National University of Computer & Emerging Sciences, Karachi
Spring 2021 CS-Department
Final Exam (Solution)
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 **7 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.
- Some relevant formulas are provided in Appendix.
- In each question, show all steps clearly.

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

Question 1: Briefly answer the following questions. Each questions should be answered in maximum *20* words including articles. Otherwise, answer will not be checked. [5 Points]

a) What is the main idea behind Inverse Random Under Sampling
Majority class will become Minority class and vice versa. FPR is controlled via Bagging. (-0.5 for not discussing FPR)

b) What is the purpose of slack variable in SVM?
Ans: Slack variables are introduced to allow certain constraints to be violated. That is, certain training points will be allowed to be within the margin. We want the number of points within the margin to be as small as possible, and of course we want their penetration of the margin to be as small as possible.

c) What is the purpose of kernel in SVM
Ans: To transform from low dimension to high dimension

d) What is Hyperplane?
An hyperplane is a generalization of a plane, in one dimension, an hyperplane is called a point in two dimensions, it is a line, in three dimensions, it is a plane, in more dimensions you can call it an hyperplane

e) Why there is a need to replace least square estimation with some alternative fitting procedure
Sol:
   a. Prediction Accuracy
   b. Model Interpretability

f) How Shrinking (Regularization) can be used for feature selection
Sol:
   a. Involves shrinking the estimated coefficients toward zero relative to the OLS estimates; has the effect of reducing variance and performs variable selection.
   b. Methods: ridge regression, lasso

g) Can you think of a real-world application in which stop words would be useful for text classification?
One example: Forensic or artistic situation.

h) Explain the difference between feature selection and feature extraction

Feature selection is to find subset of features from original features that can either reduce the cost of features or reduce the cost of extracting features. Feature extraction will transform the features into new domain

i) Explain the difference between hold out, cross validation and leave one out approach?
Hold out; separate training / test data
k Fold: data is divided into k folds; each fold is reserved for testing
when k = N, is leave one out CV

j) What is the difference between classification and regression?
Classification: Discrete output
Regression: Continuous output

k) What is the role of activation functions in neural networks?
Sol:

At the most basic level, an activation function decides whether a neuron should be fired or not. It accepts the weighted sum of the inputs and bias as input to any activation function. Step function, Sigmoid, ReLU, Tanh, and Softmax are examples of activation functions.

l) What will happen when learning rate is too high in neural networks?

Sol: Network will not converge

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

Ans: False: Depends upon output coding. For 4 class problem, two output neurons are enough

n) What is gradient descent?

Iteratively check that after assigning a value how far you are from the best values, and slightly change the assigned values values to make them better

o) Why square loss function cannot be used in Logistic Regression?
Square loss will stick in local minima

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

i. [3 Points] Find frequent Itemsets using Apriori Algorithm.
ii. [2 Points] List all the strong association rules (with support s and confidence c) matching with the following metarule, where X is a variable representing customers and $item_i$ denotes variables representing items (e.g. "M", "O", etc)

$$\forall x \in transaction,\ buys(X, item_1) \wedge buys(X, item_2) \Rightarrow buys(X, item_3) \quad [s, c]$$

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}

Table 2: Database D for Transactions.

Sol:

min_support = 2

C1

K	2
A	5
R	4
C	1
H	2
I	2
L	1
O	1
E	2
Y	1
D	2
B	1
S	1
M	1
U	1

L1

K	2
A	5
R	4
I	2
E	2
D	2
H	2

C2

K, A	2
K, R	2
~~K, I~~	~~1~~
~~K, E~~	~~0~~
~~K, D~~	~~0~~
~~K, H~~	~~0~~
A, R	4
A, I	2
A, E	2
A, D	2
A, H	2
~~R, I~~	~~1~~
~~R, E~~	~~1~~
~~R, D~~	~~1~~
R, H	2
~~I, E~~	~~0~~
~~I, D~~	~~0~~
~~I, H~~	~~0~~
~~D, H~~	~~1~~
E,R	2

L2

K, A	2
K, R	2
E, R	2
A, R	4
A, I	2
A, E	2
A, D	2
A, H	2
R, H	2
E, H	2

C3

K, A, R	OK
K, A, I	Violating Apriori Principle
K, A ,H	OK
K, A, E	Violating Apriori Principle
K, A, D	Violating Apriori Principle
A, R, I	Violating Apriori Principle
A,R, E	Violating Apriori Principle
A,R,D	Violating Apriori Principle
A,I, E	Violating Apriori Principle
A,E,D	Violating Apriori Principle
A, R, H	OK
A, I, H	Violating Apriori Principle
A ,E, H	OK
A, E, R	OK
E, H, R	OK

C3

K, A, R	2
K, A ,H	2
A, R, H	2
A ,E, H	1
A, E, R	2
E, H, R	2

And so on
Finally,
Large Itemset is {A. E. H, R}

Strong Association Rules from {A, E, H, R} = 2

Some Rules
{A,E}=>{H,R} => 2 / 2 = 100%, Accepted

{R} -> {A, E, H} = 2/4 = 50% Reject

Question 3 [4 Points]

(a) 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)?

Ans (a) [27 22; 22, 27]; eigen values are 49,5. Eigenvector from 49 is {1/sqrt(2), 1 / sqrt(2)}. Reduce dimensions are 0,0,7/sqrt(2),7/sqrt(2)

Ans (b) 49,5 plus 2 zeros

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}**

Step4: 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}.

Sol:

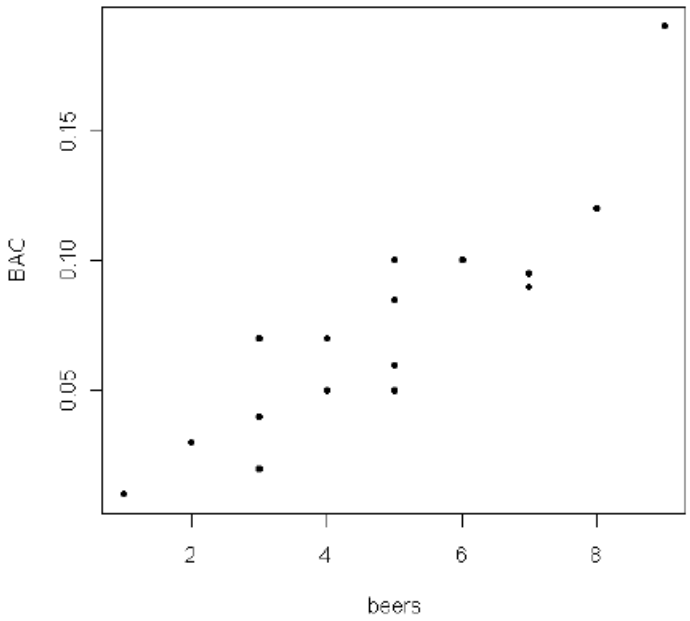
Final answers V12 = 4.051, V21 = 5.35, V22 = 8.215; 3 for all correct, 2 for calculations

D11 = 1.2, D12 = 6.79; 2 for correct approach, 1 for wrong formula

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



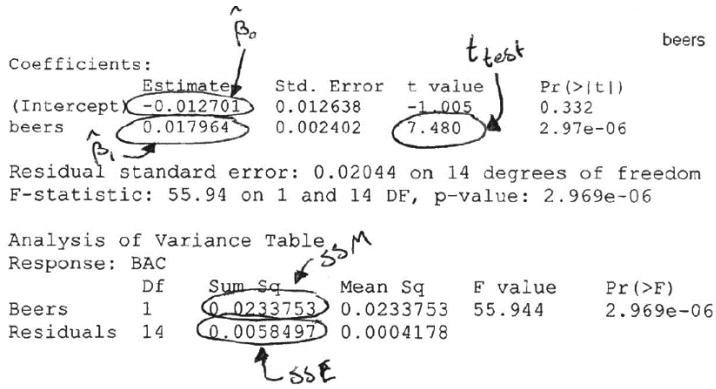

```
Coefficients:
            Estimate    Std. Error  t value     Pr(>|t|)
(Intercept) -0.012701   0.012638    -1.005      0.332
beers       0.017964    0.002402    7.480       2.97e-06

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

(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?



11. What is the estimated regression line?

$\hat{y}$ (BAC) $= -0.012701 + 0.017964x$ (beers)

12. Interpret the slope.

for every additional beer consumed, BAC increases by .017946 on average

13. Interpret y-intercept. Is this interpretation useful or reasonable?

a person who consumes 0 beers has BAC of -0.012701 on average not useful, not reasonable

14. What is the coefficient of determination? Interpret it.

$R^2 = \frac{SSM}{SST} = \frac{0.0233753}{0.0233753 + 0.0058497} = .8$

about 80% of the variability in BAC is explained by the variation in the # of beers consumed

15. What is the coefficient of correlation? Interpret it.

$r = \sqrt{R^2} = .8943$ BAC and # beers drank have a strong, positive linear relationship

16. What is the estimated average BAC for a student who drinks 7 beers?

$BAC = -0.012701 + .017964(7) = .110779$

17. What are the hypotheses for testing whether number of beers is associated with BAC?

$H_0: \beta_1 = 0$ (no association)

$H_a: \beta_1 \neq 0$ (association)

18. What is the test statistic associated with the hypotheses in question 17?

$t_{test} = 7.48$

19. What is the p-value associated with questions 17 and 18?

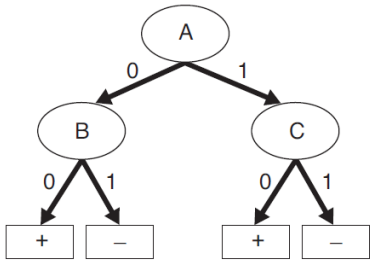
p-value = 2.97e-06 $= 2.97 \times 10^{-6}$

$= .00000297 \approx 0$

→ reject H_0, BAC & Beers are associated

Last Part: approx. 5

Question 6: Consider decision tree as shown below [3 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) [1.5 Points] Compute the generalization error rate of the tree using the training data.

(b) [1.5 Points] Compute the generalization error rate of the tree using the validation data.

Solution (a) 5/10 = 50% (b) 1/5 = 20%

Question 7: [5 Points]

(a) Explain why in ensemble learning it is important to obtain a diverse ensemble.
Different classifier can have different decisions which can led to improvement

(b) What is an unstable learner, and why does Bagging rely on having an unstable learner as the base classifier?
Ans: Not a similar output by changing few data points. Due to diversity, we need unstable learner

(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
Bagging since random distribution of points can help

(d) What is Stacking
Output of one classifier is the input of the another classifier

(e) Briefly explain the difference between Homogenous and Heterogeneous Classifiers.
Ans: Homogenous Classifiers (Same Classifiers but different training data) e.g. Bagging, Boosting etc
Heterogeneous Classifiers (Different Classifiers but same training data) e.g. Majority Voting, Mean etc)

Question 8: [3 Points]

(a) Differentiate the concept of stemming and lemmatizing using a suitable example
Sol: Word stemming means removing affixes from words and returning the root word i.e working is work. Word lemmatizing is similar to stemming, but the difference is the result of lemmatizing is a real word, actually, this is a very good level of text compression. For example stem of "increases" will be **increased**, will lemmatize will return **increase.**

(b) List 4 main reasons why Natural Language (NL) is very much more difficult to process than an Artificial Language(AL)
Sol:

NL contains a great deal of ambiguity which is controlled in AL
NL generally has more complex structure than is to be found in AL
There appears no simple universal way of representing the meaning of sentences in NL
Structure and meaning are necessarily interconnected in NL but not in AL

(c) List two real world applications of NLP

Sol: Spelling and Grammar Checking, Information Retrieval, Word Prediction

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**

For example; seaborn tsplot	Plot one or more timeseries with flexible representation of uncertainty
seaborn stripplot	Draw a scatterplot where one variable is categorical
Seaborn violin	Draw a combination of boxplot and kernel density estimate.
seaborn heatmap	Plot rectangular data as a color-encoded matrix
seaborn jointplot	Draw a plot of two variables with bivariate and univariate graphs

seaborn regplot	Plot data and a linear regression model fit
Matlabplot	provides the raw building blocks for Seaborn's visualizations
Seaborn swarmplot	Draw a categorical scatterplot with non-overlapping points
matlabplot.pyplot.xlim	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

  Sol:

```
import matplotlib.pyplot as plt
import seaborn as sns

# Plot a linear regression between 'weight' and 'hp'
sns.lmplot(x = 'weight', y = 'hp', data=auto)
sns.resigplot(x='hp', y = 'mpg', dropna = True)
# Display the plot
plt.show()
```

  Q10

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

$$tf'_{ij} = tf_{ij} * \log \frac{m}{df_i}, \tag{2.1}$$

where df_i is the number of documents in which the i^{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?

Terms that occur in every document have 0 weight, while those that occur in one document have maximum weight, i.e., $\log m$.

(b) What might be the purpose of this transformation?

This normalization reflects the observation that terms that occur in every document do not have any power to distinguish one document from another, while those that are relatively rare do.

BEST OF LUCK!