Birla Institute of Technology & Science, Pilani
Work-Integrated Learning Programmes Division
MTech. Software Engineering at DSE (FC04, FA04_1-2021) Cluster

Second Semester 2021-2022
Endsem -Semester Test
(EC-3 Makeup)

Course No.	: DSECLZG565
Course Title	: Machine Learning
Nature of Exam	: Open Book
Weightage	: 40%
Duration	: 2 Hours
Date of Exam	:

No. of Pages = 3
No. of Questions = 8

Note:

1. Please follow all the *Instructions to Candidates* given on the cover page of the answer book.
2. All parts of a question should be answered consecutively. Each answer should start from a fresh page.
3. Assumptions made if any, should be stated clearly at the beginning of your answer.

Q.1. Consider the following training set in 2-dimensional Euclidean space. [5 Marks}

Point	Coordinate	Class
X1	(-1, 1)	Negative
X2	(0, 1)	Positive
X3	(0, 2)	Negative
X4	(1, -1)	Negative
X5	(1, 0)	Positive
X6	(1, 2)	Positive
X7	(2, 2)	Negative
X8	(2, 3)	Positive

What is the class of the point (1, 1) if 7NN classifier is considered? If the value of K is reduced whether the class will change? (Consider K=3 and K=5). What should be the final class if the above 3 values of K are considered?

Solution: Euclidian distance following table – 2M

Point	Coordinate	Class	Distance from 1,1
X1	(-1, 1)	Negative	2
X2	(0, 1)	Positive	1
X3	(0, 2)	Negative	1.414
X4	(1, -1)	Negative	2
X5	(1, 0)	Positive	1
X6	(1, 2)	Positive	1
X7	(2, 2)	Negative	1.41

X8	(2, 3)	Positive	2.236

i. class of the point (1, 1) if 3NN classifier is considered x2, x5, x6 - Positive

ii. class of the point (1, 1) if 5NN classifier is considered?
x2, x5, x6, x3, x7 - Positive

iii. class of the point (1, 1) if 7NN classifier is considered?
x2, x5, x6, x3, x7, x1, x5- Negative

Final class value to be considered as Positive

Q.2 Use kernel trick and find the equation for hyperplane using nonlinear SVM. Positive Points: {(7,0), (9,0), (11,0)} Negative Points: {(0,0), (8,0), (12,0), (10,0)}. Plot the point before and after the transformation. [5 Marks]

Solution:

Φ(x) = x mod 2 [3M]
Equation of hyperplane : y=0.5 [2M]

Q.3 Consider the following dataset. Fit a regression model (lines passing through the origin) used for line fitting for the given data, [4+3=7 Marks]

x	y
1	3
1	2
2	1
3	3
5	3
7	7
6	4
7	6
6	7
4	5

A. Compute the squared error E(w) for varying values of w ={-1.0, -0.2, -0.4, 0, 0.2, 0.4, 1.0}. Mark the point of global minima on the plotted error curve.

B. Define the best hypothesis and predict the value of x=7.6. Show the computation steps clearly.

Solution:

A. $$E(w) = \sum_i \quad (y_i - wx_i)^2$$
$$= \sum_i \quad y_i^2 - 2w \sum_i \quad x_i y_i + w^2 \sum_i \quad x_i^2$$
$$= \sum_i \quad y_i^2 - 2w \sum_i \quad x_i y_i + \sum_i \quad x_i^2 w^2$$
$$= 207 - 416\, w + 226\, w^2$$

At Global minima: $d\frac{E(w)}{dt} = 0,$

W*	e(w)
0.92	15.5664

E(w) is a vertically oriented parabola with vertex (0.92, 15.56)

w	E(w)
-1	849
-0.2	299.24
-0.4	409.56
0	207
0.2	132.84
0.4	76.76
1	17

B. Best hypothesis, $w^* = 0.92$ and h(x=7.6) = 0.92 * 7.6 =6.992=7

Q.4 Consider the dataset given below where A and B are attributes which can take the values 0 and 1, and Y is the classification. The values marked "*" represent data values that are corrupted. It is known that during the construction of a decision tree to represent the clean dataset (i.e one without any "*"), the attribute B was chosen at the root instead of attribute A using information gain. Is this information enough to guess the value of the bit that must replace "*"? Give a detailed justification for your answer. [7 Marks]

A	B	Y
1	0	no
1	1	no
0	*	no
0	1	yes
0	1	yes
1	1	yes

Answer

(2M)

Let S be the given dataset. We have

$InfGain(A) = Entropy(S) - \frac{|S_{A=0}|}{|S|} Entropy(S_{A=0}) - \frac{|S_{A=1}|}{|S|} Entropy(S_{A=1}) = -(\frac{3}{6}\log_2\frac{3}{6} + \frac{3}{6}\log_2\frac{3}{6}) - \frac{3}{6}(-\frac{1}{3}\log_2\frac{1}{3} - \frac{2}{3}\log_2\frac{2}{3}) - \frac{3}{6}(-\frac{1}{3}\log_2\frac{1}{3} - \frac{2}{3}\log_2\frac{2}{3}) = -(\frac{3}{6}\log_2\frac{3}{6} + \frac{3}{6}\log_2\frac{3}{6}) + \frac{2}{3}\log_2\frac{2}{3} + \frac{1}{3}\log_2\frac{1}{3} = 1 - 0.91 = 0.09.$

(2M)

If we assume *=1, then we $InfGain(B,*=1) = Entropy(S) - \frac{|S_{B=0}|}{|S|} Entropy(S_{B=0}) - \frac{|S_{B=1}|}{|S|} Entropy(S_{B=1}) = -(\frac{3}{6}\log_2\frac{3}{6} + \frac{3}{6}\log_2\frac{3}{6}) - \frac{1}{6}(-1\log_2 1 - 0\log_2 0) - \frac{5}{6}(-\frac{3}{5}\log_2\frac{3}{5} - \frac{2}{5}\log_2\frac{2}{5}) = 1 - 0.809 = 0.191$

(2M)

If we assume *=0, then we have $InfGain(B,*=0) = Entropy(S) - \frac{|S_{B=0}|}{|S|} Entropy(S_{B=0}) - \frac{|S_{B=1}|}{|S|} Entropy(S_{B=1}) = -(\frac{3}{6}\log_2\frac{3}{6} + \frac{3}{6}\log_2\frac{3}{6}) - \frac{2}{6}(-1\log_2 1 - 0\log_2 0) - \frac{4}{6}(-\frac{1}{4}\log_2\frac{1}{4} - \frac{3}{4}\log_2\frac{3}{4}) = 1 - 0.54 = 0.46$

(2M)

Thus regardless of whether *= 0 or 1, B has a higher information gain than A and would have been chosen to be the root. Thus the information given is not sufficient to decide the value of *.

Q.5 Let say we have a dice of 4 sides. Where

{ (x,P(x)) : (0,a), (1,(1-a)/3), (2,(1-a)/2), (3,(1-a)/6) }. If the tossing event is observed as

(0,1,2,3,2,3,1,0) then what is the most probable value of *a* [3 Marks]

Ans:
As each events are independent, the likelihood of the event would be

(P(x=0)*P(x=1)*P(x=2)……)

= a*(1-a)/3*(1-a)/2*(1-a)/6*(1-a)/2*(1-a)/6*(1-a)/3*a

So we maximize (ignoring constants in denominator) = a^2*(1-a)^6

Taking gradient and equal to zero

2*a*(1-a)^6 - 6*a^2*(1-a)^5=0

⇨ (1-a)-3a =0

⇨ a = ¼

Q.6 For a linear Support Vector Machine method, positive Points are {(3, 2), (4, 3), (2, 3), (3, -1)} and Negative Points are{(1, 0), (-1, -1), (0, 2), (-1, 2)} [1+4=5Marks]

A. Find the support vectors

B. Determine the equation of hyperplane if it is changed and give a reason if it is not changed for the following two cases

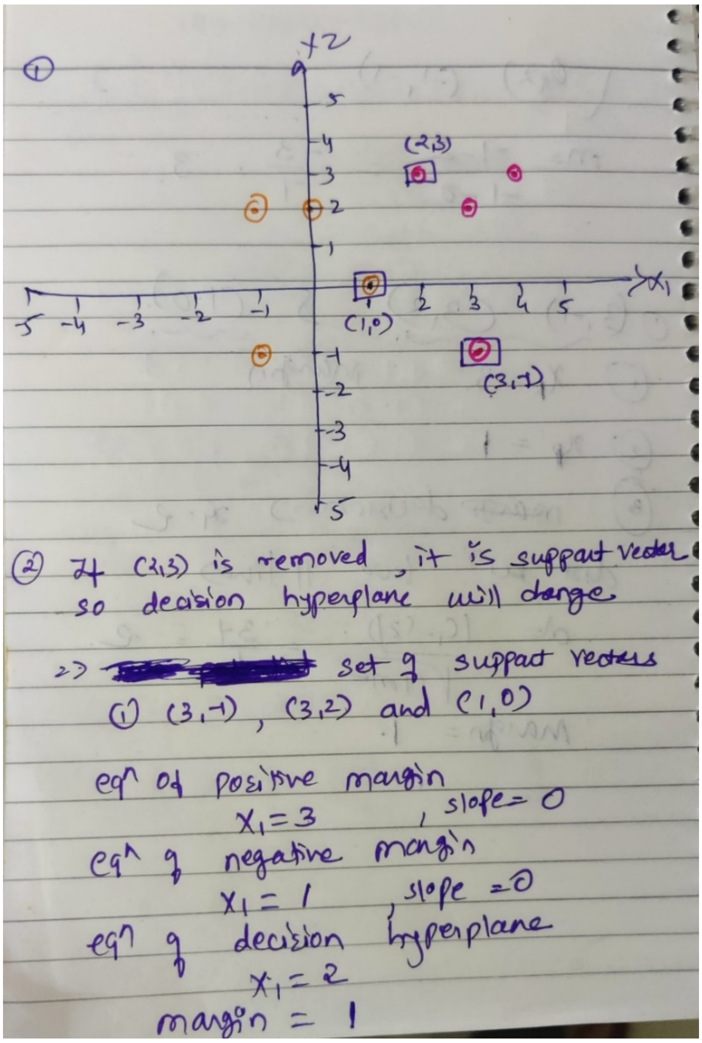
   a. If the point (2, 3) is removed.

   b. If the point (-2,-3) is added

Solution:

A. Support vectors are (2,3), (1,0) and (3,-1) [1M- if one of the SVs is wrong then 0 M]

B. a. If the point (2, 3) is removed. [3M – 1M reason, 2 M for decision boundary equation]


①
x_2
5
4
3
2
1
-1
-2
-3
-4
-5
x_1
-5
-4
-3
-2
-1
1
2
3
4
5
(2,3)
(1,0)
(3,-1)

② If (2,3) is removed, it is support vector so decision hyperplane will change

=> [redacted] set of support vectors

① (3,-1), (3,2) and (1,0)

eqn of positive margin

$X_1 = 3$, slope = 0

eqn of negative margin

$X_1 = 1$, slope = 0

eqn of decision hyperplane

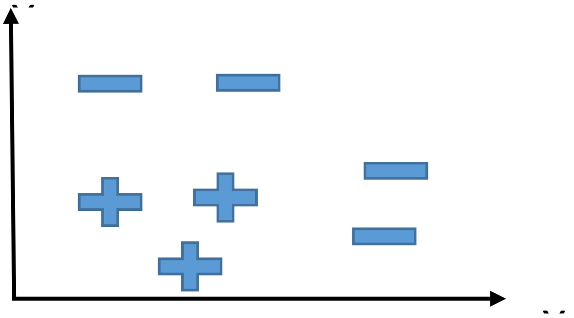
$X_1 = 2$

margin = 1

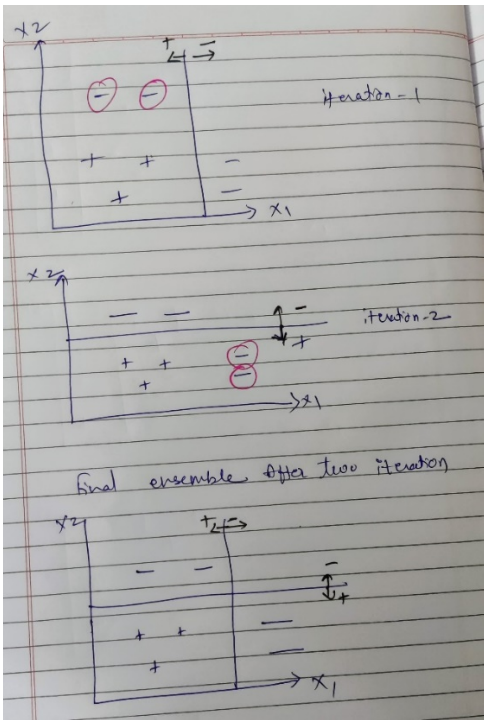
The solution obtained using the Lagrange method is equally acceptable.

b. If the point (-2,-3) is added [1M – 0M if no reason is given]: The equation of the decision hyperplane will not change as the added point is not a support vector.

Q.7 Consider training a boosting classifier using decision stumps on the following data set. Circle the examples which will have their weights increased at the end of each iteration. Run the iteration till zero training error is achieved. [3 Marks]



Solution: No of iteration - 2



Q.8 Assume that a number of points are distributed along the x-axis:

$(-R,0),(R+1,0),\ldots(-1,0),(0,0),(0,1),\ldots(R-1,0),(R,0)$ and an outlier point at $(10R,0)$.

We would like to use the K-Means algorithm to find two clusters for these points.

Initially, one cluster center is chosen to be $(0,0)$. Where should the other cluster center be placed on the x-axis initially so that one of the clusters formed has all the given data points and the other cluster has none? What will be the final locations of the cluster centers? [2+3= 5 Marks]

Answer:

The second cluster center should be chosen initially to be $(X,0)$ where either $X > 20R$ or $X < -2R$.

For these choices of the second center, the K-Means algorithm will ensure that all the given data points are assigned to the first cluster center since every data point will be closer to the first cluster center than the second one.

In the second iteration of the algorithm the first cluster center will be updated from $(0,0)$ to $(\frac{10R}{2R+2},0)$ and the second cluster center will be unchanged.

Birla Institute of Technology & Science, Pilani
Work-Integrated Learning Programmes Division
MTech. Software Engineering at DSE (FC04, FA04_1-2021) Cluster

Second Semester 2021-2022
Endsem -Semester Test
(EC-3 Regular)

Course No.	: DSECLZG565
Course Title	: Machine Learning
Nature of Exam	: Open Book
Weightage	: 40%
Duration	: 2 Hours
Date of Exam	: 25-09-22(FN)

No. of Pages = 2
No. of Questions = 8

Note:

1. Please follow all the *Instructions to Candidates* given on the cover page of the answer book.
2. All parts of a question should be answered consecutively. Each answer should start from a fresh page.
3. Assumptions made if any, should be stated clearly at the beginning of your answer.

Q.1 The results of the election are to be predicted for candidates based on dataset D. There are three different hypotheses h1, h2 and h3 are used to predict the result of candidates winning or losing an election. The probability of h1 given dataset D is 0.5, the probability of h2 given dataset D is 0.3 and the probability of h3 given dataset D is 0.2. Given a new candidate, h1 predicts that a candidate will win the election whereas h2 and h3 predict that candidate will lose the election. What's the most probable classification of a new candidate? [3 Marks]

Solution:

+ = win, - = lose

$P(h_1|D) = .5$, $P(-|h_1) = 0$, $P(+|h_1) = 1$

$P(h_2|D) = .3$, $P(-|h_2) = 1$, $P(+|h_2) = 0$

$P(h_3|D) = .2$, $P(-|h_3) = 1$, $P(+|h_3) = 0$

$$\sum_{h_i \in H} P(+|h_i)P(h_i|D)$$
$$\sum_{h_i \in H} P(-|h_i)P(h_i|D)$$

P(+|D) = 1*0.5+0*0.3+0*0.2=0.5 **[1.5M]**

P(-|D) = 0*0.5+1*0.3+1*0.2=0.5 **[1.5M]**

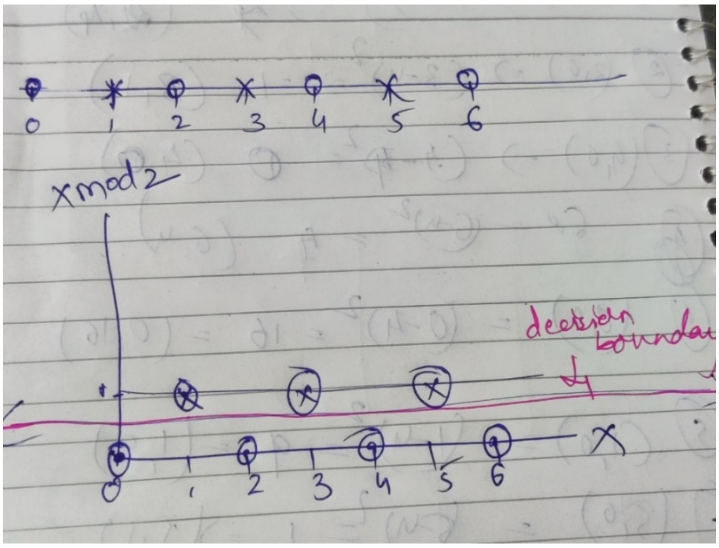
Winning and losing both are equiprobable

Q.2 Use kernel trick and find the equation for hyperplane using nonlinear SVM. Positive Points: {(1,0), (3,0), (5,0)} Negative Points: {(0,0), (2,0), (4,0), (6,0)}. Plot the point before and after the transformation. [5 Marks]

Solution:

Φ(x) = x mod 2 [3M]

Equation of hyperplane : y=0.5 [2M]



Q.3 Suppose we have the following one-dimensional data at -4.0, -.3.0, -2.0, -1.0, 0.0, 1.0, 2.0, 3.0, 4.0. Use the EM algorithm to find a Gaussian mixture model consisting of exactly one Gaussian that fits the data. Assume that the initial mean of the Gaussian is 10.0 and the initial variance is 1.0. [7 Marks]

Answer

First we note that $\pi_1 = 1$ since there is only one Gaussian in the mixture model. Computing the posterior probabilities $P(z_{n1} = 1/x_n) = \gamma(z_{n1})$ we see that the posterior probabilities are all equal to 1 since both the numerator and denominator are equal to $\pi_1 N(x_n/\mu_1, \Sigma_1)$. **[1.5 M]**

Also $N_1 = \sum \gamma(z_{n1}) = N$, the number of data points. [**0.5M]**
This completes the E-step.

In the M-step, we see that

$\mu_1^{new} = \frac{1}{N_1}\sum_{n=1}^{N}\gamma(z_{n1})x_n = \frac{\sum_{n=1}^{n=N} x_n}{N} = \frac{-4.0 + -3.0 + -2.0 + -1.0 + 0.0 + 1.0 + 2.0 + 3.0 + 4.0}{9} = 0.0$ **[2M]**

and $\Sigma_k^{new} = \frac{1}{N_1}\sum_{n=1}^{n=N}(x_n - \mu_1^{new})(x_n - \mu_1^{new})^T$.
Here the x_n and μ_1^{new} are 1×1 matrices and the expression for Σ_k^{new} simplifies to $\frac{\sum_{n=1}^{N} x_n^2}{N}$ which is $\frac{2*(4.0^2+3.0^2+2.0^2+1.0^2)}{9}$=6.66. [2M]

In the next iteration the E-step computes the posterior probabilities to be 1 and the M-step computes the same mean and covariance matrix as above, so the algorithm converges.**[1M]**

Q.4 A dataset D consists of the results of 100 independent coin tosses of the same coin where 30 turn out to be heads and 70 turn out to be tails. Let p be the probability of tossing a head. How many datasets on 100 coin tosses are possible which have the same likelihood as the given dataset D? Determine the maximum likelihood estimate of the parameter p using appropriate calculations. [5 Marks]

Answer: The likelihood of the given data D given p is $P(D/p) = p^{30}(1-p)^{70}$.

Any other dataset D' with the same number of heads and tails as D will have the same likelihood given the same probability p of tossing a head. **[1M]**

There are nCr ways of choosing r locations out of n to place heads.
Therefore the number of datasets that have the same likelihood as D is $100C30 = \frac{100!}{(70!)(30!)}$. **[2M]**

To calculate the value of p that maximizes likelihood we take log to get $\log(P(D/p) = 30\ logp + 70\log(1-p)$.
Then taking the derivative of $\log\,(P(D/p)$ and setting it to zero,
we get $0 = \frac{30}{p} - \frac{70}{1-p}$.
Solving for p we get $p = 0.3$. **[2M]**

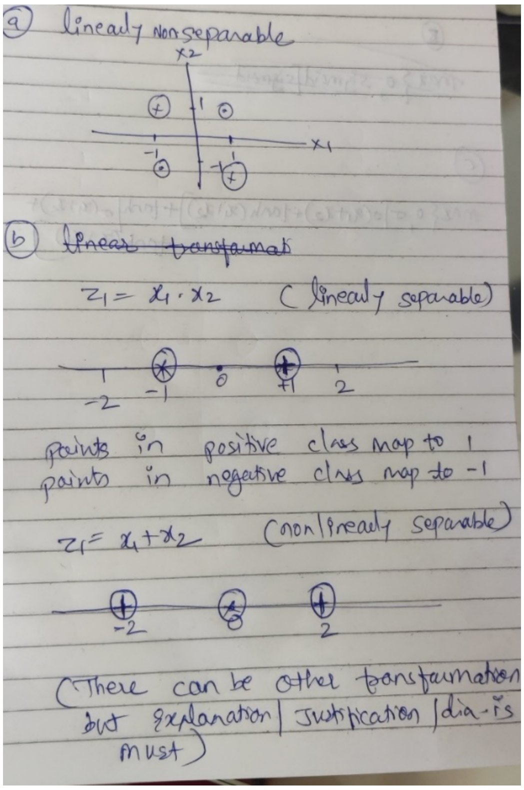
Q.5 Consider a following dataset [5 Marks]

X1	X2	Y
-1	-1	Positive class
-1	1	Negative class
1	-1	Negative class
1	1	Positive class

Answer the following with respect to above dataset

a) Comment on the separability of the dataset with an explanation.[0.5M]
b) Provide the 1-dimensional transformation of this dataset for each linearly and non-linearly separable case with justification[2M]
c) Model the above dataset with an Artificial Neural Network which has two hidden layers, each of which contains two units. Assume that the weights in each layer are set to 1 so that the top unit in each layer applies sigmoid activation to the sum of its inputs and the bottom unit in each layer applies tanh activation to the sum of its inputs. Finally, the single output node applies ReLU activation to the sum of its two inputs. Write the output of this neural network in closed form as a function of x_1 and x_2. (no need to calculate exact values) [2.5M]

Solution:
In b) just transformation with no explanation deduct 0.5M in each case.
In c) any mistake in function – 0 M



(a) linealy Non separable

x_2
1
x_1
-1
-1
1
+
+

(b) ~~linear transfamat~~

$z_1 = x_1 \cdot x_2$ (lineally separable)

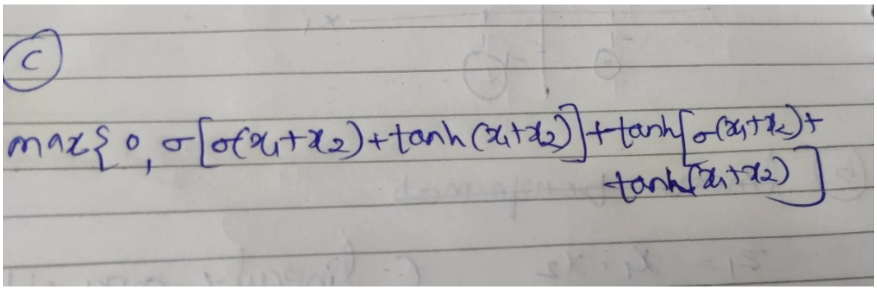
-2
-1
0
+1
2

points in positive class map to 1
points in negative class map to -1

$z_1 = x_1 + x_2$ (non lineally separable)

-2
0
2

(There can be other transfaumation but explanation | Justification | dia. is must)



(c)

$$\max\{0, \sigma[\sigma(x_1+x_2)+\tanh(x_1+x_2)]+\tanh[\sigma(x_1+x_2)+ \tanh(x_1+x_2)]$$

Q.6 Solve the below and find the equation for hyper plane using linear Support Vector Machine method. Positive Points: {(3, 2), (4, 3), (2, 3), (3, -1)} Negative Points: {(1, 0), (-1, -1), (0, 2), (-1, 2)} [5 Marks]

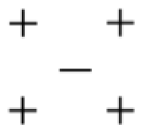
A. Find the support vectors
B. Determine the equation of hyperplane if it is changed and give a reason if it is not changed for the following two cases
   a. If the point (2, 3) is removed.
   b. If the point (5,4) is added

Solution:

A. Support vectors are (2,3), (1,0) and (3,-1) [1M- if one of the SVs is wrong then 0 M]
   Equation of decision Hyperplane - 2M
   The solution obtained using the Lagrange method or by geometrical inspection is acceptable
B. a. If the point (2, 3) is removed. [1M – 0.5M if no reason is given]:
   The equation of the decision hyperplane will change as the removed point is the support vector.

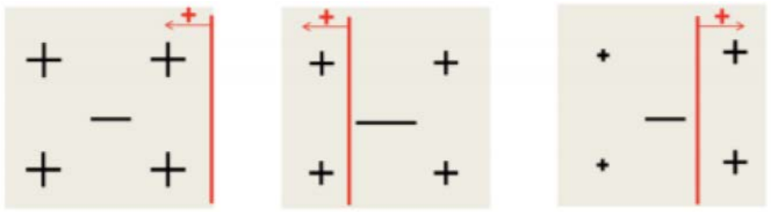
   b. If the point (5,4) is added [1M – 0.5M if no reason is given]: The equation of the decision hyperplane will not change as the added point is not support vector.

Q.7 Consider training a boosting classifier using decision stumps on the following data set. Circle the examples which will have their weights increased at the end of each iteration. Run the iteration till zero training error is achieved. [3 Marks]



Solution:

No of iteration - 3



Q.8 Consider the following dataset. [7 Marks]

x1	-1	-1	-1	-1	0	4	4
x2	2	1	-1	-2	0	2	-2
Class label	1	1	1	1	1	2	2

The given class label exhibits two natural clusters formed in the given dataset and acts as a ground truth. Now remove class labels and use the K-means clustering algorithm to find the 2 clusters by initializing two cluster centres as follows:

A. C1(-1,2) and C2(0,0)
B. C1(-0.5,0) and C2(0,0)

For both the above cases run the algorithm till centres do not change (convergence criteria) and give the final cluster assignment [2+2M]. In each case, comment on the correctness of cluster assignment. [1+1M] Also, comment in no more than 20 words on the drawback of k-means which is depicted in above two cases.[1M]

Solution:

A.

	c1				c2		
x1	-1	-1	-1	-1	0	4	4
x2	2	1	-1	-2	0	2	-2
d1	0	1	3	4	2.236068	5	6.403124
d2	2.236068	1.414214	1.414214	2.236068	0	4.472136	4.472136
cluster	1	1	2	2	2	2	2

	new c1	new c2					
x1	-1	1.2					
x2	1.5	-0.6					
d1	0.5	0.5	2.5	3.5	1.802776	5.024938	6.103278
d2	3.405877	2.720294	2.236068	2.607681	1.341641	3.820995	3.130495
cluster	1	1	2	2	2	2	2

	new c1	new c2
x1	-1	1.2
x2	1.5	-0.6

Comment on cluster assignment:

Algorithm has converged after 2 iterations but the cluster assignment does not depict the natural clusters in the datasets as given by the ground truth.

B

	c1				c2			c1
x1	-1	-1	-1	-1	0	4	4	-0.5
x2	2	1	-1	-2	0	2	-2	0
d1	2.061553	1.118034	1.118034	2.061553	0.5	4.924429	4.924429	
d2	2.236068	1.414214	1.414214	2.236068	0	4.472136	4.472136	
cluster	1	1	1	1	2	2	2	

	new c1	new c2					
x1	-1	2.666667					
x2	0	0					
d1	2	1	1	2	1	5.385165	5.385165
d2	4.176655	3.800585	3.800585	4.176655	2.666667	2.403701	2.403701
cluster	1	1	1	1	1	2	2

	new c1-mean	new c2					
x1	-0.8	4					
x2	0	0					
d1	2.009975	1.019804	1.019804	2.009975	0.8	5.2	5.2
d2	5.385165	5.09902	5.09902	5.385165	4	2	2
cluster	1	1	1	1	1	2	2
	new c1	new c2					
x1	-0.8	4					
x2	0	0					

Comment on cluster assignment:

Algorithm has converged after 3 iterations and the cluster assignment shows the natural clusters in the datasets as given by the ground truth.

The drawback of k-means as demonstrated by the above two cases:

Correctness of the K- means the algorithm is sensitive to the initialization of cluster centres.

Format of Question paper for Comprehensive Test

Birla Institute of Technology & Science, Pilani
Work Integrated Learning Programmes Division
Second Semester 2020-2021
Mid-Semester Test
(EC-3 Makeup)

SOLUTION

Course No.	: SS ZG568
Course Title	: Applied Machine Learning
Nature of Exam	: Open Book
Weightage	: 40%
Duration	: 2 Hours
Date of Exam	:

No. of Pages =
No. of Questions = 5

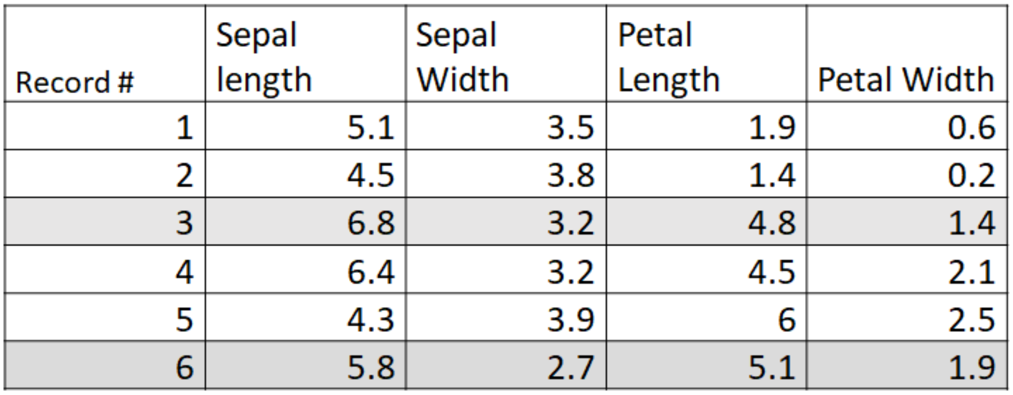
Note to Students:

1. Please follow all the *Instructions to Candidates* given on the cover page of the answer book.
2. All parts of a question should be answered consecutively. Each answer should start from a fresh page.
3. Assumptions made if any, should be stated clearly at the beginning of your answer.

Please Note:
A **maximum of 12 main questions** with Q.Nos. 1 to N**. Please do not include any objective type questions. Please do not include any reference material like data tables in the question paper, as all Mid-Semester Test are completely Open Book examinations.**

Q.1 Set. (A) Marks 3+2+3=8

Apply K-Mean Clustering algorithm for following dataset. Assume K=2 and two centres for first iteration are shown in below (shaded rows). Assume initially record # 3 belongs to class 1 and 6th one belongs to class 2.



Record #	Sepal length	Sepal Width	Petal Length	Petal Width
1	5.1	3.5	1.9	0.6
2	4.5	3.8	1.4	0.2
3	6.8	3.2	4.8	1.4
4	6.4	3.2	4.5	2.1
5	4.3	3.9	6	2.5
6	5.8	2.7	5.1	1.9

A. After the first iteration, which record #s belong to the 1st class and 2nd classes?
Record # 1, 2, 3 and 4 belong to class 1. Record # 5 and 6 belong to class 2.

B. At the next iteration, what will be new cluster centers?

Sepal length	Sepal width	Petal Length	Petal Width	
5.7	3.425	3.15	1.075	**Cluster center 1**
5.05	3.3	5.55	2.2	**Cluster center 2**

C. At the next iteration, which record #s belongs to the 1st class and 2nd class?
   Record # 1, 2, 3 belong to cluster 1
   Record # 4, 5, 6 belong to cluster 2

Q.1 Set. (B) Marks 3+2+3 = 8

Apply K-Mean Clustering algorithm for following dataset. Assume K=2 and two centres for first iteration are shown in below (shaded rows). Assume initially record # 1 belongs to class 1 and 4th one belongs to class 2.

Record #	Sepal length	Sepal Width	Petal Length	Petal Width
1	5.1	3.5	1.9	0.6
2	4.5	3.8	1.4	0.2
3	6.8	3.2	4.8	1.4
4	6.4	3.2	4.5	2.1
5	4.3	3.9	6	2.5
6	5.8	2.7	5.1	1.9

A. After the first iteration, which record #s belong to the 1st class and 2nd classes?
   Record # 1, 2 belong to class 1. Record # 3, 4, 5 and 6 belong to class 2.

B. At the next iteration, what will be new cluster centers?

4.8	3.65	1.65	0.4	**Cluster center #1**
5.825	3.25	5.1	1.975	**Cluster center #2**

C. At the next iteration, which record #s belongs to the 1st class and 2nd class?
   Record # 1, 2 belong to class 1. Record # 3, 4, 5 and 6 belong to class 2.

Q.1 Set. (C) Marks 3+2+3=8

Apply K-Mean Clustering algorithm for following dataset. Assume K=2 and two centres for first iteration are shown in below (shaded rows). Assume initially record # 2 belongs to class 1 and 5th one belongs to class 2.

Record #	Sepal length	Sepal Width	Petal Length	Petal Width
1	5.1	3.5	1.9	0.6
2	4.5	3.8	1.4	0.2
3	6.8	3.2	4.8	1.4
4	6.4	3.2	4.5	2.1
5	4.3	3.9	6	2.5
6	5.8	2.7	5.1	1.9

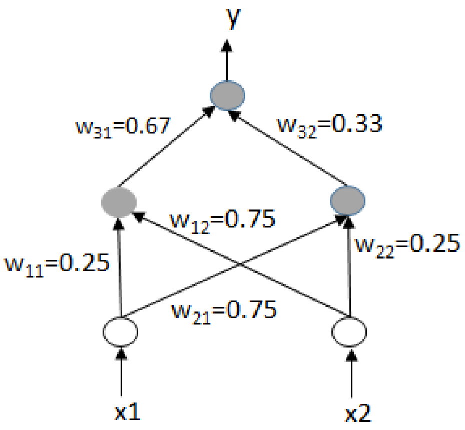
A. After the first iteration, which record #s belong to the 1st class and 2nd classes?
Record # 1, 2 belong to class 1. Record # 3, 4, 5 and 6 belong to class 2.

B. At the next iteration, what will be new cluster centers?

4.8	3.65	1.65	0.4	**Cluster center #1**
5.825	3.25	5.1	1.975	**Cluster center #2**

C. At the next iteration, which record #s belongs to the 1st class and 2nd class?
Record # 1, 2 belong to class 1. Record # 3, 4, 5 and 6 belong to class 2.

Q.2 Set. (A) Marks 1.5+1.5+2+3=8



Leaky ReLU activation function is used in the hidden nodes. Leaky ReLU activation function generates output = input, if input >= 0, and 0.1 * input if output < 0. Sigmoid activation function is used in the output node. Assume, squared difference between the actual and target output is used as the loss function, derivative of leaky ReLU activation function = 0 at input = 0, and zero bias at all hidden and output nodes. Learning rate is 0.1. x1=x2=1 and target output=0.

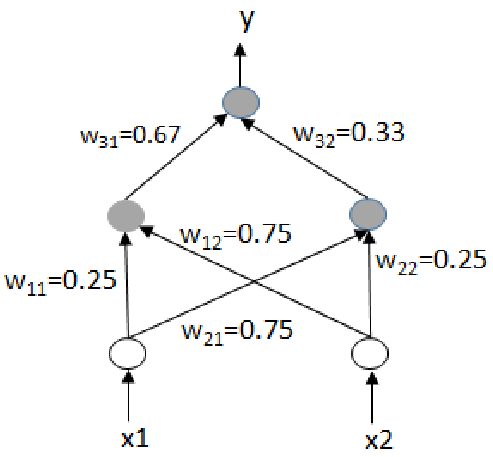
A. What is the actual output for the given weights at the current iteration?
o = e/(1+e)

B. What is the value of the loss function?
Loss = 0.5*(e/(1+e))^2

C. What will be the weight w_{31} in the next iteration?
o = output of output node, net3 = total input to the output node, o1=output of left hidden node
w_{31}=w_{31}-0.1dLoss/do*do/dnet3*dnet3/dw_{31}=0.67-0.1*o*o*(1-o)*1
=0.67-0.1e^2/(1+e)^2*1/(1+e)=0.67-0.1e^2/(1+e)^3

D. What will be the weight w_{12} in the next iteration?
w_{12}=w_{12}-0.1dLoss/dnet3*dnet3/do1*do1/w_{12}=0.75-0.1*o*o*(1-o)*w_{31}*1*x2
=0.75 – 0.067e^2/(1+e)^3

Q.2 Set. (B) Marks 1.5+1.5+2+3=8

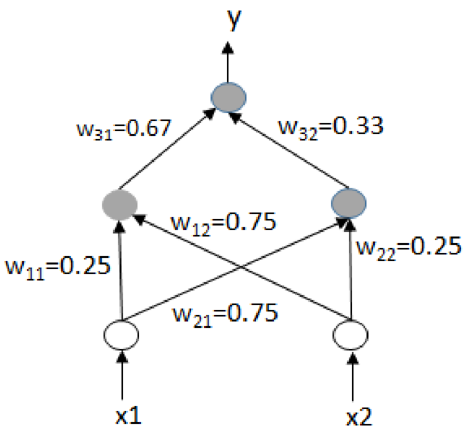


Leaky ReLU activation function is used in the hidden nodes. Leaky ReLU activation function that generates output = input, if input >= 0, and 0.1 * input if output < 0. Sigmoid activation function is used in the output node. Assume, squared difference between the actual and target output is used as the loss function, derivative of leaky ReLU activation function = 0 at input = 0, and zero bias at all hidden and output nodes. Learning rate is 0.1. x1=x2=0 and target output=1.

A. What is the actual output for the given weights at the current iteration?
   Output y = 0.5
B. What is the value of the loss function?
   Loss = $0.5*0.5^2 = 0.125$
C. What will be the weight w_{32} in the next iteration?
   $\Delta w_{32} = -0.1*\delta Loss/\delta w_{32} = -0.1*\delta Loss/\delta y* \delta y/\delta net_3 * \delta net_3/\delta w_{32}$
   $= -0.1*\delta Loss/\delta y* \delta y/\delta net_3 * o_2 = 0$
   $w_{32}=0.33$
D. What will be the weight w_{22} in the next iteration?
   $\Delta w_{22} = -0.1*\delta Loss/\delta w_{22} = -0.1*\delta Loss/\delta net_3 * \delta net_3/\delta o_2 * \delta o_2/\delta net_2 * \delta net_2/\delta w_{22}$
   $= -0.1* \delta Loss/\delta net_3 * \delta net_3/\delta o_2 * 0 * 0 = 0$
   $w_{22}=0.25$

Q.2 Set. (C) Marks 1.5+1.5+2+3=8

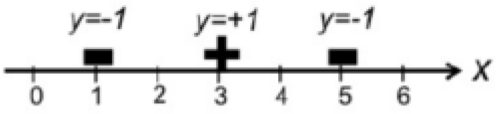
Leaky ReLU activation function is used in the hidden nodes. Leaky ReLU activation function that generates output = input, if input >= 0, and 0.1 * input if output < 0. Sigmoid activation function is used in the output node. Assume, squared difference between the actual and target output is used as the loss function, derivative of leaky ReLU activation function = 0 at input = 0, and zero bias at all hidden and output nodes. Learning rate is 0.1. x1=1, x2=0 and target output=0.



A. What is the actual output for the given weights at the current iteration?
   y= 1/(1+exp(-0.67*0.25-0.33*0.75))=0.6023
B. What is the value of the loss function?
   Loss=0.5*(0.6023-0)^2=0.1814
C. What will be the weight w_{31} in the next iteration with standard gradient descent?
   w_{31}=0.67-0.1δLoss/δw_{31}=0.67-0.1*0.6023*0.6023* (1-0.0.6023)*0.25=0.6664
D. What will be the weight w_{11} in the next iteration with standard gradient descent?
   w_{11}=0.25-0.1δLoss/δw_{11}=0.25-0.1*0.6023*0.6023* (1-0.6023)*0.67=0.2403

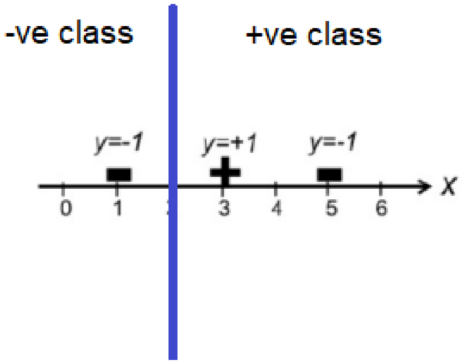
Q.3 Set. (A) 1+1+2+2+1+1=8 Marks

Consider training an AdaBoost classifier using decision stumps on the following data set. Decision stump classifier chooses a constant value c and classifies all points where x > c as one class and other points where x ≤ c as the other class.



A. What is the initial weight that is assigned to each data point? 1/3, 1/3, 1/3

B. Show the decision boundary for the first decision stump (indicate the positive and negative side of the decision boundary).



Note: Multiple valid solutions exist.

C. Calculate the importance of the first decision stump.

Data at x=5 is misclassified. So,
Error rate $\mathcal{E}_1$= 1/3*1/3=1/9
Importance $\alpha_1 = 0.5 \ln((1- \mathcal{E}_1) / \mathcal{E}_1) = 0.5\ln(8)$

D. Calculate the weights for the next iteration increases in the boosting process

w1 = 1/3 * exp(-0.5ln(8)) / Z , corresponding to data @ x=1
w2 = 1/3 * exp(-0.5ln(8)) / Z , corresponding to data @ x=3
w3 = 1/3 * exp(0.5ln(8)) / Z , corresponding to data @ x=5
Z = 1/3 * (2 exp(-0.5ln(8))+ exp(0.5ln(8))

E. How does bagging process in random forest technique impact variance?

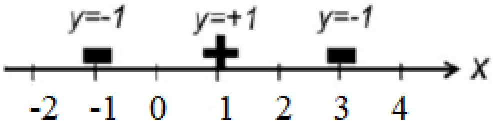
Reduces variance

F. How does feature randomization process in random forest technique impact bias?

Reduces bias

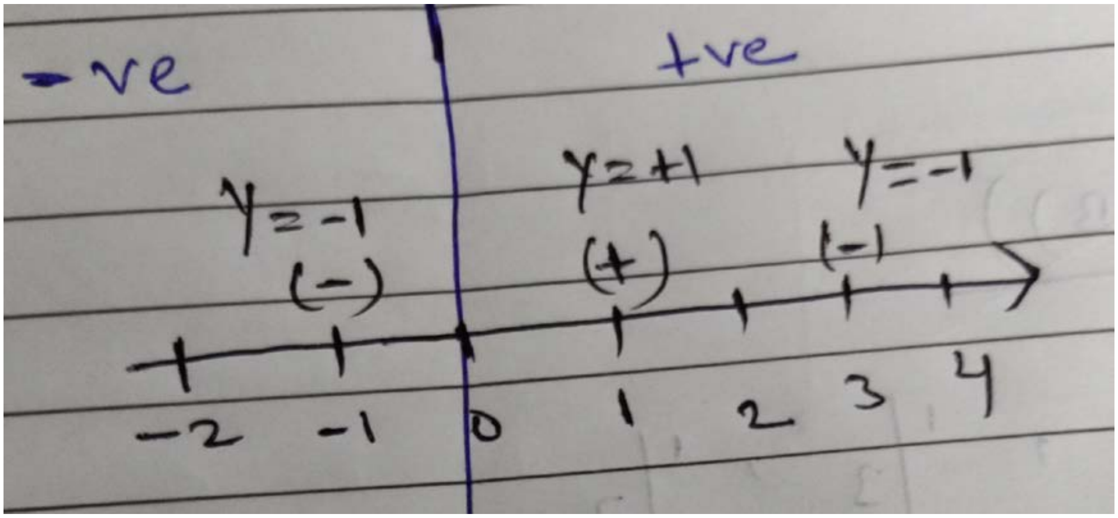
Q.3 Set. (B) 1+1+2+2+1+1=8 Marks

Consider training an AdaBoost classifier using decision stumps on the following data set. Decision stump classifier chooses a constant value c and classifies all points where x > c as one class and other points where x ≤ c as the other class.



A. What is the initial weight that is assigned to each data point? 1/3, 1/3, 1/3

B. Show the decision boundary for the first decision stump (indicate the positive and negative side of the decision boundary).



C. Calculate the importance of the first decision stump.

Data at x=3 is misclassified. So,
Error rate $\mathcal{E}1$= 1/3*1/3=1/9
Importance $\alpha 1 = 0.5 \ln((1- \mathcal{E}1) / \mathcal{E}1) = 0.5\ln(8)$

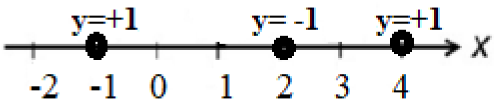
D. Calculate the weights for the next iteration increases in the boosting process.
w1 = 1/3 * exp(-0.5ln(8)) / Z , corresponding to data @ x=-1
w2 = 1/3 * exp(-0.5ln(8)) / Z , corresponding to data @ x=1
w3 = 1/3 * exp(0.5ln(8)) / Z , corresponding to data @ x=3
Z = 1/3 * (2 exp(-0.5ln(8))+ exp(0.5ln(8))

E. How does bagging process in random forest technique impact variance?
It reduces variance

F. How does feature randomization process in random forest technique impact bias?
It reduces bias

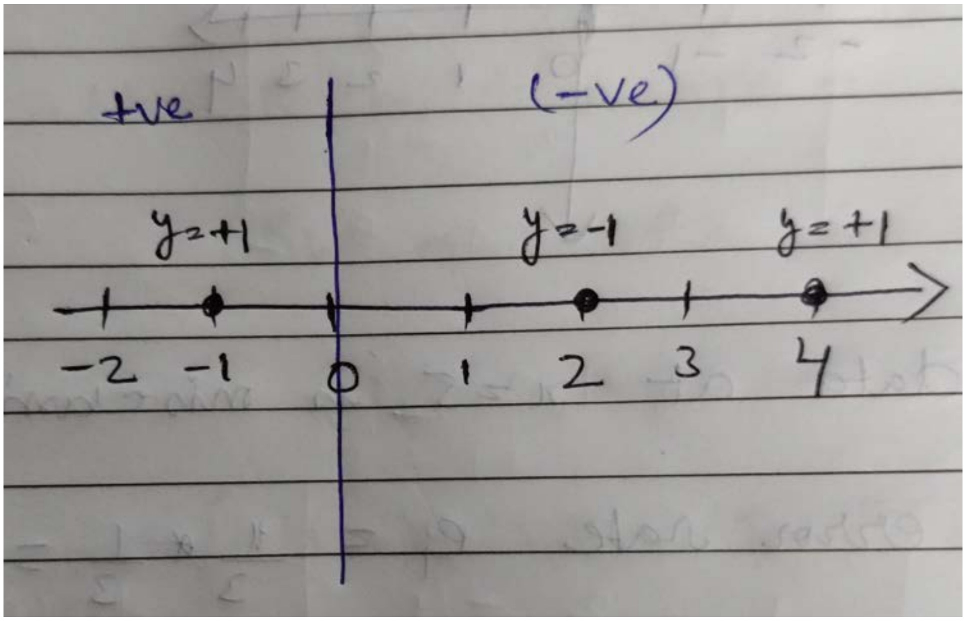
Q.3 Set. (C) 1+1+2+2+1+1=8 Marks

Consider training an AdaBoost classifier using decision stumps on the following data set. Decision stump classifier chooses a constant value c and classifies all points where x > c as one class and other points where x ≤ c as the other class.



A. What is the initial weight that is assigned to each data point?1/3, 1/3, 1/3

B. Show the decision boundary for the first decision stump (indicate the positive and negative side of the decision boundary).



C. Calculate the importance of the first decision stump.

Data at x=4 is misclassified. So,
Error rate Ɛ1= 1/3*1/3=1/9
Importance α1 = 0.5 ln((1- Ɛ1) / Ɛ1) = 0.5ln(8)

D. (D)Calculate the weights for the next iteration increases in the boosting process

w1 = 1/3 * exp(-0.5ln(8)) / Z , corresponding to data @ x= -1
w2 = 1/3 * exp(-0.5ln(8)) / Z , corresponding to data @ x=2
w3 = 1/3 * exp(0.5ln(8)) / Z , corresponding to data @ x=4
Z = 1/3 * (2 exp(-0.5ln(8))+ exp(0.5ln(8))

E. (E)How does bagging process in random forest technique impact variance?

It reduces variance

F. (F)How does feature randomization process in random forest technique impact bias?

It reduces bias

Q.4 Set. (A) 3+2+3=8 marks

Consider the following data

Input x	Output y
0	0
1	1
2	5
4	15

A. Determine the best parameters of the regression model given by output = w0+w1*x^2 + w2*x^3.

$$\begin{pmatrix} 1 & 0 & 0 \\ 1 & 1 & 1 \\ 1 & 4 & 8 \\ 1 & 16 & 64 \end{pmatrix} \begin{pmatrix} w0 \\ w1 \\ w2 \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \\ 5 \\ 15 \end{pmatrix}$$

[w0 w1 w2] = [-0.16957 1.58773 -0.15978]

B. Assume learning rate α=0.5. At iteration t, w0=2, w1=1 and w2=0. What will be the values of w0, w1 and w2 at next iteration t+1?

$$\theta_j = \theta_j - \frac{\alpha}{n}\sum_{i=1}^{n}\left(h_{\theta}\left(x^{(i)}\right) - y^{(i)}\right)x_j^{(i)} - \alpha\lambda\,\text{sign}(\theta_j)$$

Input x	Output y	$x_1=x^2$	$x_2=x^3$	$h_\theta(x)=w0+w1x_1+w2x_2$
0	0	**0**	**0**	**2**
1	1	**1**	**1**	**3**
2	5	**4**	**8**	**6**
4	15	**16**	**64**	**18**

w0 = 2 – 0.5/4 [2+2+1+3]=1
w1=1-0.5/4[2*0+2*1+1*4+3*16] = -5.75
w2=0-0.5/4[2*0+2*1+1*8+3*64]= -25.25

C. Assume L1 regularization with regularization constant λ=1 and learning rate α=0.5. At iteration t, w0=2, w1=1 and w2=0. What will be the values of w0, w1 and w2 at the next iteration t+1? Assume zero is a positive integer.
w0=1, no regularization applied
w1=-5.75-0.5=-6.25
w2=-25.25-0.5=-25.75

Q.4 Set. (B) 3+2+3=8 Marks

Consider the following data

Input x	Output y
0	0
1	1
2	10
4	60

A. Determine the best parameters of the regression model given by output = w0+w1*x + w2*x^3.
[w0 w1 w2] = [-0.41 1.51 0.85]

B. Assume learning rate α=0.5. At iteration t, w0=2, w1=0 and w2=1. What will be the values of w0, w1 and w2 at next iteration t+1?

Input x	Output y	$x_1=x$	$x_2=x^3$	$h_\theta(x)=w0+w1x_1+w2x_2$
0	0	**0**	**0**	**2**
1	1	**1**	**1**	**3**
2	10	**2**	**8**	**10**
4	60	**4**	**64**	**66**

w0 = 2 – 0.5/4 [2+2+0+6] = 1.875 w1=0-0.5/4[2*0+2*1+0*2+6*4] = -3.25
w2=1-0.5/4[2*0+2*1+0*8+6*64] = -47.25

C. Assume L1 regularization with regularization constant λ=1 and learning rate α=0.5. At iteration t, w0=2, w1=0 and w2=1. What will be the values of w0, w1 and w2 at next iteration t+1? Assume zero is a positive integer.

w0=1.875, no regularization applied
w1=-3.25-0.5=-3.75
w2=-47.25-0.5=-47.75

Q.4 Set. (C) 2+4+2 = 8 Marks

Consider the following data

Input x	Output y
0	0
1	1
2	3
4	3

A. Determine the best parameters of the regression model given by output = w0+w1*x + w2*x^2.
[w0 w1 w2] = [-0.191 1.995 -0.295]

B. Assume learning rate α=0.5. At iteration t, w0=2, w1=1 and w2=0. What will be the values of w0, w1 and w2 at next iteration t+1?

Input x	Output y	x_1=x	$x_2=x^2$	$h_\theta(x)=w0+w1x_1+w2x_2$
0	0	**0**	**0**	**2**
1	1	**1**	**1**	**3**
2	3	**2**	**4**	**4**
4	3	**4**	**16**	**6**

w0 = 2 – 0.5/4 [2+2+1+3] = 1
w1=1-0.5/4[2*0+2*1+1*2+3*4] = -1
w2=0-0.5/4[2*0+2*1+1*4+3*16] = -6.75

C. Assume L1 regularization with regularization constant λ=1 and learning rate α=0.5. At iteration t, w0=2, w1=1 and w2=0. What will be the values of w0, w1 and w2 at next iteration t+1? Assume zero is a positive integer.

w0=1, no regularization applied
w1=-1-0.5=-1.5
w2=-6.75-0.5=-7.25

Q.5 Set. (A) 3+1+2+2=8 Marks

Training data is represented by the following matrix **A**, whose rows corresponding to training records and columns corresponding to input dimensions.

$$\mathbf{A} = \begin{pmatrix} 2 & 0 \\ -3 & 1 \\ 1 & -3 \\ 0 & 2 \end{pmatrix}$$

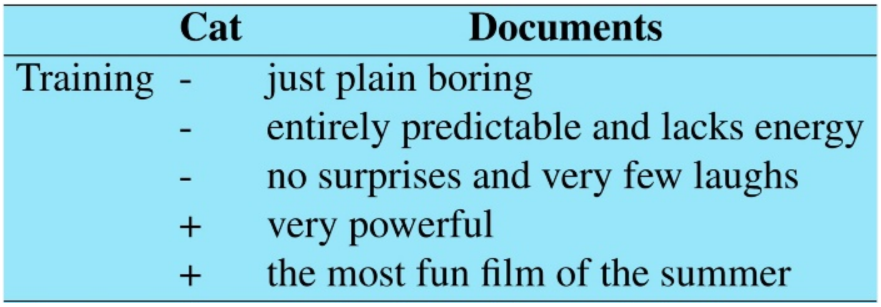
A. What is the best possible projection of this data in one-dimension that minimizes the loss of information?

Covariance matrix = A'*A (note: A is zero mean)
Eigenvector of A'*A corresponding to largest eigenvalue=20 is e1 = [-0.707 0.707]
Best possible projection: A*e1'= [-1.4142 2.8284 -2.8284 1.4142]'

B. Show the 3x3 convolution kernel for implementing mean pooling in gray scale images.

[1/9 1/9 1/9; 1/9 1/9 1/9; 1/9 1/9 1/9]

Consider an LSTM network with one hidden layer of 20 nodes used for sentiment prediction of the following data. No bias is used in any of the nodes. One-hot representation is used for representing the input and output.



	Cat	Documents
Training	-	just plain boring
	-	entirely predictable and lacks energy
	-	no surprises and very few laughs
	+	very powerful
	+	the most fun film of the summer

C. How many input and output nodes will be needed in the LSTM network? Assume each training data starts with a <SOS> and ends with a <EOS> token.

Number of Unique words = 21. So, # of input nodes = 21+2=23, including the <SOS> and <EOS> tokens

# of output nodes = 2

D. What will be the total number of trainable weights in the LSTM network? (Show individual numbers for all LSTM components for partial marking)

4*(23+20)*20+20*2

Q.5 Set. (B) 3+2+1+2=8 Marks

Training data is represented by the following matrix **A**, whose rows corresponding to training records and columns corresponding to input dimensions.

$$\mathbf{A} = \begin{pmatrix} 2 & 0 \\ -2 & 1 \\ 1 & -3 \\ -1 & 2 \end{pmatrix}$$

A. What is the best possible projection of this data in one-dimension that minimizes the loss of information?

P= [-1.2044 2.0027 -2.9972 2.1989]’

B. What is that minimum value of total loss over the training data?

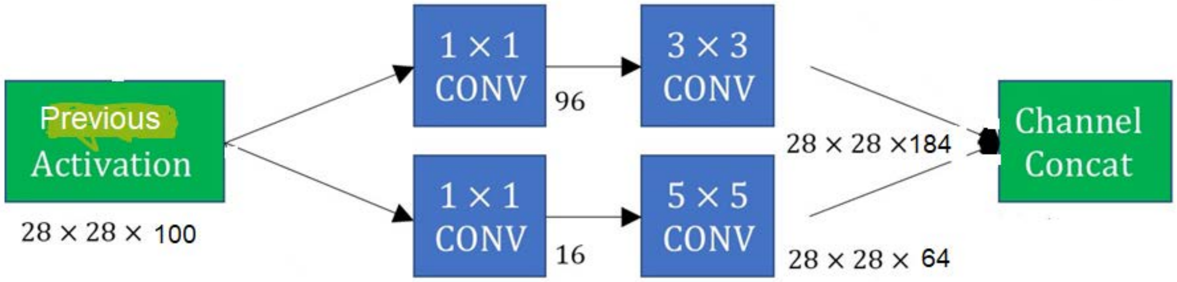
||A-P*e1’||=2.1725

C. Show the 5x5 convolution kernel for implementing mean pooling in a gray scale image.

[1/25 1/25 1/25 1/25 1/25; 1/25 1/25 1/25 1/25 1/25; 1/25 1/25 1/25 1/25 1/25]

Consider the following inception blocks in a hypothetical GoogLeNet architecture.

D. What is the size of the output post concatenation of the channels?
   28x28x248
E. What is the total number of trainable parameters? Assume no bias is used.
   100*96+100*16+96*3*3*184+16*5*5*64



Q.5 Set. (C) 3+3+2=8 Marks

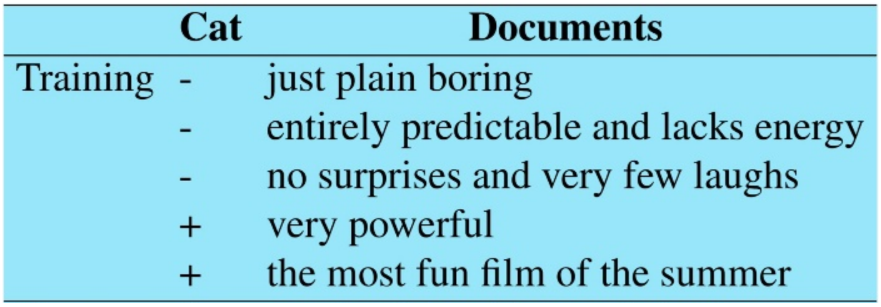
Training data is represented by the following matrix **A**, whose rows corresponding to training records and columns corresponding to input dimensions.

$$\mathbf{A} = \begin{pmatrix} 2 & 0 \\ -3 & 2 \\ 1 & -1 \\ 0 & -1 \end{pmatrix}$$

A. What is the best possible projection of this data in one-dimension that minimizes the loss of information?

   Covariance matrix = A'*A (note: A is zero mean)
   Eigenvector of A'*A corresponding to largest eigenvalue=18.1 is e1 = [-0.865 0.502]
   Best possible projection: A*e1'= [-1.72982 3.59858 -1.36684 -0.50193]'

Consider a vanilla RNN network with one hidden layer of 10 nodes used for sentiment prediction of the following data. One-hot representation is used for representing the input and output.



	Cat	Documents
Training	-	just plain boring
	-	entirely predictable and lacks energy
	-	no surprises and very few laughs
	+	very powerful
	+	the most fun film of the summer

B. How many input and output nodes will be needed in the RNN network? Assume each training data starts with a <SOS> and ends with a <EOS> token.

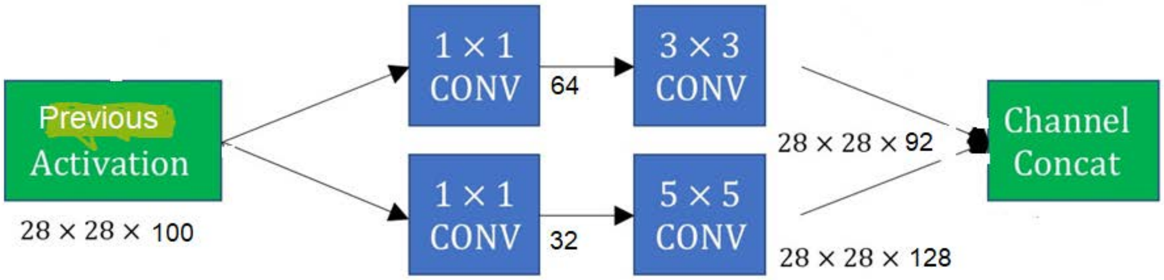
Number of Unique words = 21. So, # of input nodes = 21+2=23, including the <SOS> and <EOS> tokens
# of output nodes = 2

C. What will be the total number of trainable parameters (including bias) in the RNN network? (Show individual numbers for all RNN components for partial marking)
(23+10)*10 + 10*2 + 10+2

Consider the following inception blocks in a hypothetical GoogLeNet architecture.

D. What is the size of the output post concatenation of the channels?
28x28x220 = 172480

E. What is the total number of trainable parameters? Assume no bias is used.
100*64 + 100*32 + 64*3*3*92 + 32*5*5*128 = 164992


Previous Activation
28 × 28 × 100
1 × 1 CONV
64
3 × 3 CONV
1 × 1 CONV
32
5 × 5 CONV
28 × 28 × 92
28 × 28 × 128
Channel Concat