MANIPAL INSTITUTE OF TECHNOLOGY
MANIPAL
(A constituent unit of MAHE, Manipal)

VII SEMESTER B.TECH. (Mechatronics)

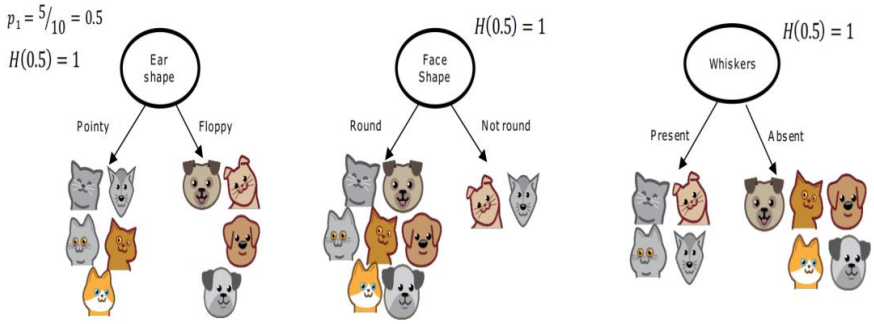
End Sem Examination

SUBJECT: Machine Learning [MTE 4073]
Date: 9/12/2023

Time: 3 Hour Exam time: 2:30PM-05:30PM **MAX. MARKS: 50**

Instructions to Candidates:

- Answer ALL the questions.
- Missing data may be suitably assumed and justified.

Q. No	Question	M	CO	PO	LO	BL
1A	Consider the cat and dog dataset to take a decision for tree learning algorithm to classify as cat or not as shown in the Fig 1A.  Fig 1A: Decision trees with different root nodes There are few training examples at the root note, comprising cat and dog. If the algorithm can choose from among four features, resulting in four corresponding splits, which would you choose based on entropy criteria.	4	4	2	2	4
1B	Consider the classification of flower dataset for predicting the color based on 2 features, i.e, brightness, and saturation with class as colors red or blue as shown in the Table 1B. Classify the color with brightness of 50 and saturation of 80 using KNN classifier. Choose the best K value for prediction. Justify the selection of K value for given dataset. **Brightness** / **Saturation** / **Class** 40 / 20 / Red 50 / 50 / Blue 60 / 90 / Blue 10 / 25 / Red 70 / 70 / Blue 60 / 10 / Red 25 / 80 / Blue Table 1B: flower dataset for color prediction	4	4	2	2	4
1C	The Pima Indians Diabetes Dataset involves predicting the onset of diabetes within 5 years in Pima Indians given medical details. It is a binary (2-class) classification problem. Consider a Indians_diabetics dataset for classification using logistic regression method and obtained the accuracy of 77%. Explore and explain the effect	2	3	2	2	4

	of regularization method, and use of stochastic gradient descent algorithm to accumulate updates for each epoch.					
2A	Assume that you are working in the factory that produces wind turbines. Consider a set of features such as temperature, wind, energy, etc. You are asked to test the system whether it is faulty or not. Model the anomaly detection algorithm by fitting Gaussian distribution for the given data and how the threshold parameter is selected for the optimal result.	4	4	2	2	3
2B	A new virus is affecting the population. People who have the virus will normally have specific symptoms such as a cough and the loss of the sense of taste and/or smell. It is estimated that 1 in 5 of people who suffer these symptoms have the virus and 1 in 2000 people without these symptoms have the virus. A test for the virus has the following accuracy using SVM classifier. For people with symptoms, the true positive rate is 90% and the false positive rate is 5%. For people without symptoms, the true positive rate is 80% and the false positive rate is 1%. Construct the confusion matrix, discuss accuracy, F1 score, precision and recall of given classifier.	4	1	1	1	3
2C	Consider a wine quality test case study that employs linear regression to predict the quality. Explore and explain the effect of tuning the learning rate, number of epochs on given data.	2	3	2	2	4
3A	You are asked to collect the dataset of people performing yoga for classification of yoga poses. What are the ethical concerns that you need to pay attention to before you start your project (provide 4 points).	4	5	8	8	3

3B	You work as a lead data scientist for a bio-sciences company and under your supervision you have a junior Machine Learning engineer. You asked them to develop a model for recognizing 10 different classes of bacteria from 28x28 RGB images. The ML engineer came back to you with their proposed architecture, shown in Fig 3B. Is there anything wrong with it? Provide the reasoning behind your answer.
	
	`class MyNet(torch.nn.Module):`
	`  def __init__(self, num_inputs, C1, C2, num_outputs):`
	`  super(MyNet, self).__init__()`
	`  self.num_inputs = num_inputs`
	`  self.num_outputs = num_outputs`
	`  if C1 != C2:`
	`  self.expand_channels = nn.Conv2d(C1, C2, 1)`
	`  self.relu = nn.ReLU()`
	`  self.stem = nn.Conv2d(num_inputs, C1, kernel_size = 5, padding = 2)`
	`  self.bn1 = nn.BatchNorm2d(C1)`
	`  self.conv1 = nn.Conv2d(C1, C1, kernel_size = 5, padding = 2)`
	`  self.sigmoid = nn.Sigmoid()`
	`  self.bn2 = nn.BatchNorm2d(C1)`
	`  self.conv2 = nn.Conv2d(C1, C2, kernel_size = 5, padding = 2)`
	`  self.conv_last = nn.Conv2d(C2, C2, kernel_size = 3, stride = 4)`
	`  self.fl = nn.Flatten()`
	`  self.lin = nn.Linear(784, num_outputs)`
	`  def forward(self, x):`
	`  x = self.stem(x)`
	`  identity = x`
	`  x = self.bn1(x)`
	`  x = self.relu(x)`
	`  x = self.conv1(x)`
	`  x = self.bn2(x)`
	`  x = self.relu(x)`
	`  x = self.conv2(x)`
	`  if self.expand_channels is not None:`
	`  identity = self.expand_channels(identity)`
	`  x = identity`
	`  x = self.relu(x)`
	`  out = self.conv_last(x)`
	`  out = self.fl(out)`
	`  out = self.lin(out)`
	`  return out`
	
	Fig 3B: sample pytorch code for classification of images
		4	3	3	5	4

3C	The Sonar Dataset involves the prediction of whether or not an object is a mine or a rock given the strength of sonar returns at different angles. It is a binary (2-class) classification problem. A sonar dataset is used using perceptron algorithm and obtained accuracy of 73%. Explain the effect of data normalisation, and learning rate in the accuracy of the algorithm.	2	3	2	2	4
4A	The set of input training vectors to a perceptron is as follows: $X_1 = \begin{bmatrix} 1 \\ -2 \end{bmatrix} \quad X_2 = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$ The initial weight vector W1 is assumed to be: $W_1 = \begin{bmatrix} 2 \\ -1 \end{bmatrix}$ The learning constant is assumed to be 0.1. Update the weights based on forward and backward propagation with MSE as a loss function. (Perform one iteration)	4	3	1	1	3
4B	Assume that you are working as a datascience engineer and asked to develop a model for selecting the best product. It is important that your company has to make a profit and can recognize interest of customers. You have been given the dataset of people who purchased similar items and rated the product. Which method you would adopt to find the best product for the customer. Justify the answer with mathematical proof.	4	4	2	2	5
4C	Construct the confusion matrix for the data given below: actual predicted 0 0 0 1 0 1 0 0 0 0 1 1 1 0 1 1 1 1 1 1 Table 4C: Output of a classifier and ground truth data	2	1	1	1	3
5A	Construct a bagged decision trees to make an ensemble predictions. Create 3 decision trees from training data and split points as Model1 : X1 ≤ 5:38660215 Model2:X1≤4:090032824 Model3 : X2 ≤ 0:925340325 X1 X2 Y 2.309572387 1.168959634 0 1.500958319 2.535482186 0 3.107545266 2.162569456 0 4.090032824 3.123409313 0 5.38660215 2.109488166 0 6.451823468 0.242952387 1 6.633669528 2.749508563 1 8.749958452 2.676022211 1 4.589131161 0.925340325 1 6.619322828 3.831050828 1 Table 5A: Dataset	4	4	2	2	5
5B	Discuss the AdaBoost classifier by taking an example of face detection algorithm.	4	4	2	2	4
5C	Discuss the ethical concerns in big data collection with a case sample.	2	5	8	8	4