MID Semester Examination (CS 361)

1st March 2023

Time: 120 Minutes

Full Marks: 50

No clarifications of the question will be given at the examination hall. In case of any doubt, write your assumption and answer the question accordingly. **Please submit the question paper along with the answer sheet.**

Attempt all questions.

1. Suppose you have a dataset of emails classified as spam or not spam. The dataset contains the following features:
   - F1: The email contains the word "buy"
   - F2: The email contains the word "discount"
   - F3: The email contains the word "limited time offer"
   - F4: The email contains the word "urgent"
   - F5: The email contains the word "moneyback guarantee"

   You want to use the Naive Bayes classifier to predict whether a new email is a spam or not spam based on these features. You have the following training data (refer to **Table 1**):

Email	F1	F2	F3	F4	F5	Spam
E1	1	0	1	0	0	YES
E2	1	1	1	1	1	YES
E3	0	1	0	1	1	YES
E4	0	1	1	0	0	NO
E5	1	0	0	1	1	NO

Table 1

   Find whether a new email with the features: F1=1, F2=0, F3=1, F4=0, F5=0 is spam or not spam. **[10]**

2. Answer all the questions.
   a. Consider a univariate linear regression model. Which of the following(s) is/are true?
      i. Changing the input variable by 1 unit always affects the output by 1 unit too.
      ii. Considering Mean Squared Error to compute the loss is a good idea as it reduces the effect of outliers.
      iii. Since it is univariate, we need to estimate one coefficient for modeling the data.
      iv. None of the above.
   b. Mention two distance matrices (One for continuous variable and one for categorical variable) used in KNN Algorithm.
   c. Which of the following statement(s) is/are TRUE about KNN?
      i. KNN can be used in both classification and regression
      ii. KNN algorithm cannot be used for assigning missing values of categorical and continuous variables.
      iii. When you increase the k, the bias will decrease, and variance will increase
      iv. KNN algorithm does more computation on test time rather than training time.

d. Two attributes have been considered to classify whether a fruit is an apple or not apple. The following table (refer to **Table 2**) shows the set of attributes and class belongingness for four data samples.

A1	A2	Y = Classification
8	8	Not Apple
8	5	Not Apple
3	5	Apple
2	5	Apple

Table 2

Now, a new unknown fruit having attributes (A1 = 3 and A2 = 7) is given. Find the class belongingness of the new fruit (whether Apple or not) using the KNN classification method using k=3. Justify your answer.

e. When we query a node in a Bayesian network, the result is often referred to as the marginal. What do you mean by 'marginal"?

f. Which of the following statement(s) is/are FALSE

i. A Bayesian network is a factorized representation of the full joint distribution.

ii. For a discrete Bayesian network with n variables, the amount of space required to store the "joint" distribution table is O(n).

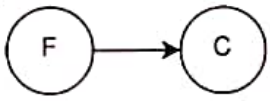
iii. While creating a Bayesian network, a node, and its predecessors are conditionally dependent.

iv. Bayesian learning requires the computation of the posterior distribution over the model parameters, which can be computationally expensive.

v. In a Bayesian Network, the Conditional Probability Table is the Local Probability Distribution at each node.

vi. Bayesian Learning is Unsupervised Learning.

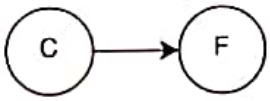
g. Consider the following Bayesian network, where F = having Fever and C = Infected with Corona:



$P(F) = 0.1 \quad P(C|F) = 0.8 \quad P(C|\neg F) = 0.4$

i. Write down the joint probability table specified by the Bayesian network.

ii. Determine the probabilities for the following Bayesian network so that it specifies the same joint probabilities as in the given one.



[1 + 1 + 1 + 3 +1 + 1 + (1 + 1) = 10]

3. Consider **Table 3**, which consists of land prices based on the area and the proximity to the city center. We want to build a linear regression estimator on these data. We will use the Mean Square Error Cost function.

a. Write the hypothesis function and the equations for updating the different parameters of this estimator.

b. Use the training data (first four rows of **Table 3**) to calculate the parameters after one iteration of gradient descent. Take the initial values of the parameters as 0.5 each, and the learning rate as 0.001. Show your steps.

c. Use the trained model to predict the price of the unknown data point (last row of **Table 3**) to the nearest million Rupees.

Sr. No.	Area of Land (m^2)	Dist. to City center (Km)	Price (million INR)
1	25	12	13
2	30	10	18
3	21	25	9
4	28	2	17
5	**35**	**15**	**?**

Table 3 **[4 + 5 + 1 = 10]**

4. Choose the correct alternatives:
   a. When a decision tree is grown to full depth, it is more likely to fit the noise in the data.
      [TRUE/FALSE/CAN'T SAY]
   b. When the feature space is larger, overfitting is more likely.
      [TRUE/FALSE/CAN'T SAY]
   c. When the hypothesis space is richer, overfitting is more likely.
      [TRUE/FALSE/CAN'T SAY]
   d. Assume that we try to fit a linear and 8th-degree polynomial to data distribution coming from a cubic function corrupted by standard Gaussian noise. Let M1 and M2 denote the models corresponding to the linear and 8th-degree polynomials. Then
      i. Bias(M1) ≤ Bias(M2), Variance(M1) ≤ Variance(M2)
      ii. Bias(M1) ≤ Bias(M2), Variance(M1) ≥ Variance(M2)
      iii. Bias(M1) ≥ Bias(M2), Variance(M1) ≤ Variance(M2)
      iv. Bias(M1) ≥ Bias(M2), Variance(M1) ≥ Variance(M2).
   e. What is the use of regularization parameters while performing a regularized linear regression (RLR)?
      i. Until some point, increasing it reduces the variance of the model significantly without significant addition of bias to the model.
      ii. It reduces the bias in the model and hence reduces overfitting.
      iii. Controls the trade-off between the need for the model to fit the training set well and also having a large number of model parameters.
      iv. Helps to find the exact decision boundary regardless of its complexity.
      v. Consider a self-driving car that learns an RLR model X that gives the best driving performance based on 10 attributes – e.g., road curvature, steering angle, and speed. After a month, you get data about 15 more attributes like weather, driver experience, and the car model - and incorporate them into X. Suppose the re-trained RLR model is Y. A high regularization parameter increases the inability of Y to capture the true relationship between the 25 attributes in the dataset. **[1 + 1 + 1 + 1 + 1 = 5]**

5. In the case of the k-NN classifier, how increase of feature dimensionality affects the classification performance? Justify your answer. **[1 + 4 = 5]**

6. A few students from your school won a lucky contest and are visiting the Harry Potter Studio in England where you can get your personality attributes tested by the "sorting hat" and get placed into either the *Gryffindor (G)* or *Slytherin (S)* house. Each attribute can take a value Low (L), Medium (M) 4or High (H).

Player ID	Compassion	Emotional stability	Self-discipline	Ambition	House
1	L	M	M	M	S
2	M	L	M	L	S
3	L	H	L	H	S
4	H	H	M	M	G
5	H	L	L	H	S
6	L	M	L	L	S
7	M	H	L	H	G
8	M	L	L	L	S
9	M	H	H	L	G
10	L	L	L	H	S
11	M	M	M	H	G
12	H	H	L	L	S
13	M	H	M	H	G

a. The "sorting hat" uses a **CART algorithm**. Considering only the attributes *Emotional stability* and *Ambition*, find the reduction in impurity for these attributes and finally state which attribute should be selected as a splitting attribute (among these two attributes only), showing what are the attribute values in each branch of the splitting attribute.

b. When choosing one feature from X1……, Xn (with class label denoted with Y) while building a Decision Tree, which of the following criteria is the most appropriate to maximize? (H() = entropy, P() = probability)

i. $P(Y|X_j)$
ii. $P(Y) - P(Y|X_j)$
iii. $H(Y) - H(Y|X_j)$
iv. $H(Y|X_j)$
v. $H(Y) - P(Y)$

c. Assume two decision trees are trained on the same data. A tree with a depth of 3 has _____ (higher/lower) variance and _____ (higher/lower) bias than a tree with a depth of 1.

d. Assume we have two equal vectors X and Z in our training set (that is, all attributes of X and Z including the labels are exactly the same). Can removing Z from our training data change the decision tree we learn for this dataset? Explain briefly. **[7 + 1 + 1 + 1 = 10]**