PES University, Bengaluru

(Established under Karnataka Act No. 16 of 2013)

UE18CS303

OCTOBER 2020: IN SEMESTERASSESSMENT B Tech CSE5th SEMESTER TEST – 1

UE18CS303 –MACHINE INTELLIGENCE

Time: 2 Hrs	Answer All Questions	Max Marks: 60

1.	a)	1. What is the difference between supervised and unsupervised learning? Explain with examples *(1+1)* 2. What do model based learning algorithms search for? (1) 3. What is the most common strategy they use to succeed? (1) 4. How do model based algorithms make predictions? (1)	5
	b)	Given the table below predict TPR, TNR, FPR, FNR for thresholds of 0.6, 0.7 and 0.8 for all values and draw the ROC-AUC curve.	5

ID	Actual	Prediction Probability
1	0	0.98
2	1	0.67
3	1	0.58
4	0	0.78
5	1	0.85
6	0	0.86
7	0	0.79
8	0	0.89
9	1	0.82
10	0	0.86

2.	a)	Given a feature set in which one feature can take 3 distinct values while the other 5 features can take 2 distinct values how many syntactically and semantically different concepts can be learnt? 1+1	2
	b)	Prove that under a choice of P(h) and P(D|h) every consistent hypothesis has a posterior probability of $1/\|VS_{H,D}\|$ given that the symbols of P(h) and P(D|h) and $1/\|VS_{H,D}\|$ have their usual meanings as used in the machine learning literature	3
	c)	The following table contains training examples that help predict whether a patient is likely to have a heart attack.	5

PID	CHEST PAIN	MALE	SMOKES	EXERCISE	HEART ATTACK
1	Yes	Yes	No	Yes	Yes
2	Yes	Yes	Yes	No	Yes
3	No	No	Yes	No	Yes
4	No	Yes	No	Yes	No
5	Yes	No	Yes	Yes	Yes
6	No	Yes	Yes	Yes	No

Use information theory to construct a minimal decision tree that predicts whether or not a patient is likely to have a heart attack. **Show each step of the computation.**

3. a) 3

1. If a Decision Tree is overfitting the training set, is it a good idea to try decreasing the maximum depth of the tree? (1)
2. If a Decision Tree is underfitting the training set, is it a good idea to try scaling the input features?(1)
3. If it takes one hour to train a Decision Tree on a training set containing 1 million instances, roughly how much time will it take to train another Decision Tree on a training set containing 10 million instances?(1)

b) 7

Consider a dataset with 3 points in 1-D:

(class)	x
+	0
–	−1
–	+1

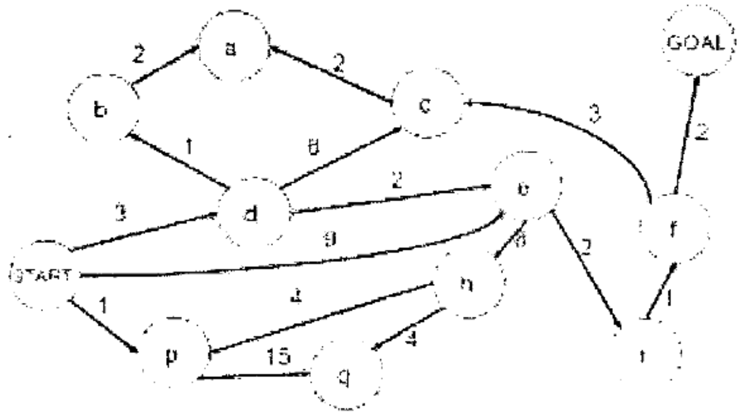
Are these linearly seperable? (1)

Consider mapping each point to 3-D using new feature vectors $\varphi(x) = [1, \sqrt{}2x, x^2]^T$.

Are the classes now linearly separable? If so, find a separating hyperplane. (1)

4. a) 2

Perform a Uniform Cost Search where A is the start state and G and C are both perfectly viable goal states. **Show all steps** and the final backtracking path.



b) 2

Suppose you have an MLP composed of one input layer with 10 passthrough neurons, followed by one hidden layer with 50 artificial neurons, and finally one output layer with 3 artificial neurons. All artificial neurons use the ReLU activation function.

1. What is the shape of the input matrix X? (0.5)
2. What are the shapes of the hidden layer's weight vector W and its bias vector b? (0.5)
3. What are the shapes of the output layer's weight vector W and its bias vector b ? (0.5)
4. What is the shape of the network's output matrix Y? (0.5)

c) 2

1. Is it OK to initialize all the weights to the same value as long as that value is selected randomly using Xavierinitialization? (0.5)
2. Is it OK to initialize the bias terms to 0? (0.5)
3. Name two ways you can produce a sparse neural network model (i.e. network where most of the weights are zero) (0.5)
4. Name two ways that can help you in climbing out of a local minimum in gradient descent in neural networks (0.5)

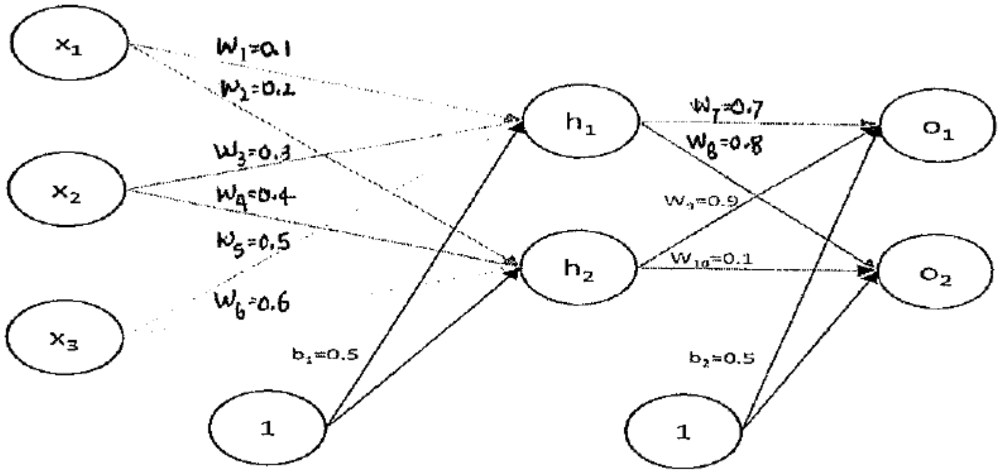
d) Given below is a neural network with one input layer, one hidden layer and one output layer. Assume the activation function used everywhere is **sigmoid** and error function is **MSE.** 4

Input values of **x1, x2, x3** are **1,4 and 5**

Target valuest**1** and **t2** are **0.1 and 0.05**

Forward propagate through the network to get the output values, and calculate the error and backpropagate to find the following error derivatives.

$$\frac{dE}{dw_7} \quad \frac{dE}{dw_9} \quad \frac{dE}{dw_2}$$



5. a) A certain point X_i has 7 nearest neighbor with the following class set={A,A,A,A,B,B,B} and the decision boundary is derived with k=7. This is a weighted KNN algorithm with the weights being inversely proportional to class frequency. The class frequency of A is 95% and B is 5%. 2

Choose the label that X_i needs to be assigned to and provide the calculations.

b) Derive that for all correctly classified examples and all incorrectly classified examples that the sum of their respective weights in the weight updation rule in Adaboost must add up to 0.5 i.e. 3

$$\sum_{correct} w_i^{s+1} = \frac{1}{2}\cdot\frac{\sum_{correct} w_i^s}{1-E^s} = \frac{1}{2}$$

$$\sum_{incorrect} w_i^{s+1} = \frac{1}{2}\cdot\frac{\sum_{incorrect} w_i^s}{E^s} = \frac{1}{2}$$

c) State as the following statements being true or false (Please write the statement and then answer True of False) 2

1. Weak Classifiers in bagging prevents overfitting
2. Bagging increases the variance of a classifier
3. Using Sampling with replacement prevents overfitting
4. Bagging can help robust classifiers from unstable classifier

d) I. Which of the following is / are true about weak learners used in ensemble models? 3

A) They have low variance and they don't usually overfit

B) They have high bias, so they cannot solve hard learning problems

C) They have high variance and they don't usually overfit (1)

2. Suppose, you are working on a binary classification problem and there are 3 models each with 70% accuracy. If you want to ensemble these models using majority voting method. What will be the **maximum** accuracy you can get? Will the **minimum** accuracy always be 70% or above? (1)

3. Which among the following is/are some of the assumptions made by the k-means algorithm (assuming Euclidean distance measure)? (a) Clusters are spherical in shape (b) Clusters are of similar sizes (1)

6 a) 6

Define an HMM H with three states {A, B, C} and alphabet {0, 1, 2}. The initial stable probabilities are πA = 1 and πB = πC = 0. The transition and emission probabilities are as follows:

	A	B	C	0	1	2
A	0.2	0.8	0.0	0.8	0.2	0.0
B	0.0	0.8	0.2	0.0	0.6	0.4
C	0.4	0.0	0.6	0.2	0.0	0.8

1. Draw the state diagram of this HMM and show the transition probabilities. (2)
2. Give all state paths with non-zero probability for the sequence O = 0, 1, 2, 0 (2)
3. What is P(O)? (Use the brute force approach) (2)

b) You are given the following five training instances 2

- x1 = 2, x2 = 1, y = 4
- x1 = 6, x2 = 3, y = 2
- x1 = 2, x2 = 5, y = 2
- x1 = 6, x2 = 7, y = 3
- x1 = 10, x2 = 7, y = 3

We want to model this function using the K-nearest neighbor regressor model. When we want

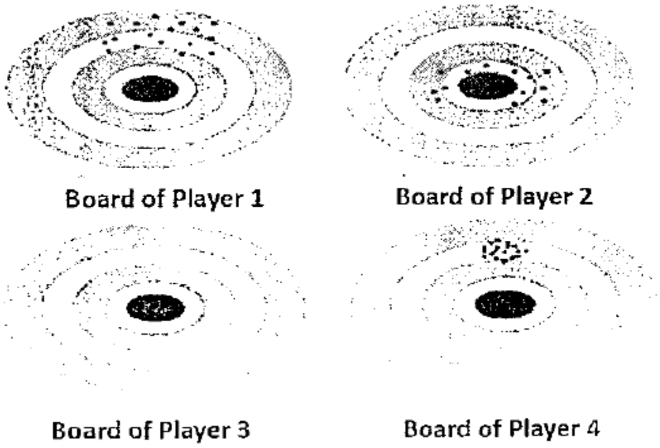
to predict the value of y corresponding to (x1, x2) = (3, 6)

Find the values for k = 2 and k=3

c) Bias and Variance can be visualized using a classic example of a dart game. We can think of the true value of the parameters as the bull's-eye on a target, and the arrow's value as the estimated value from each sample. Consider the following situations of four players, and select the correct option(s) 2

(a) Player 1 has low variance compared to player 4

(b) Player 1 has higher variance compared to player 4



Board of Player 1

Board of Player 2

Board of Player 3

Board of Player 4