National University of Computer & Emerging Sciences, Karachi

Spring 2021 CS-Department

Final Exam

Monday, 14th June 2021, 9:00 am – 12:00 pm

Course Code: CS481	**Course Name:** Data Science
Instructor Name: Dr Muhammad Atif Tahir	
Student Roll No:	**Section No:**

Instructions:

- Return the question paper.
- Read each question completely before answering it. There are **10 questions and 5 pages.**
- In case of any ambiguity, you may make assumption. But your assumption should not contradict any statement in the question paper.
- In each question, show all steps clearly.

Time: 180 minutes. **Total Marks**: 50 points

Question 1 [7.5 Points]: Briefly answer the following questions. Each questions should be answered in maximum *20* words including articles. Otherwise, answer will not be checked.

a) What is the main idea behind Inverse Random Under Sampling?
b) What is the purpose of slack variable in SVM?
c) What is the purpose of kernel in SVM?
d) What is Hyperplane?
e) Why there is a need to replace least square estimation with some alternative fitting procedure?
f) How Shrinking (Regularization) can be used for feature selection?
g) Can you think of a real-world application in which stop words would be useful for text classification?
h) Explain the difference between feature selection and feature extraction?
i) Explain the difference between hold out, cross validation and leave one out approach?
j) What is the difference between classification and regression?
k) What is the role of activation functions in neural networks?
l) What will happen when learning rate is too high in neural networks?
m) Is it true that the number of neurons in output layer should match the number of classes i.e. where is the number of classes are greater than 2
n) What is gradient descent?
o) Why square loss function cannot be used in Logistic Regression?

Question 2 [5 Points]: Consider a Database D below which consists of 5 transactions. Let min_sup= 2 and min_conf = 80%.

(a) [3 Points] Find frequent Itemsets using Apriori Algorithm.
(b) [2 Points] List all the strong association rules from the itemset with the highest number of items

TID	Items_bought
1	{K, A, R, C, I}
2	{A, H, O, R, E]
3	{H, Y, D, E, R, A, B}
4	{I, L, M, A, D}
5	{U,K,A,R}

Question 3 [4 Points] Consider the following data matrix (M) consists of 2 features and 4 instances.

Feature 1	Feature 2
-1	1
1	-1
3	4
4	3

(i) Use Principal Component Analysis to find the first principal component of the above data matrix (M). In other words, reduce the number of dimensions to 1 (3 Points)

(ii) Let's assume that instead of starting with M^TM, you would like to examine eigenvalues using MM^T. What do you will be the eigenvalues using MM^T? Are they going to be same or different (1 Point)?

Question 4 [5 Points] Clustering is an unsupervised machine learning technique which divides the given data into different clusters based on their distances (similarity) from each other. The unsupervised k-means clustering algorithm gives the values of any point lying in some particular cluster to be either as 0 or 1 i.e., either true or false. But the fuzzy logic gives the fuzzy values of any particular data point to be lying in either of the clusters. Here, in fuzzy c-means clustering, the centroid of the data points are first find out and then calculate the distance of each data point from the given centroids until the clusters formed becomes constant. Let's assume the data points are **{(1, 3), (2, 5), (6, 8), (7, 9)}**

Following are the major steps in Fuzzy Clustering and you are requested to perform missing steps below

Step 1: Initialize the data points into desired number of clusters randomly:

Let's assume there are 2 clusters in which the data is to be divided, initializing the data point randomly. Each data point lies in both the clusters with some membership value which can be assumed anything in the initial state. The table below represents the values of the data points along with their membership (gamma) in each of the cluster.

Cluster	(1, 3)	(2, 5)	(4, 8)	(7, 9)
1	0.8	0.7	0.2	0.1
2	0.2	0.3	0.8	0.9

Step 2: Find out the centroid

The formula for finding out the centroid (V) is:

$$V_{ij} = (\sum_{1}^{n}(\gamma_{ik}^{m} * x_k)/\sum_{1}^{n}\gamma_{ik}^{m}$$

where, ɣ is fuzzy membership value of the data point, m is the fuzziness parameter (generally taken as 2), and x_k is the data point.

For example, V_{11} is being computed as follows

V_{11} = =$(0.8^2*1+0.7^2*2+0.2^2*4+0.1^2*7)/(0.8^2+0.7^2+0.2^2+0.1^2)$ = 1.568

Now calculate V_{12}, V_{21}, and V_{22}

Step 3: Now compute the distance of each point from centroid. You only need to compute **D_{11}, and D_{12}**

Step 4: Updating membership values again using the formulas above (You do not need to do anything, just for learning)

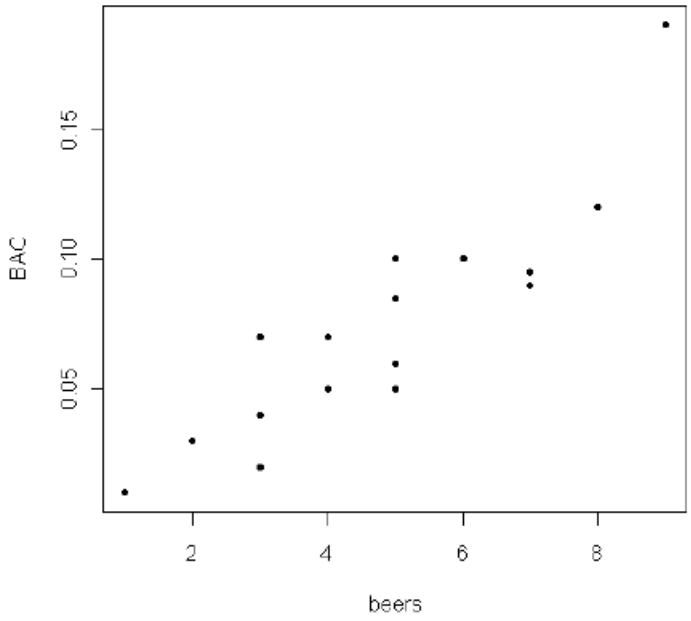
Step 5: Repeat Step2 – Step4 until there is no significant change in membership values like in Kmeans where we stop when there is no significant change in centroids (You do not need to do anything, just for learning)

So, in summary, you need to calculate V_{12}, V_{21}, V_{22}, D_{11}, D_{12}.

Question 5 [5 Points]:

In a study of alcohol consumption and related blood alcohol content, 16 student volunteers at Ohio State University drank a randomly assigned number of cans of beer. Thirty minutes later, a police officer measured their percent blood alcohol content (BAC):

# of Beers	BAC
5	0.10
2	0.03
9	0.19
8	0.12
3	0.04
7	0.095
3	0.07
5	0.06
3	0.02
5	0.05
4	0.07
6	0.10
5	0.085
7	0.09
1	0.01
4	0.05

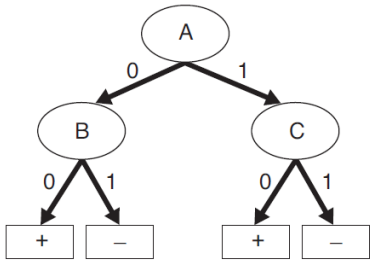



```
Coefficients:
             Estimate    Std. Error  t value     Pr(>|t|)
(Intercept) -0.012701    0.012638    -1.005       0.332
beers        0.017964    0.002402    7.480        2.97e-06

Residual standard error: 0.02044 on 14 degrees of freedom
F-statistic: 55.94 on 1 and 14 DF, p-value: 2.969e-06

Analysis of Variance Table
Response: BAC
            Df    Sum Sq     Mean Sq    F value     Pr(>F)
Beers       1     0.0233753  0.0233753  55.944      2.969e-06
Residuals   14    0.0058497  0.0004178
```

Answer the following questions

(a) What is the estimated regression line?
(b) Interpret the slope
(c) Interpret y-intercept. Is this interpretation useful or reasonable?
(d) What is the coefficient of determination? Interpret it.
(e) What is the coefficient of correlation? Interpret it.
(f) What is the estimated average BAC for a student who drinks 7 beers?
(g) What are the hypotheses for testing whether number of beers is associated with BAC?
(h) What are the hypotheses for testing whether number of beers is associated with BAC?
(i) What is the p-value associated with questions 17 and 18?
(j) What is the estimated number of beers for BAC of 0.08?

Question 6: Consider decision tree as shown below [2.5 Points]



Training:

Instance	A	B	C	Class
1	0	0	0	+
2	0	0	1	+
3	0	1	0	+
4	0	1	1	–
5	1	0	0	+
6	1	0	0	+
7	1	1	0	–
8	1	0	1	+
9	1	1	0	–
10	1	1	0	–

Validation:

Instance	A	B	C	Class
11	0	0	0	+
12	0	1	1	+
13	1	1	0	+
14	1	0	1	–
15	1	0	0	+

(a) [1.25 Points] Compute the generalization error rate of the tree using the training data.
(b) [1.25 Points] Compute the generalization error rate of the tree using the validation data.

Question 7: [8 Points]

(a) Explain why in ensemble learning it is important to obtain a diverse ensemble.
(b) What is an unstable learner, and why does Bagging rely on having an unstable learner as the base classifier?
(c) Which of the two ensemble learners, Bagging or AdaBoost, would you expect to be more robust to noise in the data? Provide a short (2-3 sentence) justification of your answer.
(d) What is Stacking?
(e) Briefly explain the difference between Homogenous and Heterogeneous Classifiers.

Question 8: [3 Points]

(a) Differentiate the concept of stemming and lemmatizing using a suitable example
(b) List 4 main reasons why Natural Language (NL) is very much more difficult to process than an Artificial Language (AL)
(c) List two real world applications of NLP

Question 9: [6 Points]

(a) [4 Points] Write down in the cells below seaborn and matlabplot functions according to their correct definitions (Some may not be used): **matplotlib.axes.Axes.hexbin, seaborn swarmplot, seaborn regplot, seaborn jointplot, matlabplot.pyplot.xlim, seaborn heatmap, matplotlib.axes.Axes.hist2d, seaborn stripplot, seaborn violin, matlabplot, opencv; [You can fill this directly in Question paper]**

For example; seaborn tsplot	**Plot one or more timeseries with flexible representation of uncertainty**
	Draw a scatterplot where one variable is categorical
	Draw a combination of boxplot and kernel density estimate.

	Plot rectangular data as a color-encoded matrix
	Draw a plot of two variables with bivariate and univariate graphs
	Plot data and a linear regression model fit
	provides the raw building blocks for Seaborn's visualizations
	Draw a categorical scatterplot with non-overlapping points
	Set the x-axis range

(b) [2 Points] Seaborn is a Python visualization library based on matplotlib. It provides a high-level interface for drawing attractive statistical graphics. One of the simplest things one can do using seaborn is to fit and visualize a simple linear regression between two variables using seaborn lmplot**.** One difference between seaborn and regular matplotlib plotting is that you can pass pandas DataFrames directly to the plot and refer to each column by name. For example, if you were to plot the column 'price' vs the column 'area' from a DataFrame df, you could call lmplot(x='area', y='price', data=df) from seaborn.

Residuals on the other hand visualizie how far datapoints diverge from the regression line. Seaborn residplot function will regress y on x. Below is the syntax of that function
seaborn.residplot(x, y, data=None, lowess=False, x_partial=None, y_partial=None, order=1, robust=False, dropna=False, label=None, color=None, scatter_kws=None, line_kws=None, ax=None)

Based on the above information and following instructions, write down the program regression.py; that will do the following

- Import matplotlib.pyplot and seaborn using the standard names plt and sns respectively.
- Plot a linear regression between the 'weight' column (on the x-axis) and the 'hp' column (on the y-axis) from the DataFrame auto.
- Generate a green residual plot of the regression between 'hp' (on the x-axis) and 'mpg' (on the y-axis). Ignore observations with missing data

Question 10 [4 Points]: Consider a document-term matrix where tf_{ij} is the frequency of the i^{th} word (term) in the j^{th} document and m is the number of documents. Consider the variable transformation that is defined by $tf_{ij}' = tf_{ij} * \log(m/df_i)$, where df_i is the number of documents in which the j^{th} term appears and is known as the document frequency of the term. This transformation is known as the inverse document frequency transformation

(a) What is the effect of this transformation if a term occurs in one document? In every document? [2 Points]
(b) What might be the purpose of this transformation? [2 Points}

BEST OF LUCK!