National University of Computer & Emerging Sciences, Karachi

Spring 2019 CS-Department

Final Exam (Solution)

Monday, 13th May 2019, 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. [10 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) Why it is important to scale attributes in kNN classifier?

Ans: Attributes may have to be scaled to prevent distance measures from being dominated by one of the attributes

c) What are the two main advantages of Decision Tree based Classifiers?

Ans: Inexpensive to construct

Extremely fast at classifying unknown records

Easy to interpret for small-sized trees

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

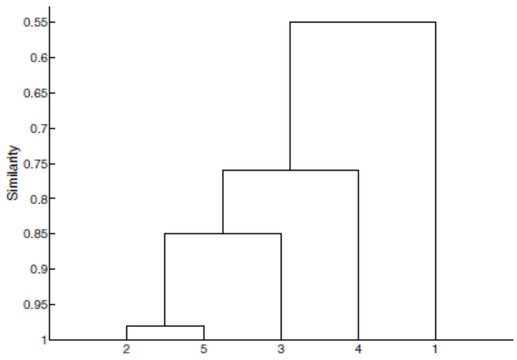
j) What is the difference between classification and regression?
Classification: Discrete output
Regression: Continuous output

Question 2 **[7 Points]**

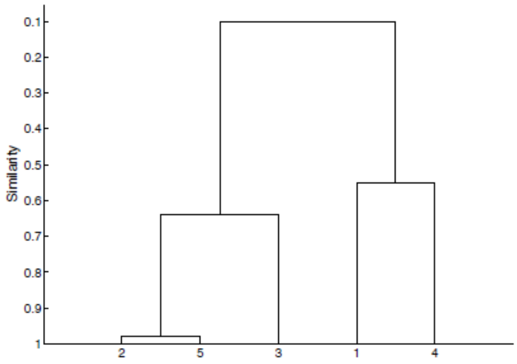
(a) **[4 Points]** Table 1 shows an example of a Distance Matrix. Draw Dendrogram using Hierarchical Clustering Process using
(a) single link clustering [2 Points]
(b) complete link clustering [2 Points]

	p1	p2	p3	p4	p5
p1	1.00	0.10	0.41	0.55	0.35
p2	0.10	1.00	0.64	0.47	0.98
p3	0.41	0.64	1.00	0.44	0.85
p4	0.55	0.47	0.44	1.00	0.76
p5	0.35	0.98	0.85	0.76	1.00

Table1: Distance Matrix

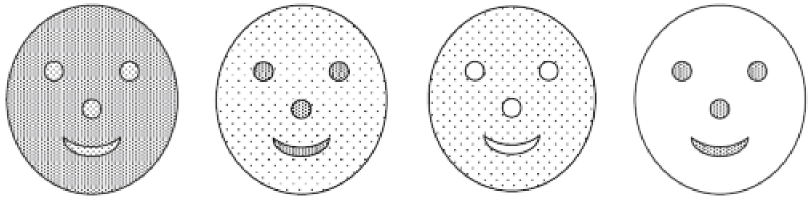


(a) Single link.



(b) Complete link.

(c) **[3 Points]** Consider the following four faces shown in the Figure below. Darkness or number of dots represents density. Lines are used only to distinguish regions and do not represent points

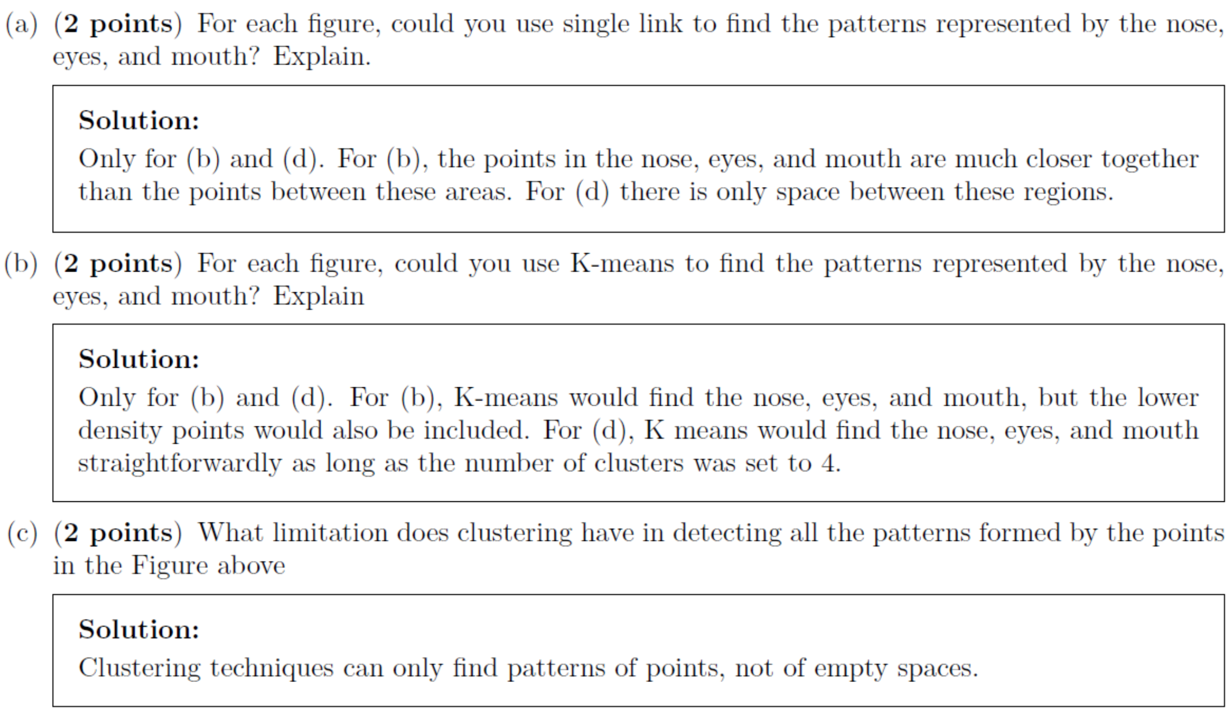


(a) (b) (c) (d)

i. For each figure, could you use single link to find the patterns represented by the nose, eyes, and mouth? Explain

ii. For each figure, could you use K-means to find the patterns represented by the nose, eyes, and mouth? Explain

iii. What limitation does clustering have in detecting all the patterns formed by the points in the Figure above



(a) (**2 points**) For each figure, could you use single link to find the patterns represented by the nose, eyes, and mouth? Explain.

> **Solution:**
> Only for (b) and (d). For (b), the points in the nose, eyes, and mouth are much closer together than the points between these areas. For (d) there is only space between these regions.

(b) (**2 points**) For each figure, could you use K-means to find the patterns represented by the nose, eyes, and mouth? Explain

> **Solution:**
> Only for (b) and (d). For (b), K-means would find the nose, eyes, and mouth, but the lower density points would also be included. For (d), K means would find the nose, eyes, and mouth straightforwardly as long as the number of clusters was set to 4.

(c) (**2 points**) What limitation does clustering have in detecting all the patterns formed by the points in the Figure above

> **Solution:**
> Clustering techniques can only find patterns of points, not of empty spaces.

Question 3: **[5 Points]**

Consider a Database D (Table 2) consists of 5 transactions. Let min_sup= 60% and min_conf = 80%.

i. [3 Points] Find frequent Itemsets using Apriori Algorithm. [Hint: Ignore items that appears twice in same transactions]

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, A, C, H, I}
2	{L, A, H, O, R, E]
3	{K, O, T, R, I}
4	{L, A, R, K, A, N, A}
5	{T, H, A, T, T, A}

Table 2: Database D for Transactions.

Sol:

Table after removing similar items

TID	Items_bought
1	{K, A, R, C, H, I}
2	{L, A, H, O, R, E]
3	{K, O, T, R, I}
4	{L, A, R, K, N,}
5	{T, H}

min_support = 3

C1

K	3
A	3
R	4
C	1
H	3
I	2
L	2
O	2
E	1
T	2
N	1

L1

K	3
A	3
R	4
H	3

C2

K, A	2
K, R	3
K, H	1
A, R	3
A, H	2
R, H	2

L2

K, R	3
A, R	3

C3

None since {K,A} is violating Aprori Principle

No associative rules according to given rules

Question 4: **[10 Points]**

The following are the value of shares of company X in Karachi Stock at the end of last three weeks

3rd Sept 2017	5.0
27th Oct 2017	4.0
20th Nov 2017	10.0
1st Dec 2017	9.0

The following three events are responsible for the change of shares of company X

Date	Event1 in Million Rupees (New Investment)	Event2 in Million Rupees (Loan Return)
3rd Sept 2017	3	4
27th Oct 2017	4	3
20th Nov 2017	2	1
1st Dec 2017	1	2

What is the predicted share value on 1st Jan 2018 (show all steps with illustration) if following events are going to happen on 1st Jan 2018 [Hint: Use PCA to reduce the dimensions to 1, then apply linear regression on reduced dimensions and share value] Formulas for linear regression are given as appendix

Event1 in Million Rupees (New Investment)	Event2 in Million Rupees (Loan Return)
2	4

First Apply PCA, square matrix, [[30 28], [28 30]]

Eigvectors

[58.   2.]

For Lambda = 1

Now, 30x + 28y = 58x => x = y

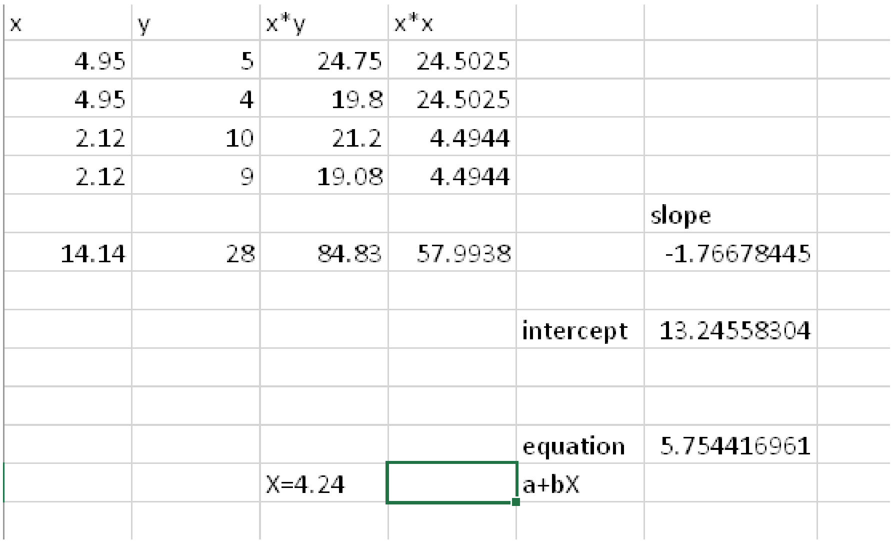
28x + 30y = 58y => x = y i.e. x = 1, y = 1

Normalize [x,y] = [1/sqrt(2), 1/sqrt(2)]

New dimensions [4.95,4.95, 2.12,2.12]

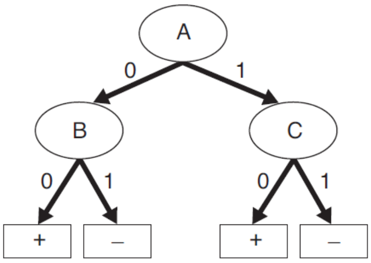
For test, [2,3] * normalize(x,y) = 4.24

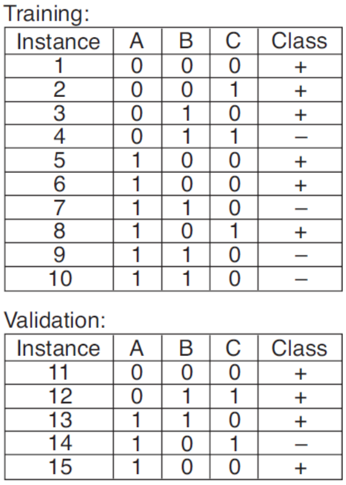
Now apply linear regression: 5.74 is the anwer



x	y	x*y	x*x		
4.95	5	24.75	24.5025		
4.95	4	19.8	24.5025		
2.12	10	21.2	4.4944		
2.12	9	19.08	4.4944		
					slope
14.14	28	84.83	57.9938		-1.76678445
				intercept	13.24558304
				equation	5.754416961
		X=4.24		a+bX	

Question 5: 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 6: **[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 7: **[10 Points]**

(a) [2 Points] What is the advantage of using seaborn library when compared with Pandas visualization library.

(b) [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

(c) [4 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)

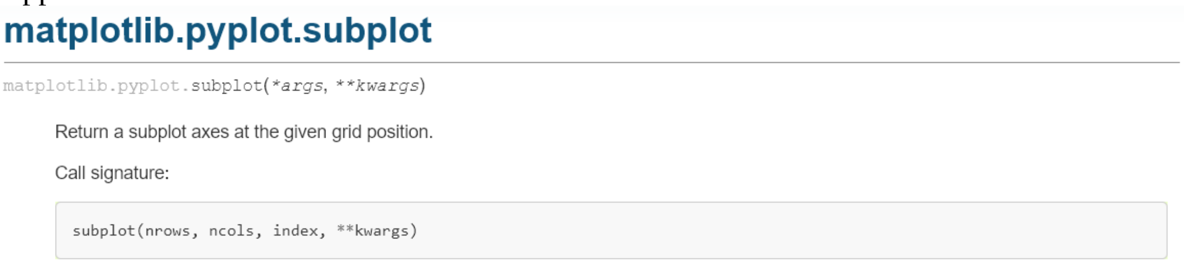
Based on the above information and following instructions, write down the program regression.py; that will do the following

- 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

```
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()
```

Appendix:



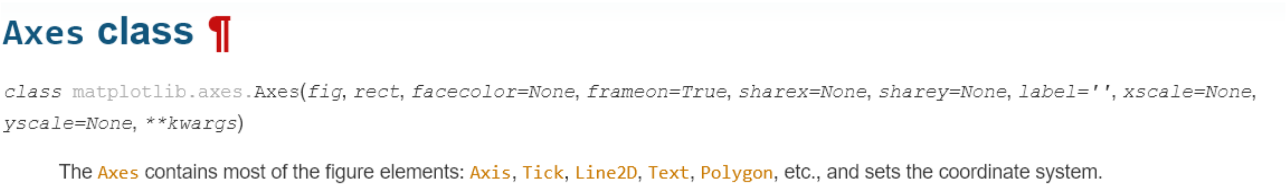
matplotlib.pyplot.subplot

matplotlib.pyplot.subplot(*args, **kwargs)

Return a subplot axes at the given grid position.

Call signature:

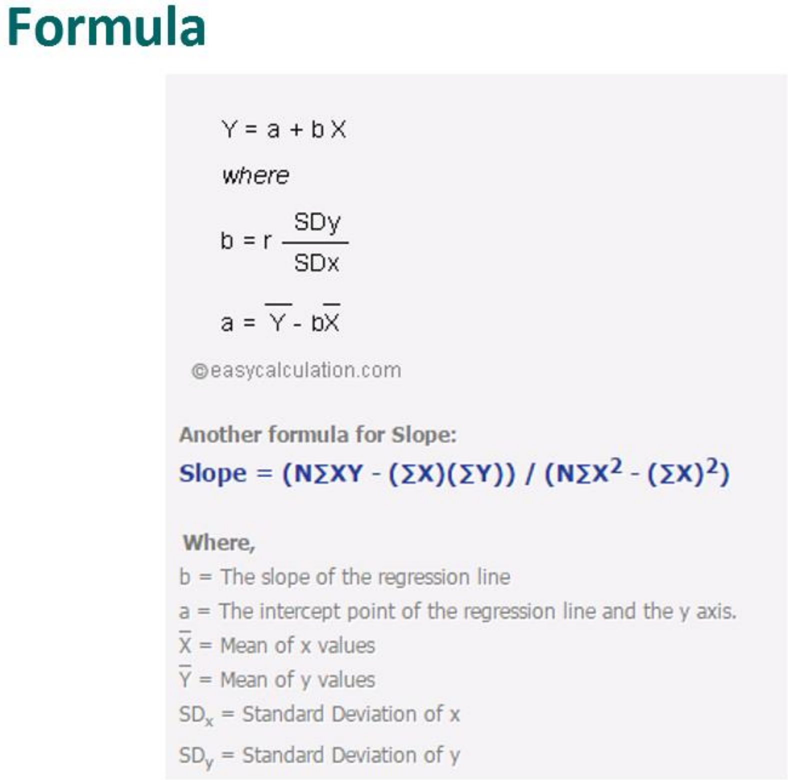
```
subplot(nrows, ncols, index, **kwargs)
```



Axes class ¶

class matplotlib.axes.Axes(fig, rect, facecolor=None, frameon=True, sharex=None, sharey=None, label='', xscale=None, yscale=None, **kwargs)

The Axes contains most of the figure elements: Axis, Tick, Line2D, Text, Polygon, etc., and sets the coordinate system.

 **Formula** $Y = a + bX$ *where* $b = r\frac{SDy}{SDx}$ $a = \overline{Y} - b\overline{X}$ ©easycalculation.com **Another formula for Slope:** **$\text{Slope} = (N\Sigma XY - (\Sigma X)(\Sigma Y)) / (N\Sigma X^2 - (\Sigma X)^2)$** **Where,** b = The slope of the regression line a = The intercept point of the regression line and the y axis. $\overline{X}$ = Mean of x values $\overline{Y}$ = Mean of y values SD_x = Standard Deviation of x SD_y = Standard Deviation of y	

BEST OF LUCK!