymbol{\lambda_i}$.

iii. What proportion of the total variance in the data does the first principal component account for?

iv. Now, construct a matrix *E*, the matrix of eigenvectors $\boldsymbol{V_i}$ for the matrix $AA^{T,}$ and use the concept of PCA to compute the resultant matrix $\boldsymbol{E} \bullet \boldsymbol{A}$.

Q. No. 5 Consider the following study of sugar consumption in a particular cold drink for two months and its related Hemoglobin A1c (HbA1c), of 5 student volunteers at the State University. After 45 days the HbA1c Test for Diabetes was conducted to observe the change from the normal range. [2 + 1 + 1 + 2 = 6 points]

Number of Cold Drinks Taken (X)	5	2	9	8	3
HbA1c (Y)	8.20	5.90	9.24	8.95	6.14

i. The regression equation is a linear equation of the form: $\hat{y} = b_0 + b_1x$. Show the computation steps for the regression coefficient (b_1) and slope (b_0).

ii. Interpret the slope (how many units of Y are changing by changing how many units of X)

iii. Find the coefficient of determination R^2 , if the sum of square due to regression (SSR) = 9.1585 and the sum of square due to error (SSE) = 0.6982. What coefficient of determination R2 indicates in this example?

iv. Calculate a Pearson's correlation on the data *xi* and *yi* given in the above table.

Q. No. 6 a) Consider the following four transactions. [3 + 3 = 6 Points]

TID	Items_bought
001	A, B, K, D
002	A, B, C, D, E
003	A, E, C, B
004	D, A, B
005	A, C, D, E

Suppose a minimum level of support min_sup= 3 and a minimum level of confidence min_conf= 80%:

i. If the support threshold is 60%, and a minimum level confidence min_conf= 80%, find all frequent itemsets using the Apriori algorithm. For each iteration show the candidate and acceptable frequent itemsets.

ii. List all strong association rules, along with their support and confidence values.

b) Consider the following two Documents: [2 Points]

Document 1: The bus is driven on the motorway by YFN. Document 2: The truck is driven on the highway by FiN

Calculate the TF-IDF for the above two documents, which represent our corpus.

Q. No. 7 Draw a histogram on paper for bivariate analysis of data in the below figure. [2 Points]

Hair Color	Eye Color				
	Blue	Green	Brown	Black	Total
Blonde	2	1	2	1	6
Red	1	1	2	0	4
Brown	1	0	4	2	7
Black	1	0	2	0	3
Total	5	2	10	3	20

Programming Part: 30-35 Minutes

Q. No. 8 We‘ve grabbed some web page contents and saved it in a data frame. Then, we will analyze the text to see what the page is about by performing the following NLP operations.

text="""Hi Mr. Mohsin, how are you doing today? The weather is great, and the city of Karachi is awesome today. The sky is blue. You should visit the seaside."""

i. Convert the text into tokens. [0.5 Points]

ii. Print the total number of sentences in the file. [0.5 Points]

iii. Remove stop words from the above text. Hint: clean_tokens.remove(token) [1 Point]

iv. Part-of-speech (POS) tagging is used to assign parts of speech to each word of a given text (such as nouns, verbs, pronouns, adverb, conjunction, adjectives, interjection) based on its definition and its context. Write code to assign parts of speech to the above text. [1 Point]

v. Differentiate the concept of stemming and lemmatizing and *apply* it on the extracted text. [1 Point]

vi. Consider the following two Documents:

Document 1: The bus is driven on the motorway by Mohsin. Document 2: The truck is driven on the highway by Mohsin.

Write python code to find the TF-IDF for the above two documents, which represent our corpus. [1 Point]

Question 9: This Problem is about calculating different aggregates using data visualization techniques. Consider the following financial data of K-Electric between June 13, 2020 to June 18, 2020. Also, consider financial data stored in **fdata.csv**, which is loaded into the Pandas DataFrame df as:

```
# For this problem you can use Jupytor notebook for coding

import matplotlib.pyplot as plt

import pandas as pd

df = pd.read_csv('fdata.csv')

============================================================

Sample Financial data (fdata.csv):
Date,        Open,        High,        Low,         Close
06-13-16,    778.23,      776.065002,  769.50,      772.559998
06-14-16,    776.030029,  778.710022,  772.890015,  776.429993
06-15-16,    779.309998,  782.070007,  789.236,     776.469971
06-16-16,    779.0214     780.47998,   775.539978,  776.859985
06-17-16,    779.659973,  779.659973,  770.75,      770.080017

===========================================================
```

a) Write a Python code to draw a line charts considering the above data. [1 Point]

b) Suppose we are now working on another dataset. We found that the survey *response* data is categorical, and we might want to count how many times each category appears. Plot the number of times a particular value appears in the Response data. [1 Point]

c) Suppose our data has many outliers, in that situation, we might also want to plot the median. Plot the median of the Response column. [1 Point]

BEST OF LUCK!

Reference Materials Sheet

$R^2 = \{ (1 / N) * \Sigma [(x_i - x) * (y_i - y)] / (\sigma_x * \sigma_y) \}^2$

$\sigma_x = \text{sqrt} [\Sigma (x_i - x)^2 / N]$
$\sigma_y = \text{sqrt} [\Sigma (y_i - y)^2 / N]$

Natural language Processing:

Select appropriate function/s from the following list to perform NLP related tasks:

nltk.tokenize	sent_tokenize	clean_tokens.remove(token)
split()	tokenized_text	FreqDist
Plot	Show	filtered_sent.append
wordnet.synsets	stemmer.stem	lemmatizer.lemmatize

```
df1 = pd.DataFrame({'Java': [data1], 'Python': [data2], 'Go': [data2]})

vectorizer = TfidfVectorizer()  # Initialize

doc_vec = vectorizer.fit_transform(df1.iloc[0])

 # Create dataFrame

df2 = pd.DataFrame(doc_vec.toarray().transpose(),index=vectorizer.get_feature_names())

df2.columns = df1.columns
```

```
vectorizer.fit_transform(corpus)

vectorizer.get_feature_names())

Pipeline([('count', CountVectorizer(vocabulary=vocabulary)),('tfid', TfidfTransformer())]).fit(corpus)
```

IDF:

idf(t) = log [n / (df(t) + 1)])

seaborn.barplot(*, x=None, y=None, hue=None, data=None, order=None, hue_order=None, estimator=<function mean at 0x7fecadf1cee0>, ci=95, n_boot=1000, units=None, seed=None, orient=None, color=None, palette=None, saturation=0.75, errcolor='.26', errwidth=None, capsize=None, dodge=True, ax=None, **kwargs)

seaborn.lineplot(*, x=None, y=None, hue=None, size=None, style=None, data=None, palette=None, hue_order=None, hue_norm=None, sizes=None, size_order=None, size_norm=None, dashes=True, markers=None, style_order=None, units=None, estimator='mean', ci=95, n_boot=1000, seed=None, sort=True, err_style='band', err_kws=None, legend='auto', ax=None, **kwargs)

Euclidean distance formula:

$$d(\mathbf{p}, \mathbf{q}) = \sqrt{\sum_{i=1}^{n} (q_i - p_i)^2}$$