INDIAN INSTITUTE OF TECHNOLOGY KHARAGPUR

Stamp / Signature of the Invigilator

EXAMINATION (End Semester)	SEMESTER (Autumn)

Roll Number										Section		Name	

Subject Number	C	S	6	0	0	5	0	Subject Name	*MACHINE LEARNING*

Department / Center of the Student		Additional sheets	

Instructions and Guidelines to Students Appearing in the Examination

1. Ensure that you have occupied the seat as per the examination schedule.
2. Ensure that you do not have a mobile phone or a similar gadget with you even in switched off mode. Note that loose papers, notes, books should not be in your possession, even if those are irrelevant to the paper you are writing.
3. Data book, codes or any other materials are allowed only under the instruction of the paper-setter.
4. Use of instrument box, pencil box and non-programmable calculator is allowed during the examination. However, exchange of these items is not permitted.
5. Additional sheets, graph papers and relevant tables will be provided on request.
6. Write on both sides of the answer script and do not tear off any page. Report to the invigilator if the answer script has torn page(s).
7. Show the admit card / identity card whenever asked for by the invigilator. It is your responsibility to ensure that your attendance is recorded by the invigilator.
8. You may leave the examination hall for wash room or for drinking water, but not before one hour after the commencement of the examination. Record your absence from the examination hall in the register provided. Smoking and consumption of any kind of beverages is not allowed inside the examination hall.
9. After the completion of the examination, do not leave the seat until the invigilator collects the answer script.
10. During the examination, either inside the examination hall or outside the examination hall, gathering information from any kind of sources or any such attempts, exchange or helping in exchange of information with others or any such attempts will be treated as adopting 'unfair means'. Do not adopt 'unfair means' and do not indulge in unseemly behavior as well.

Violation of any of the instructions may lead to disciplinary action.

Signature of the Student

To be filled in by the examiner

Question Number	1	2	3	4	5	6	7	8	9	10	Total
Marks Obtained											

Marks obtained (in words)	Signature of the Examiner	Signature of the Scrutineer

Instructions: Answer all TEN questions. Time = 3hrs. Total marks = 60. Question 1 has one mark per answer box. Questions 2-10 have two marks per answer box. Write your answers only in the space provided. The question paper has total 12 pages.

Rough Work

1. State whether the following statements are true/false (T/F).

i. Kernel matrices are always positive definite.	T/F
ii. The backpropagation algorithm employs gradient descent.	T
iii. The perceptron learning rule converges for linearly separable data.	T
iv. Complete linkage clustering is computationally cheaper compared to single linkage.	F
v. K-Means clustering is computationally cheaper compared to single linkage clustering.	T
vi. Sequential Forward Search always generates the optimal feature subset.	F
vii. Classifiers having lower bias have a higher variance.	T
viii. A weak learner for a binary classification problem has error probability more than 0.5.	F
ix. The hypothesis class of half planes shatter any three non-collinear points in two-dimension.	T
x. Sample complexity of learning a hypothesis class increases with its VC-dimension.	T

2(a). A hypothetical SVM model has the following values of lagrange multipliers α and support vectors:

α	*Support vector*	*y*
1	(1, −1, 1)	+1
0.5	(0, 2, −1)	−1
1	(−1, 0, 2)	−1

Suppose that the linear kernel is used. Compute the output *y* of this SVM model when the input feature vector is (0.3, 0.8, 0.6).

-1/-1.3

2(b). Suppose we have four training examples in two dimensions, positive examples at *X1 = [0, 0]*, *X2 = [2, 2]*, and negative examples at *X3 = [h, 1]*, *X4 = [0, 3]*, where we treat $0 \leq h \leq 3$ as a parameter.

i. How large can *h* be so that the training points are still linearly separable?

<= 1

ii. What is the margin achieved by the maximum margin boundary as a function of *h*?

$$\frac{1-h}{\sqrt{2}}$$

3(a). A kernel function *K(x, z)* measures the similarity between two instances *x* and *z* in a transformed space. For a feature transform $x \rightarrow \phi(x)$ the kernel function is $K(x, z) = \phi(x).\phi(z)$. Consider the two dimensional input vectors $x = (x_1, x_2)$. For each of the kernel function below what is the corresponding feature transform?

i. $K(x, z) = 1 + x.z$

B

(A) $\phi(x) = (x_1, x_2)$ (B) $\phi(x) = (1, x_1, x_2)$ (C) $\phi(x) = (x_1^2, x_2^2)$ (D) $\phi(x) = (1, x_1^2, x_2^2)$

ii. $K(x, z) = (x.z)^2$

C

(A) $\phi(x) = (x_1^2, x_2^2)$ (B) $\phi(x) = (1, x_1^2, x_2^2)$

(C) $\phi(x) = (x_1^2, x_2^2, \sqrt{2}\, x_1x_2)$ (D) $\phi(x) = (1, x_1^2, x_2^2, \sqrt{2}\, x_1x_2)$

iii. $K(x, z) = (1 + x.z)^2$

C

(A) $\phi(x) = (1, x_1^2, x_2^2)$ (B) $\phi(x) = (1, x_1^2, x_2^2, \sqrt{2}\, x_1x_2)$

(C) $\phi(x) = (1, x_1^2, x_2^2, \sqrt{2}\, x_1x_2, \sqrt{2}\, x_1, \sqrt{2}\, x_2)$ (D) $\phi(x) = (1, x_1^2, x_2^2, \sqrt{2}\, x_1x_2, x_1, x_2)$

3(b). Multiple kernels can be combined to produce new kernels. For example, $K(x, z) = K_1(x, z) + K_2(x, z)$ is a valid combination. Suppose kernel K_1 has the associated feature transformation ϕ_1 and K_2 has the associated feature transformation ϕ_2. What is the feature transform associated with the combinations given below?

i. $K(x, z) = \alpha K_1(x, z)$

D

(A) $\phi(x) = \phi_1(x)$ (B) $\phi(x) = \alpha^2\phi_1(x)$

(C) $\phi(x) = \alpha\phi_1(x)$ (D) $\phi(x) = \sqrt{\alpha}\phi_1(x)$

ii. $K(x, z) = \alpha K_1(x, z) + \beta K_2(x, z)$

(A) $\phi(x) = \alpha\, \phi_1(x) + \beta\, \phi_2(x)$ (B) $\phi(x) = \sqrt{\alpha}\, \phi_1(x) + \sqrt{\beta}\, \phi_2(x)$ D

(C) $\phi(x) = [\alpha\, \phi_1(x), \beta\, \phi_2(x)]$ (D) $\phi(x) = [\sqrt{\alpha}\, \phi_1(x), \sqrt{\beta}\, \phi_2(x)]$

4. We are given the following four data points in two dimension: $X_1 = (2, 2)$, $X_2 = (8, 6)$, $X_3 = (6, 8)$, $X_4 = (2, 4)$. We want to cluster the data points into two clusters C_1 and C_2 using the K-Means algorithm. Manhattan distance is used for clustering. To initialize the algorithm we consider $C_1 = \{X_1, X_3\}$, and $C_2 = \{X_2, X_4\}$. After two iteration of the K-Means algorithm, the cluster memberships are –

$C_1 = \{X_1, X_4\}$, and $C_2 = \{X_2, X_3\}$.

5. We would like to cluster the natural numbers from *1* to *1024* into two clusters using hierarchical agglomerative clustering. We will use Euclidean distance as our distance measure. We break ties by merging the clusters in which the lowest natural number resides. For example, if the distance between clusters *A* and *B* is the same as the distance between clusters *C* and *D*, we would choose *A* and *B* as the next clusters to merge if $min\{A, B\} < min\{C, D\}$, where $\{A, B\}$ are the set of natural numbers assigned to clusters *A* and *B*. For each of the clustering methods mentioned below, specify the number of elements assigned to each of the two clusters obtained by cutting the dendogram at the root.

i. Single linkage: 1023 , 1

ii. Complete linkage: 512 , 512

iii. Average linkage: 512 , 512

6. In a course the probability that a student gets a grade "A" is $P(A) = ½$, a "B" grade is $P(B) = \mu$, a grade "C" is $P(C) = 2\mu$, and a grade "D" is $P(D) = ½ - 3\mu$. We are told that *c* students get "C" and *d* students get "D". We do not know how many students got exactly an "A" or exactly a "B". But we do know that *h* students got either "A" or "B", i.e., $a + b = h$. Our goal is to use the Expectation Maximization algorithm to obtain an estimate of μ.

i. Expectation step: Which formula compute the expected value of *a* and *b* given μ? B

(A) $\hat{a} = \frac{\frac{1}{2}}{\frac{1}{2}+h}\mu$ $\hat{b} = \frac{\mu}{\frac{1}{2}+h}\mu$ (B) $\hat{a} = \frac{\frac{1}{2}}{\frac{1}{2}+\mu}h$ $\hat{b} = \frac{\mu}{\frac{1}{2}+\mu}h$

(C) $\hat{a} = \frac{\mu}{\frac{1}{2}+\mu}h$ $\hat{b} = \frac{\frac{1}{2}}{\frac{1}{2}+\mu}h$ (D) $\hat{a} = \frac{\frac{1}{2}}{1+\frac{\mu}{2}}h$ $\hat{b} = \frac{\mu}{1+\frac{\mu}{2}}h$

ii. Maximization step: Given the expected values of *a* and *b*, which formula computes the maximum likelihood estimate of μ?

A

(A) $\hat{\mu} = \frac{h-a+c}{6(h-a+c+d)}$

(B) $\hat{\mu} = \frac{h-a+d}{6(h-2a-d)}$

(C) $\hat{\mu} = \frac{h-a}{6(h-2a+c)}$

(D) $\hat{\mu} = \frac{2(h-a)}{3(h-a+c+d)}$

7. Given three data points in two-dimensional space: *(1, 1)*, *(2, 2),* and *(3, 3).*

i. What is the first principal component?

$(1/\sqrt{2}, 1/\sqrt{2})$

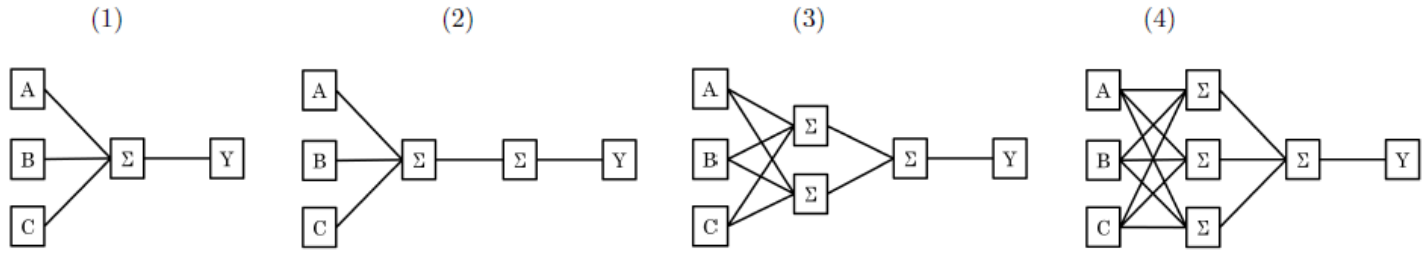
ii. If we want to project the original data points into one dimensional space using the principal component, what is the variance of the projected data?

4/3 = 1.33 or 2

iii. For the projected data above, now, if we represent them in the original two-dimensional space, what is the reconstruction error?

0

8(a). In the following questions i-iv, mark ALL neural networks (1/2/3/4) that can compute the same function as the b6oolean expression mentioned. If none of the neural nets can do this, mark None. Booleans will take values 0, 1, and each perceptron will output values 0, 1. You may assume that each perceptron also has as input a bias feature that always takes the value 1. Connection weights are allowed to take on any values. It may help to write out the truth table for each expression.



i. *A*

1, 2, 3, 4

ii. *A OR B*

1, 2, 3, 4

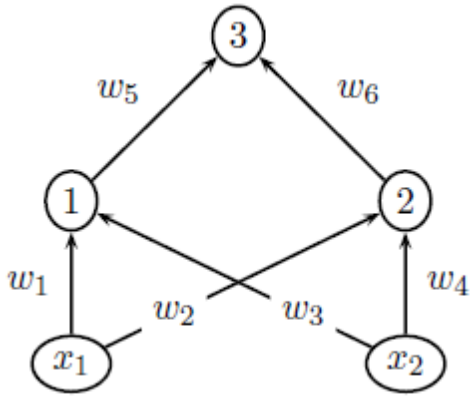
iii. *B XOR C*

3, 4

iv. *(A XOR B) XOR C*

4

8(b). Consider the neural network shown below:



Assume that all the internal nodes and the output nodes use the *tanh* activation function. Note that derivative of $tanh(x) = 1 - tanh^2(x)$. Backpropagation is applied on this network to minimize the squared error. Let o_1, o_2, and o_3 be the output of the neurons 1, 2, and 3 respectively, and x_1, x_2 be the input values. Let δ_1, δ_2, and δ_3 be the values backpropagated by the neurons 1, 2 and 3 respectively. Complete the three expressions below:

i. $\delta_3 = (1-o_3^2)$ x	$(y-o_3)$
ii. $\delta_2 = (1-o_2^2)\ \delta_3$ x	w_6
iii. $\delta_1 = (1-o_1^2)\ \delta_3$ x	w_5

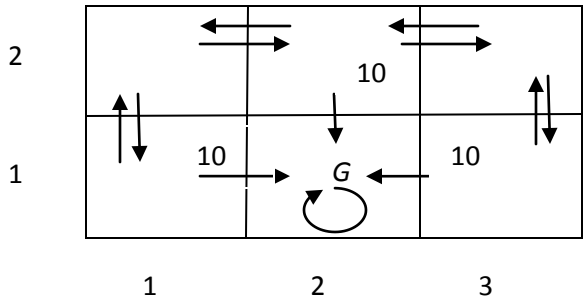
9. Imagine that you are given the following set of training examples (E1-E5). All the features are Boolean-valued.

	F1	F2	F3	Class
E1:	T	T	F	+
E2:	F	T	T	+
E3:	T	F	T	-
E4:	F	T	F	-
E5:	F	F	T	-

Assume that we are using a very weak learner within the AdaBoost algorithm. This simplistic learner simply chooses for its learned model the *lowest-numbered feature that has not yet been used*. Its only intelligent aspect is that it decides whether or not to *negate* this feature, depending on which option works best. I.e., the first time called it will return either *F1* or *NOT(F1)* as its model. We perform two rounds of AdaBoost. In each round five examples are used for training. In the first round the original data set is used. The second round training examples are obtained by re-sampling the original data using the model learned in first round. What are the training examples that likely to be boosted for the third round?

E3/E4

10. Consider the deterministic grid world shown below with the absorbing goal-state *G*. Here the immediate rewards are *10* for the labeled transitions and *0* for all unlabelled transitions. Use discount factor $\gamma = 0.8$.



i. What is the value of V^* for the state *(1, 2)*?

8 (0)

ii. Consider applying Q-learning to this grid world. The Q-table values are all initialized to 0's. Assume that the agent begins in the bottom left grid square and then travels clockwise along the perimeter of the grid until it reaches the absorbing goal state, completing a training episode. Specify which Q values are updated as a result of two such episodes, and give their revised values.

Q((3,1),L) = 10, Q((3,2),D) = 8 (Q((1,3),L) = 10, Q((2,3),D) = 8)

---- END ----

Rough Work

Rough Work

Rough Work

Rough Work

Rough Work