Set 1

Name: Roll number:

Partial/Step marking is not there in most of the questions. Please answer accordingly. Internet in your laptop must be turned off. Make reasonable assumptions whenever necessary.

Don't do any rough work in the answer sheet. Write only the final answers. Make sure your handwriting is legible.

The expected value of indulging in unfair means during the examination is negative. Please refrain from indulging in unfair means.

Write your name and roll number on top of each sheet of the paper.

Unless explicitly stated in any question, don't do scaling of the data.

Total marks :

Time : 2 hours 15 minutes

Q.1

Consider the following data and answer the following questions.

X1	X2	Y
0	0	1
0	40	0
40	0	0
40	40	1

X1 and X2 are the input features, Y is the output feature.

Q1a.) [5 marks] –

The above situation can be represented by a neural network with no hidden layers.

a.) True

[✓ b.) False]

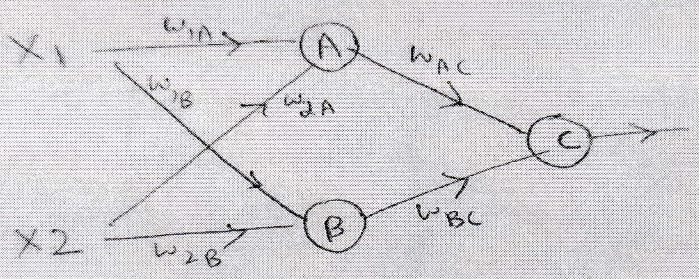
Q1b.) [15 marks]

Consider the following Artificial Neural Network that attempts to perfectly represent the above situation. Assume the activation function to be threshold function at each node,

i.e. f(x) = 1 , x>=0; 0 otherwise.

Fill in the blanks. [You get full credit only if all the 9 blanks are filled correctly, otherwise the score is zero]

Name: Roll number:



For the bias term, use the notations as w_A, w_B and w_C

W_A : ________, W_B : _________, W_C : ________,

W_{1A} : _________, W_{1B} : _________, W_{AC} : _________,

W_{2A} : _________, W_{2B} : _________, W_{BC} : _________

Q.2 – (5 marks)

If you have 600 documents and 800 unique terms, how many elements are there in the "Document Term Matrix" ? [We include all the terms in building DTM and use only unigrams]

480,000

Q.3 – (15 marks)

Consider the following 3 sentences.

S1 – Data Mining course is a easy course

S2 – It is difficult to drop the course

S3 – It is difficult to ignore the course

From these 3 sentences, make a TF-IDF matrix.

remove only the following stopwords – {is, a, It, in, to, the}

Formula used to obtain IDF from DF is (1+ln((1+n)/(1+df(t)).

Write only the TF-IDF matrix, don't write any rough work here.

ame: Roll number:

	course	data	difficult	drop	easy	ignore	mining
0	0.5634	0.4769	0	0	0.4769	0	0.4769
1	0.4254	0	0.5478	0.7203	0	0	0
2	0.4254	0	0.5478	0	0	0.7203	0

Q.4 Consider the following output from a machine learning model used to classify an article into sports, movie or politics.

S.No	1	2	3	4	5	6	7	8	9
Actual Class	Sports	Movie	Politics	Sports	Movie	Politics	Sports	Movie	Politics
Predicted Class	Movie	Movie	Politics	Sports	Politics	Sports	Sports	Politics	Politics

4a.) [5 marks]

What is the accuracy of the model?

55.5%

4b.) [10 marks]

Write the confusion matrix for the above classification model.

True

$$\begin{array}{c|ccc} & S & M & P \\ \hline S & 2 & 1 & 0 \\ M & 0 & 1 & 2 \\ P & 1 & 0 & 2 \end{array}$$

Predicted

4c.) [10 marks]

What is the F-1 score for class – Movie ?

0.40

Name:

Roll number:

Q.5 [15 marks]

The output of a Gaussian Mixture Model for classification of a one-dimensional data in three clusters is as follows:
Pi_1, mu_1, sigma_1 = (0.5,100,20)
Pi_2, mu_2, sigma_2 = (0.2,50,20)
Pi_3, mu_3, sigma_3 = (0.3,150,20)

The symbols have their usual meaning. Pi_i is the prior probability a point belongs to sample i, mu_i and sigma_i is the mean and standard deviation respectively of the ith normal distribution.

Now, we want to cluster a point {125} using the model. What is the probability that the point belongs to cluster 3?

0.3748

Q.6 [10 marks]

Use the k-means algorithm to cluster the following 1-dimensional data into 3 clusters.
[1,3,5,7,8,10,12,13,15,16,18]
Take the initial center of the 3 clusters to be 4,8,12.
If there is a tie in deciding which cluster should a point go to, let the point go in the cluster with the least value of the center.

4 – 1,3,5
8 – 7,8,10

the points in cluster 1 are – 12 – 12,13,15,16,18 (1,3,5)

3 – 1,3,5
8.33 – 7,8,10

the points in cluster 2 are – 14.8 – 12,13,15,16,18 (7,8,10)

the points in cluster 3 are – 3 8.33 14.8 (12,13,15,16,18)

Q.7[10 marks]

The eigenvalues of a 2*2 matrix (M) is 60 and 40, the respective eigenvectors are [1,1] and [1,-1]. What is M? If there are multiple possibilities for M, write any one of them.

50	10
10	50

Name: Roll number:

Q.8 [5 marks]

Consider the following output from the VADER analysis of a sentence. The number corresponding to "pos" is missing. What should be that number?

{'neg': 0.231, 'neu': 0.396, 'pos': 0·373, 'compound': 0.3182}

Q.9 (5 marks)
For data from social media having emojis etc, which of the following options is better for conducting sentiment analysis.

a.) AFINN

[✓ b.) VADER]

Q.10 (5 marks)
After India's loss to England in semifinals of T-20 world cup, we extract the data from Twitter in the form of 1 lakh recent tweets on this issue. This is an example of –

a.) Structured Data

b.) Unstructured Data

Q.11 (5 marks) If the end goal of a text analysis is viewing the "word cloud", which of the following steps must be avoided –

a.) Converting to lower case

b.) Removing custom stopwords

[✓ c.) Stemming]

d.) Lemmatization

Q.12 [10 marks]

Refer to the file by the name "startup.csv" sent to you. It contains the data of 50 US startups about their R&D spend, marketing spend and advertising spend. Make 3 clusters using Gaussian Mixture Model.

Let a, b, c be the number of elements in the three clusters in no particular order.

a: 15, b: 30, c: 5 [different ans. for different roll no]

{For the GMM model, use an additional argument as random_state = your roll number.
Eg - model = GMM(n_components=6, random_state = 20022212)}

↳ 21918016

Name:

Roll number:

Q.13 (10 marks + 10 marks + 10 marks + 10 marks)
Refer to the advertisement data sent to you and answer the following questions accordingly.

There are 1000 records in the data.

13a.) Build a logistic regression model using the first "900-last two digits of your roll number" in the training set.

Eg. if your roll number 20201863, use the first 837 records to build the model. [Hint: You may use the argument shuffle = False in train_test_split]

What is the accuracy of the model in the records used for training.

0.955 [different ans. for different roll no.]

Accuracy = (records correctly classified by the model assuming 0.5 as the threshold)/(total no. of records used to build the model)

[Answer should be correct upto 4 decimal places, you can choose whether to round off the final result or not, both the answers will be accepted]

13b.) For the above question, take the remaining records, i.e, the records not used in building the model as the test set.

In the test set, what is the accuracy of the model?

0.9310 [different ans. for different roll no.]

[Answer should be correct upto 4 decimal places, you can choose whether to round off the final result or not, both the answers will be accepted]

13c.) Build a "**SVM model using kernel as polynomial of degree 2**" model using all the 1000 records.

What is the accuracy of the model?

Accuracy = (records correctly classified by the model)/(total no. of records used to build the model = 1000)

0.956

[Answer should be correct upto 4 decimal places, you can choose whether to round off the final result or not, both the answers will be accepted]

Name: Roll number:

13d.) Ignore the last column of the original dataset now, and using the remaining columns, cluster the data using k-means clustering.

What is the optimal number of clusters using DB-Index? Optimal k is defined as : the value of k having the least value and k<=10

2

Q.14 [5 marks]
Logistic Regression (for a 2 class problem) can do perfect classification with 100% accuracy only if the points are linearly separable.

[✓ a.) True]
b.) False

Q.15 [5 marks]
Consider a dataset having 3 classes and 10 features. We apply PCA to reduce it to 3 features.
If a dataset is linearly separable in higher dimension, Principle Component Analysis(PCA) will ensure that the data is separable in lower dimensions.

a.) True
[✓ b.) False]

Name:

Roll number:

Q.16 [5+10+10 marks]

Consider the following dataset having 4 features

S.No	X1	X2	X3	X4
1	10	1	10	40
2	11	5	50	40
3	12	9	90	40
4	13	13	130	40
5	14	17	165	50
6	15	21	200	60
7	16	25	250	70

Use PCA on the dataset to convert it into a dataset having 2 features.

16a.) What is the largest eigenvalue?

7345.13

16b.) What is the percentage of variance explained by the top 2 features?

99.58, 0.41

16c.) Fill the table below with the corresponding entries for the 2 features

S.No	Feature 1	Feature 2
1	-118.63	-6.49
2	-78.74	-1.4
3	-38.85	3.6
4	1.02	8.77
5	37.25	3.31
6	73.47	-2.15
7	124.49	-5.72