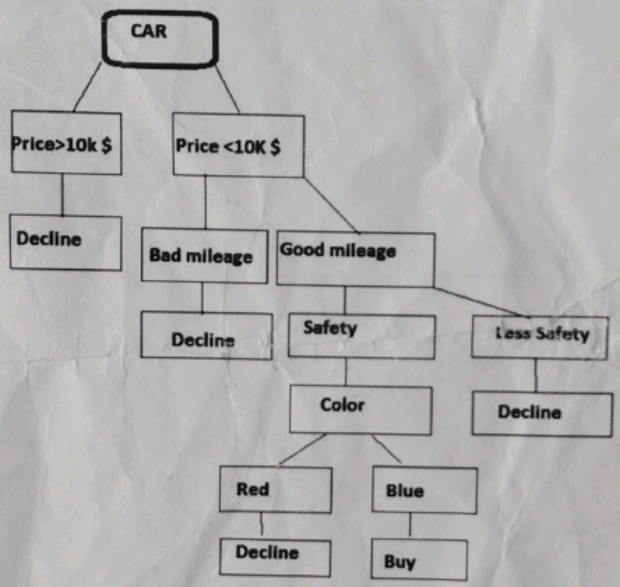
Department of Computer Science and Engineering
Indian Institute of Technology Jodhpur
CSL7620 - Machine Learning

October 3, 2023

Instructions:
1. Read the questions carefully.
2. If you find anything unclear/incorrect in any question, make a reasonable assumption and proceed.

Time: 45 min | *Quiz-2* | Maximum Marks: 10

1. Consider the decision tree shown below. [2]



Will the following car(s) be bought or declined and why?

1. A safe, blue car having good mileage with price 11K$.
2. A safe, red car having good mileage with price 9.5K$.

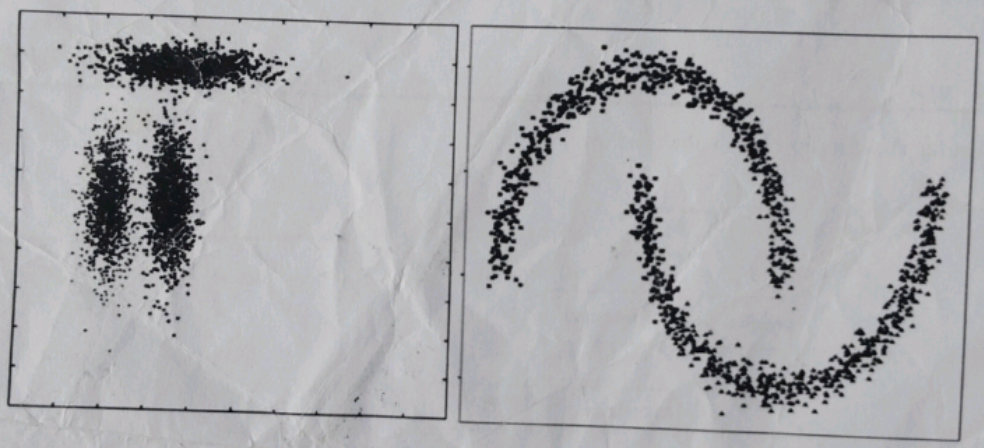
2. If the data matrix X contains samples arranged in columns, as below, show that the principal components A are given by $A = UX$ where U is the matrix of left singular vectors of X. [2]

$$X = \begin{bmatrix} | & | & \dots & | \\ x^{(1)} & x^{(2)} & \dots & x^{(n)} \\ | & | & \dots & | \end{bmatrix}_{d \times n}$$

3. [True/False] The best way to initialize the parameters of GMM is random initialization. Justify your answer. [2]

4. You have been asked to perform K-means clustering in your second assignment using cosine similarity instead of the Euclidean distance. What are the changes you need to do in your objective function and parameter estimation in this case? [2]

5. Consider the two sets of 2D data points shown below. Our objective is to divide the data points into different clusters. Given that the data does not contain any noise, which clustering technique among K-means, DBSCAN, and GMM would be the best choice for the given sets and why? [2]



Department of Computer Science and Engineering
Indian Institute of Technology Jodhpur

CSL7620 - Machine Learning

September 8, 2023

Instructions:
1. Read the questions carefully.
2. All questions are mandatory.
3. If a question requires justification, zero mark will be awarded in absence of the justification.
4. If you find anything unclear/incorrect in any question, make a reasonable assumption and proceed.

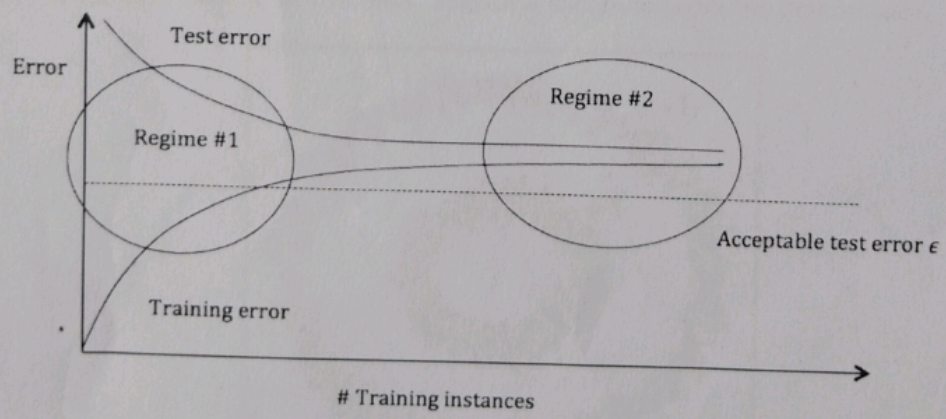
Time: 1 hour | *Minor-I* | Maximum Marks: 20

1. [True/False] l_2 regularization in polynomial regression is useful for parameter selection. Justify your answer. [2]

2. [True/False] GDA is more robust and less sensitive to incorrect modeling assumptions as compared to logistic regression. Justify your answer. [2]

3. Consider a simple function $f(x, y, z) = q(x, y) \times z$, where $q(x, y) = x + y$. Now let us assume that we are evaluating this function at $x = -2, y = 5,$ and $z = -4$. In addition let the value of the upstream gradient (gradient of the loss with respect to our function, $\frac{\partial L}{\partial f}$) is equal to 1. If we use gradient descent to update $x, y,$ and z with a learning rate of 0.01, find out the values of the parameters after each of those is updated once. [2]

4. Consider the error dynamics of an ML model shown below. Which of the two regimes marked in the figure correspond to high bias or high variance? [2]



5. A 10 class classification task is to be solved using Bayes decision rule. Given that the value of reject class classification risk ($R(\alpha_{C+1}|x)$) for an input x is 0.4, consider the posterior [4]

probability vectors corresponding to two input samples and explain which of the two will be classified in the reject class.

1. 0.1 0.01 0.7 0.04 0.02 0.03 0.05 0.01 0.01 0.03
2. 0.5 0.01 0.02 0.1 0.07 0.05 0.05 0.1 0.04 0.06

6. Show that for a binary classification task the maximum likelihood estimate for covariance matrix in LDA is the weighted average of the covariance matrices corresponding to the two classes. What is the significance of the associated weights? [4]

7. If we apply naive conditional independence assumption in GDA and also assume that each feature (x_j) has the same variance then show that the decision boundary is orthogonal to the line joining the class-wise means of input samples (μ_0 and μ_1) for a binary classification task. Further if we also assume that the prior probabilities of both the classes are equal then show that the classification can be performed by simply measuring the Euclidean distance from the means. [4]

Department of Computer Science and Engineering
Indian Institute of Technology Jodhpur

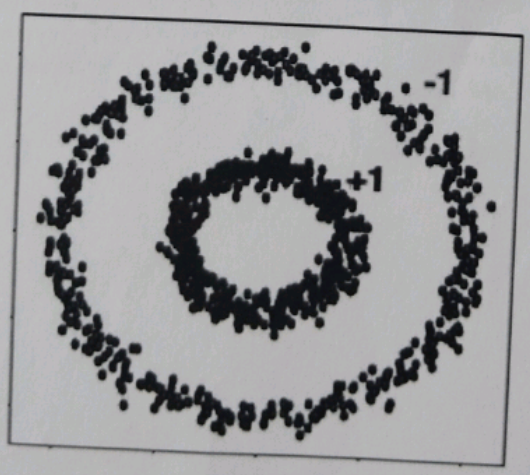
CSL7620 - Machine Learning

October 17, 2023

Instructions:
1. Read the questions carefully.
2. All questions are mandatory.
3. If a question requires justification, zero mark will be awarded in absence of the justification.
4. If you find anything unclear/incorrect in any question, make a reasonable assumption and proceed.

Time: 1 hour — *Minor-II* — Maximum Marks: 20

1. [True/False:] A non-parametric classifier is the one which does not contain any parameter. [1]
2. [True/False:] The average performance of any pair of algorithms across all possible problems is identical. Justify your answer. [1]
3. [True/False:] The main reason behind a better performance of boosting as compared to bagging is the sequential training of base classifiers. Justify your answer. [1]
4. Why the non-Gaussianality assumption is critical for ICA? [1]
5. Show that the first principal components produced by PCA (projections on the axis corresponding to the largest eigenvalue) has largest variance. [2]
6. When do we observe the high variance phenomenon in decision trees? How do we address it? [2]
7. What are the ways to improve the performance of standard k-NN classifier. [2]
8. Consider a binary classification problem where the data points belonging to two classes {+1, -1} appear in form of concentric circles, as shown below. Comment on the performance of Fisher linear discriminant on this data. Suggest a way to improve the performance. [2]



9. Consider the data available on buying a car shown below. Show the construction of a decision tree with details using ID3 algorithm. [8]

Colour	No. of Airbags	Cost	Sunroof	Buy
Red	6	$> 10L$	Yes	Yes
Blue	6	$< 10L$	No	Yes
Red	4	$< 10L$	No	No
Blue	6	$> 10L$	No	No
Black	6	$> 10L$	Yes	Yes
Black	4	$< 10L$	Yes	No
Blue	4	$< 10L$	Yes	No
Black	4	$> 10L$	Yes	No
Red	2	$< 10L$	No	No
Blue	4	$> 10L$	Yes	No

OR

Assume that you have two coins, C1 and C2, where the bias of C1 is θ_1 (*i.e.* probability of getting heads with C1) and the bias of C2 is θ_2. You randomly select (with probability 0.5) one of the coins for 5 times and toss the selected coin 10 times. The following outcome of the coin tosses is observed.

H T T T H H H H H H
T T H H H T H T H H
T T T T H T T T T T
T T H H H H T T T T
T H T H T T H T T T

You decide to use Expectation-Maximization to estimate the values of θ_1 and θ_2. Show the E and M step calculations and corresponding values of θ_1 and θ_2 for two iterations while assuming the initial values as 0.5 and 0.4, respectively.

Department of Computer Science and Engineering
Indian Institute of Technology Jodhpur
CSL7620 - Machine Learning

November 17, 2023

Instructions:
1. Read the questions carefully.
2. If you find anything unclear/incorrect in any question, make a reasonable assumption and proceed.

Time: 45 min | *Quiz-3* | Maximum Marks: 10

1. [True/False] The logistic function ($\sigma(z)$) when used as activation functions in a neural network may lead to the problem of vanishing gradients but the hyperbolic tangent function ($tanh(z)$) will not. Justify your answer. [2]

2. [True/False] For any function $f(x) : R^n \rightarrow R^m$, we can always find a neural network (with 1 hidden layer containing enough neurons) whose output $g(x)$ satisfies $|g(x) - f(x)| < \epsilon$. Justify your answer. [2]

3. Consider the modified perceptron learning algorithm shown on the right. Let us assume that the data points {(-2, 2), (-2, 1), (-1, 1)} have the true label of 1 and {(1, 3), (1, 2), (2, 1)} have the true label of 0. Answer the following. [6]

   1. Given that the $\boldsymbol{w}$ is initialized to $[w_0\ w_1\ w_2]^T = [0\ 1\ -1]^T$, find the value of $\boldsymbol{w}$ after two iterations of **while** loop.
   2. How do we interpret this modified version of the perceptron learning algorithm geometrically?
   3. What are the advantages and disadvantages of the modified algorithm over the conventional perceptron learning algorithm?

$P \leftarrow$ inputs with label 1;
$N \leftarrow$ inputs with label 0;
Initialize $\boldsymbol{w}$ randomly;
$\boldsymbol{a} = \mathbf{0}$;
while ***not*** *convergence* **do**
  $\boldsymbol{a} = \mathbf{0}$;
  for $i = 0$ *to* $|P \cup N|$ **do**
    if $\boldsymbol{x}_i \in P$ *and* $\boldsymbol{w}^T\boldsymbol{x}_i < 0$ **then**
      $\boldsymbol{a} = \boldsymbol{a} + \boldsymbol{x}_i$;
    end if
    if $\boldsymbol{x}_i \in N$ *and* $\boldsymbol{w}^T\boldsymbol{x}_i > 0$ **then**
      $\boldsymbol{a} = \boldsymbol{a}$ - $\boldsymbol{x}_i$;
    end if
  end for
  $\boldsymbol{w} = \boldsymbol{w} + \boldsymbol{a}$;
end while