National University of Computer & Emerging Sciences, Karachi

Spring 2020 CS-Department

Final Exam (Solution)

22nd June 2020, 09:00 am – 12:00 pm

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

Instructions:

- Start of Exam: 9:00 am; End of Exam: 12:30 pm including submission time
- Read each question completely before answering it. There is **7 questions and 6 pages**.
- In case of any ambiguity, you may make assumptions. But your assumption should not contradict any statement in the question paper.
- You will attempt this paper **offline**, in your **hand writing**.
- You may use **cam-scanner**, **MS lens** or any equivalent application to scan and convert your hand-written answer sheets in **a single PDF file**
- The paper should be submitted using Google Form (link at the end of the paper). You are given 30 minutes for this purpose, which is already included in the exam time mentioned above. Additionally, after submitting, you should email it to your instructor which should be exactly same pdf as uploaded earlier.
- **WRITE YOUR ID ON TOP OF EVERY PAGE by your hand.** Write also **page # on every page. You should also sign on every page.**
- Please fill the below table with your details. A sample value for a male student having roll number K16-3689 and name Zeshan Khan Alvi is provided.

Sr#	Key	Description	Sample Value	Value for you
1	@fullname	Your Full Name	Zeshan Khan Alvi	
2	@fname	Your First Name	Zeshan	
3	@lname	Your Last Name	Alvi	
4	@gender	Your Gender [male,female]	Male	
5	@nameparts	Number of words/parts in your full name	3	
6	@serial	The last 4 digits of your roll number	3689 @serial[0]=3 @serial[1]=6 @serial[2]=8 @serial[3]=9	

Time: 180 minutes **Max Marks**: 50 Points

Question 1 [1 (Tokenize)+ 4(VSM)+ 2(ranking) = 7 Points]: Given paragraphs as documents and your @fullname as a query string. Tokenize the words and take first letter of each word in lower case as a token. Apply word vector space model (VSM), to compute the similarity between query vector and document vector and return the ranked list of documents for the provided query. For the ease of computation, you can use $\mathrm{similarity}(\mathrm{q},\mathrm{d}) = \sum_{\mathrm{t}=1}^{\mathrm{tokens}}(\mathrm{q_t} * \mathrm{d_t})$ for similarity between document vector d and query vector q where $\mathrm{d_t}$ is the counting of t^{th} token in document d.

Document 1) Shahzaib Yousuf Bilal Hyder Saad Umar Imtiaz Ali Issam Ahmed Neha

Document 2) Nadeem Hassan Afzal Abdullah Mujeeb Doulat Singh Murtaza Ali
Document 3) Dawar Hasnain Hamza Ashfaq Aliakber Madni Hussain Ashar Ali
Document 4) Ramchand Muhammad Ushay Murtaza Fakhruddin Ammar Rizwan
Document 5) Mujtaba Usama Vasnani Mehdi Raza Subash Kumar Shumail Steve

Note: Tokens for a document/query “Zeshan Khan Alvi” are [z, k, a].

Solution:

Tokenize [1 Point]

Doc1	s	y	b	h	s	u	i	a	i	a	n
Doc2	n	h	a	a	m	d	s	m	a		
Doc3	d	h	h	a	a	m	h	a	a		
Doc4	r	m	u	m	f	a	r				
Doc5	m	u	v	m	r	s	k	s	s		
Query	z	k	a								

Vectors [2 Points]

Doc/Tokens	a	b	h	i	n	s	y	m	n	d	r	u	f	r	v	k	z
Doc1	2	1	1	2	1	2	1					1					
Doc2	3		1			1		2	1	1							
Doc3	4		3					1		1							
Doc4	1							2			1	1	1	1			
Doc5						3		2				1		1	1	1	
Query	1															1	1

Vector Space Model (VSM) [2 Points]

$$\text{Similarity}(\text{Doc1}, \text{Query}) = 0 * 1 + 0 * 1 + 2 * 1 = 2$$
$$\text{Similarity}(\text{Doc2}, \text{Query}) = 0 * 1 + 0 * 1 + 3 * 1 = 3$$
$$\text{Similarity}(\text{Doc3}, \text{Query}) = 0 * 1 + 0 * 1 + 4 * 1 = 4$$
$$\text{Similarity}(\text{Doc4}, \text{Query}) = 0 * 1 + 0 * 1 + 1 * 1 = 1$$
$$\text{Similarity}(\text{Doc5}, \text{Query}) = 0 * 1 + 1 * 1 + 0 * 1 = 1$$

Ranking [2 Points]

$\text{Doc3}, \text{Doc2}, \text{Doc1}, \text{Doc4}, \text{Doc5}$

Question 2 [1 (Data Completion) + 4 (Training)+ 2 (Testing)= 7 Points]: Apply the naive bayes classifier on the data provided below. There are five columns (3 for X and one for the output Label and Sr# is only a serial number) of the data. The data-set provided in “Table 1” have 10 training samples and two testing samples. You are required to provide the labels for the test samples (Sr# 11 and Sr# 12) in the test data for classification.

Note: “**if(@fname[0]==vowel)**” is true if your first name starts from a vowel e.g. Owais, Ali, Imran etc.

Table 1: Training data for classification

Sr#	gender	is_even	Name starts at Vowel	Grade

		(@serial)		(SU/Letter)
1	male	Yes	yes	SU
2	male	No	if(@fname[0]==vowel)	SU
3	@gender	Yes	yes	SU
4	male	No	if(@lname[0]==vowel)	Letter
5	male	Yes	no	Letter
6	@gender	No	if(@fname[0]==vowel)	Letter
7	female	Yes	if(@lname[0]==vowel)	SU
8	male	@serial%2==0	yes	SU
9	male	Yes	no	SU
10	female	@serial%2==0	yes	SU

Table 2: Test data for Classification

Sr#	Gender	is_even (@serial)	Name starts at Vowel	Grade (SU/Letter)
11	@gender	@serial%2==0	if(@fname[0]==vowel)	?
12	@gender	@serial%2==0	if(@lname[0]==Consonant)	?

Solution

Data Completion [1 Points]

Table 2: Training data for classification

Sr#	gender	is_even (@serial)	Name starts at Vowel	Grade (SU/Letter)
1	male	Yes	yes	SU
2	male	No	no	SU
3	male	Yes	yes	SU
4	male	No	yes	Letter
5	male	Yes	no	Letter
6	male	No	no	Letter
7	female	Yes	yes	SU
8	male	No	yes	SU
9	male	Yes	no	SU
10	female	No	yes	SU

Table 2: Test data for Classification

Sr#	Gender	is_even (@serial)	Name starts at Vowel	Grade (SU/Letter)
11	male	No	no	?
12	male	No	no	?

Training [4 Points]

$\text{Gender}(\text{male}, \text{SU}) = 5/7$

$\mathrm{Gender(female, SU)} = 2/7$

$\mathrm{Gender(male, Letter)} = 3/3$

$\mathrm{Gender(female, Letter)} = 0/3$

$\mathrm{Is\_even(Yes, SU)} = 4/7$

$\mathrm{Is\_even(No, SU)} = 3/7$

$\mathrm{Is\_even(Yes, Letter)} = 1/3$

$\mathrm{Is\_even(No, Letter)} = 2/3$

$\mathrm{Vowel(Yes, SU)} = 5/7$

$\mathrm{Vowel(No, SU)} = 2/7$

$\mathrm{Vowel(Yes, Letter)} = 1/3$

$\mathrm{Vowel(No, Letter)} = 2/3$

Testing [2 Points]

$\mathrm{P(male, no, no: SU)} = \mathrm{Gender(male, SU)} * \mathrm{Is\_even(No, SU)} * \mathrm{Vowel(No, SU)}$

$\mathrm{P(male, no, no: SU)} = 5/7 * 3/7 \; * 2/7 * 7/10 = 210/3430 = 0.061$

$\mathrm{P(male, no, no: Letter)} = 3/3 * 2/3 \; * 2/3 * 3/10 = 36/270 = 0.13$

Letter Grade for both

Question 3 [6 Points]: Using hierarchical clustering algorithms (Single link and Distance b/w centroids) and City-block distance ($\mathrm{d} = (|\mathrm{x}_2 - \mathrm{x}_1|) + (|\mathrm{y}_2 - \mathrm{y}_1|)$) to cluster the following 5 points into 3 clusters. Using A1 = (@serial[1],10), A2 = (2,@serial[0]), A3 = (8,4), A4 = (5,@serial[2]), A5 = (7,5).

Solution:

Table

Point	X	Y
A1	5	10
A2	2	3
A3	8	4
A4	5	0
A5	7	5

Centroid [1.5 Points]

Point	X	y	Distance				
A1	5	10	0	10	9	10	7
A2	2	3	10	10	9	10	7
A3	8	4	9	7	0	7	2
A4	5	0	10	6	7	0	7
A5	7	5	7	7	2	7	0

The minimum distance between Points A3 and A5.

Combine point A3 and point A5

Point	X	Y	Distance			
A1	5	10	0	10	8	10
A2	2	3	10	0	7	6
A3,A5	7.5	4.5	8	7	0	7
A4	5	0	10	6	7	0

The minimum distance between Points A3 and A5.

Combine point A2 and cluster A3,A5

Point	X	Y	Distance		
A1	5	10	0	6.5	5
A2,A3,A5	4.75	3.75	6.5	0	4
A4	5	0	5	4	0

Three clusters are:

Cluster 1: Point A1

Cluster 2: Points A2,A3,A5

Cluster 3: Point A4

Single Link [1.5]

Points	X	Y	Distance				
A1	5	10	0	10	9	10	7

A2	2	3	10	0	7	6	7
A3	8	4	9	7	0	7	2
A4	5	0	10	6	7	0	7
A5	7	5	7	7	2	7	0

The minimum distance is in between points A3 and A5

Point	X	Y	Distance			
A1	5	10	0	10	7	10
A2	2	3	10	0	6	6
A3,A5	8,7	4,5	7	6	0	7
A4	5	0	10	6	7	0

The minimum distance is between points A2 and A3,A4

Three clusters are:

Cluster 1: Point A1

Cluster 2: Points A2,A3,A5

Cluster 3: Point A4

Question 4 [1.5+1.5+1.5+1.5=6 Points]: Consider the data set shown in Table below:

w = ceil(@serial[0]/2)
x = ceil(@serial[1]/2)
y = ceil(@serial[2]/2)
z = ceil(@serial[3]/2)

Let x = 4, y = 2, z = 5, w = 5

Customer Number	Items Bought
A	{1, **x**, 5}
A	{1,2,3,5}
B	{1,2,4,5}
B	{1,3,4,5}
C	{2,3,**w**}
C	{2,4,5}
D	{3,4}
D	{1,**y**,3}
E	{1,4,5}
E	{1,2,**z**}

(a) Compute the support for itemsets *{5}*, *{2, 4}*, and *{2, 4, 5}* by treating each transaction ID as automobile shop basket.

Sol: Support ((5}) = 8 / 10 = 0.8

Support ({2,4}) = 2 / 10 = 0.2

Support ({2,4,5}) = 2 / 10 = 0.2

(b) Computer the confidence for the association rule (i) {2,4} -> {5} (ii) {5} -> {2,4}. Is the confidence a symmetric measure?
confidence(2,4 -> 5) = 0.2 / 0.2 = 1
confidence(5->2,4) = 0.2 / 0.8 = 0.25

No, it is not symmetric measure

(c) Repeat (a) by treating each customer number as automobile parts basket. Similar customer numbers should be treated as one customer.
Support(5) = 4/5

Support(2,4) = 5/5 = 1

Support(2,4,5) = 4/5 = 0.8

(d) What is the maximum number of association rules that can be extracted from this data?

3^d – 2 ^{d+1} +1 = 3^5 – 2^6+1 = 180 rules

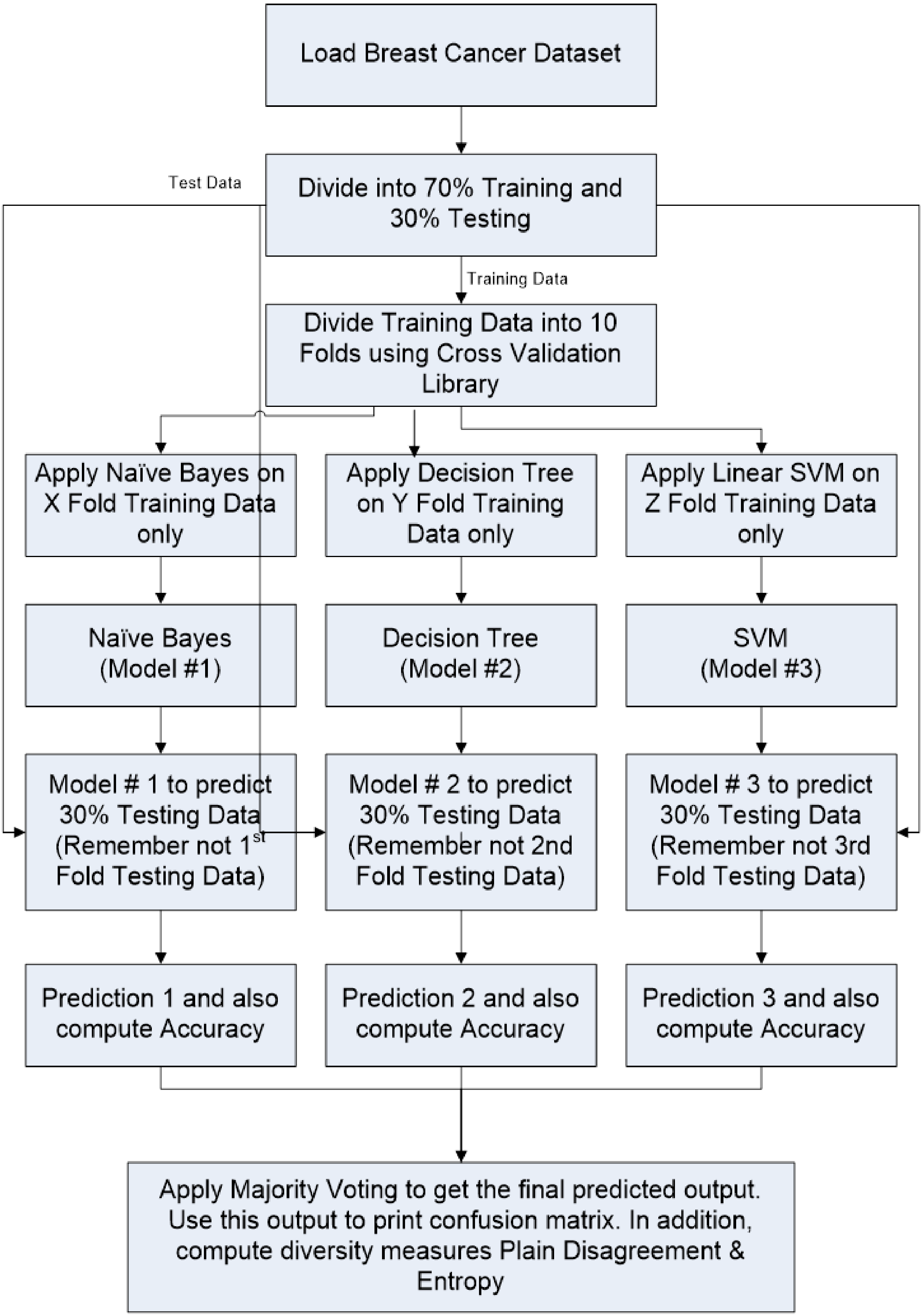
For d=6

3^d – 2 ^{d+1} +1 = 3^5 – 2^6+1 = 602 rules

Question 5 [10 Points]: Implement the model shown in Figure below. Upload the source code only. Here X = @serial[1]+1, Y = @serial[2]+1, Z = @serial[3]+1

Consider the following Initial Coding:

```
# Load libraries
…………………… // import necessary libraries
from sklearn.datasets import load_breast_cancer
breast_data = load_breast_cancer()
#…… Complete the program as explained in block diagram
```


Load Breast Cancer Dataset
Divide into 70% Training and 30% Testing
Test Data
Training Data
Divide Training Data into 10 Folds using Cross Validation Library
Apply Naïve Bayes on X Fold Training Data only
Apply Decision Tree on Y Fold Training Data only
Apply Linear SVM on Z Fold Training Data only
Naïve Bayes (Model #1)
Decision Tree (Model #2)
SVM (Model #3)
Model # 1 to predict 30% Testing Data (Remember not 1st Fold Testing Data)
Model # 2 to predict 30% Testing Data (Remember not 2nd Fold Testing Data)
Model # 3 to predict 30% Testing Data (Remember not 3rd Fold Testing Data)
Prediction 1 and also compute Accuracy
Prediction 2 and also compute Accuracy
Prediction 3 and also compute Accuracy
Apply Majority Voting to get the final predicted output. Use this output to print confusion matrix. In addition, compute diversity measures Plain Disagreement & Entropy

```
breast_data = load_breast_cancer()

# Divide iris_data into training and testing (70% training, and 30% testing)

# assign all data to variable X
# assign target class to variable Y
```

[1 Points to assign data into X and Y]

```
X = breast_data.data
Y = breast_data.target
```

[1 Points to divide into 70 / 30 split]

```
X_train, X_test, y_train, y_test = train_test_split(X, Y, test_size=0.3, random_state = 1)
```

[0.5 Points; Print Size of Training and Testing Data]

```
print("Size of Training Data")
print(X_train.shape)

print("Size of Testing Data")
print(X_test.shape)
```

[0.5 Points; Divide X_train into 3 Folds]

```
# Divide X_Train into 3 Folds
kf = KFold(n_splits=3) # Define the split - into 3 folds

i = 1;

for train_index, test_index in kf.split(X_train):
```

[1 Point for this for loop]

[0.5 Point to correctly assign train label]

```
#print(“TRAIN:”, train_index, “TEST:”, test_index)
train, test = X_train[train_index], X_train[test_index]
label_train, label_test = y_train[train_index], y_train[test_index]
```

[1 Points to correctly Run Naïve Bayes]

```
if i ==1:
#Run naive bayes
nb = GaussianNB()
nb.fit(train, label_train)
prediction1 = nb.predict(X_test)
print("Fold" + str(i) + ": Accuracy using Naive Bayes: " + str(accuracy_score(y_test, prediction1)))
```

[0.75 Points to correctly Run Decision Tree]

```
elif i==2:
dt = tree.DecisionTreeClassifier()
dt.fit(train, label_train)
prediction2 = dt.predict(X_test)
print("Fold" + str(i) + ": Accuracy using Decstion Tree " + str(accuracy_score(y_test, prediction2)))
```

[0.75 Points to correctly Run SVM with correct parameters]

```
elif i==3:
```

```
svm = svm.SVC(kernel='linear', C=1)
svm.fit(train, label_train)
prediction3 = svm.predict(X_test)
print("Fold" + str(i) + ": Accuracy using SVM: " + str(accuracy_score(y_test, prediction3)))
i = i+1;
```

[2 Points for Majority Voting implementation and 2 points for Diversity measure

```
# Calculate Majority Voting from prediction1, prediction2, prediction3, prediction4
count1 = 0
count2 = 0

final_prediction = prediction1 # dummy assign to prediction

#print(len(prediction1))

for i in range(len(prediction1)):
if prediction1[i] == 1:
count1 = count1+1;
else:
count2 = count2+1;
if prediction2[i] == 1:
count1 = count1+1;
else:
count2 = count2+1;
if prediction3[i] == 1:
count1 = count1+1;
else:
count2 = count2+1;

#print(count1)
#print(count2)
if count1 > count2:
final_prediction[i] = 1;
else:
final_prediction[i] = 0;

# reset bout count1 and count2

count1 = 0;
count2 = 0;


# Now caculate Accuracy using Ensemble

print("Accuracy using Majority Voting: " + str(accuracy_score(y_test, final_prediction)))

/* For diversity measure, check the logic only */
```

Question 6 [8 Marks]: The following table shows the value of shares of company in Karachi Stock at the end of last four weeks:

Date	**Share Value (Target Variable)**
3rd Sept 2017	@serial[0]
27th Oct 2017	@serial[1]
20th Nov 2017	@serial[2]
1st Dec 2017	@serial[3]

The following two events in Table below are responsible for the change of shares of company

Table: Events

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 of 2 events to 1, then apply linear regression on 1st dimension as independent variable and share value as target variable]

Event1 in Million Rupees (New Investment)	Event2 in Million Rupees (Loan Return)
@serial[1]	@serial[2]

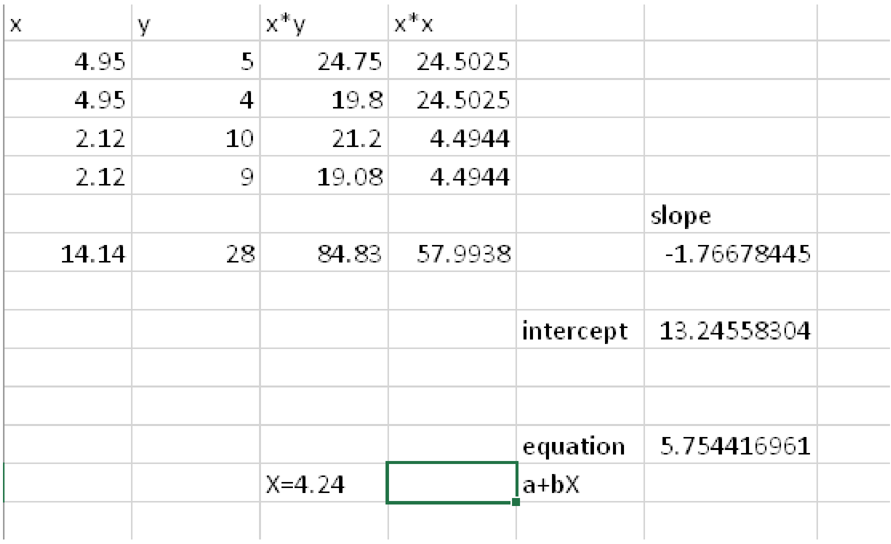
Sol: PCA: 4 Points and LR: 4 Points
First Apply PCA, square matrix, [[30 28], [28 30]]
Eigvectors
[58.   2.]
For Lambda = 1
Now, 30x + 28y = 58x => x = y
28x + 30y = 58y => x = y i.e. x = 1, y = 1

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

New dimensions [4.95,4.95, 2.12,2.12] [3 Points upto here]

For test, [2,3] * normalize(x,y) = 4.24 [1 Point]

Now apply linear regression: 5.74 is the answer



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 7 [6 Points]:

(a) **[4 Points]** In this Problem, you will work on the **Error Bars** to display error visually in a bar chart. For someone who is learning about the different drink types at Macdonald, a bar chart of milk amounts in each drink may be useful. We have provided the ounces_of_milk list, which contains the amount of milk in each 14oz drink in the drinks list. According to different barista styles and measurement errors, there might be variation on how much milk actually goes into each drink. We have included a list error on each amount of milk. You need to write program in your answer sheet.

drinks = ["cappuccino", "latte", "chai", "americano", "mocha", "espresso"]
ounces_of_milk = [w, x, y, z, 9, 10]
error = [1.1, 0.7, 1.2, 1.0, 0, 1.7]
where, w = @serial[0], x = @serial[1], y = @serial[2], z = @serial[3]

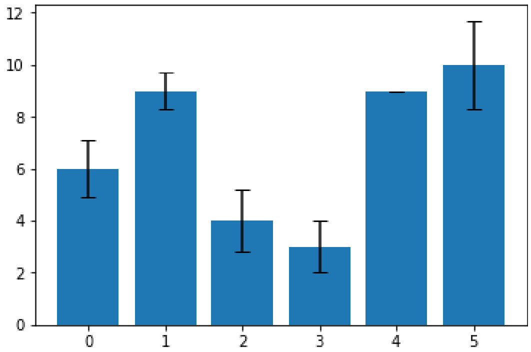
i. Plot this information as a bar chart. [1 point]

ii. Display this error as error bars on the bar graph and add caps of size 5 to your error bars. [1 point]

iii. Set the axis to go from ‘cappuccino’ to ‘espresso’ on the x-axis and 2 to 14 on the y-axis. [1 point]

iv. Add the title "Drinks to milk ratio", x-axis label "Drinks", and y-axis label "Milk amount in ounces. [1 point]

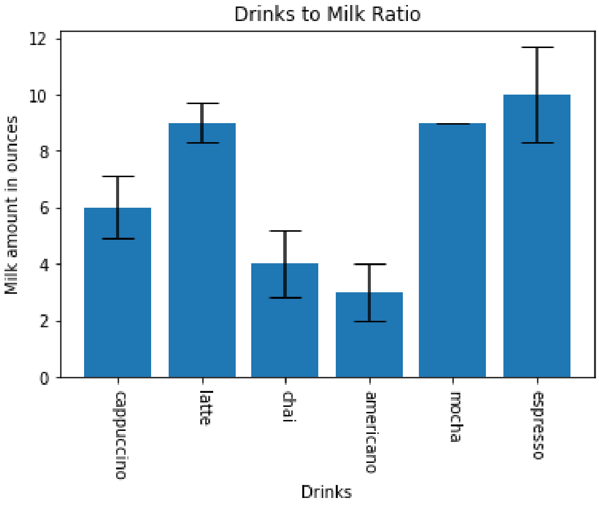
```
from matplotlib import pyplot as plt
drinks = ["cappuccino", "latte", "chai", "americano", "mocha", "espresso"]
ounces_of_milk = [6, 9, 4, 3, 9, 10]
error = [1.1, 0.7, 1.2, 1.0, 0, 1.7]
plt.bar(range(len(drinks)),ounces_of_milk , yerr=error, capsize=5)
plt.show()
```



v.

Q 4 b (iii and iv)

```
from matplotlib import pyplot as plt
drinks =["cappuccino","latte","chai","americano","mocha","espresso"]
ounces_of_milk = [6, 9, 4, 3, 9, 10]
error = [1.1, 0.7, 1.2, 1.0, 0, 1.7]
ax = plt.subplot()
```



```
plt.bar(drinks,ounces_of_milk,yerr=error,capsize=10)
ax.set_xticklabels(drinks, rotation=270)
plt.title("Drinks to Milk Ratio")
ax.set_xlabel("Drinks")
ax.set_ylabel("Milk amount in ounces")
plt.show()
```

(b) **[2 Points]** Write in your own words difference between seaborn heatmap, seaborn stripplot, seaborn violin, and seaborn heatmap.

seaborn stripplot [0.75]	Draw a scatterplot where one variable is categorical
Seaborn violin [0.75]	Draw a combination of boxplot and kernel density estimate.
seaborn heatmap [0.5]	Plot rectangular data as a color-encoded matrix

Concluding Remarks

You need to prepare a pdf file of all the question as per the question ordering. The orientation should be portrait for each page. It should be clearly visible for each and every text written on the page. You suppose to upload it on the provided form as an assignment submission. You have good 30 minutes for it.
Form URL
https://docs.google.com/forms/d/e/1FAIpQLSefZ3vTJHuudiEDlSu3ok7t1kHXEDlkEkdTUHNxIbVYtby1gw/viewform?usp=sf_link

BEST OF LUCK!