CSL7620: Machine Learning
Major Examination

September 25, 2024

Time: 3 hrs **Max Marks: 50**

Instructions:

1. This Question paper has two parts Part-A and Part-B. Part-A has MCQ questions with a single correct choice.
2. Part- B has descriptive type questions.Write your answers on A-4 size sheets, and after completing the exam, scan the answer sheets and upload it in PDF format named as Name RollNumber.pdf on Google Classroom.
3. Maximum marks for every question are written in a [ ] at the end of the section.
4. Be very careful and do not answer in a hurry.

1 Part-A

1.1 Multiple Choice Questions [1 mark each]

1. What is overfitting in machine learning?
   (a) When a statistical model describes random error or noise instead of an underlying pattern
   (b) Robots are programmed so that they can perform the task based on data they gather from
   (c) While involving the process of learning, 'overfitting' occurs.
   (d) A set of data is used to discover the potentially predictive relationship

2. Which of the below is not a supervised machine learning algorithm?
   (a) K-means
   (b) Naïve Bayes
   (c) SVM for classification problems
   (d) Decision tree

3. Match the items in Column 1 with the items in Column 2 in the following table given below.

Column 1	**Column 2**
(p) Principal Component Analysis	(i) Discriminative Model
(q) Naïve Bayes Classification	(ii) Dimensionality Reduction
(r) Logistic Regression	(iii) Generative Model

Table 1: Match the items in Column 1 with the items in Column 2

   (a) (p) - (iii), (q) - (i), (r) - (ii)
   (b) (p) - (ii), (q) - (i), (r) - (iii)

(c) (p) - (ii), (q) - (iii), (r) - (i)

(d) (p) - (iii), (q) - (ii), (r) - (i)

4. Which of the following statements is correct regarding bias and variance in machine learning models?

   (a) High bias can lead to underfitting, and high variance can lead to overfitting

   (b) Increasing bias improves model accuracy

   (c) Low bias and low variance lead to better ML models

   (d) Increasing variance improves model generalization

5. Which of the following statements is false about boosting?

   (a) It mainly increases the bias and the variance

   (b) It tries to generate complementary base learners by training the next learner on the mistakes of the previous learners

   (c) It is a technique for solving two-class classification problems

   (d) It uses the mechanism of increasing the weights of misclassified data in preceding classifiers

2 Part-B

2.1 Short Answer Questions [2 marks each]

1. Consider a set of points

$$\begin{pmatrix} 2 \\ 2 \end{pmatrix}, \begin{pmatrix} 3 \\ 1 \end{pmatrix}, \begin{pmatrix} 3 \\ -1 \end{pmatrix}, \begin{pmatrix} -1 \\ 1 \end{pmatrix}, \begin{pmatrix} -2 \\ -2 \end{pmatrix}$$

   embedded in a 2-dimensional space. Our goal is to reduce the dimensionality of these data points to 1 using the PCA algorithm. Find the reduced dimensionality data points and calculate the reconstruction error.

2. Given a dataset with K binary-valued attributes (where $K > 2$) for a two-class classification task, justify that the number of parameters to be estimated for learning a naïve Bayes classifier is $2^K + 1$.

3. Given a hypothesis class $\mathcal{H}$ for a binary classification problem, prove that the $VC(\mathcal{H}) \leq \log_2 |\mathcal{H}|$.

4. Consider a Multi-Layer Perceptron (MLP) model with two hidden layers and one output layer. The first hidden layer has 16 neurons, the second hidden layer has 8 neurons, and the output layer has 4 neurons. The input to the MLP is a 6-dimensional vector. Each neuron is connected to every neuron in the previous layer, and a bias term is included for each neuron. The activation function used is the ReLU function.

5. Let $A \in \mathbb{R}^{n \times n}$ be a matrix and $x, y \in \mathbb{R}^n$ be two vectors. Then, show that $x^T A y = \sum_{i=1}^{n} \sum_{j=1}^{n} x_i y_j a_{ij}$.

3 Long Answer Questions [5 marks each]

1. Consider the following dataset with three features and a binary target variable "Class":

Feature A	Feature B	Feature C	Class
Sunny	Hot	High	0
Sunny	Hot	Normal	1
Overcast	Hot	High	1
Rainy	Mild	High	0
Rainy	Cool	Normal	1
Rainy	Cool	High	1
Overcast	Cool	Normal	1
Sunny	Mild	High	0
Sunny	Cool	Normal	1
Rainy	Mild	Normal	1

   (a) Compute the entropy of the target variable "Class" for the entire dataset.

   (b) Calculate the entropy and information gain for splitting the dataset based on "Feature A".

   (c) Determine the entropy and information gain for splitting the dataset based on "Feature C".

   (d) Compare the information gains obtained from "Feature A" and "Feature C". Which feature provides a better split for the dataset?

2. Consider a dataset consisting of the following 2D data points:

$$\{(2,3),(3,3),(6,5),(8,8),(5,8),(7,7)\}$$

   Apply the K-means clustering algorithm with $K = 2$ clusters.

   (a) Initialize the cluster centroids by selecting (2, 3). Compute the Euclidean distance between each data point and the initial centroids, and assign each data point to the nearest centroid. Calculate the new centroids based on these assignments.

   (b) Repeat the assignment and centroid update steps until convergence is reached. Determine the final cluster assignments and the final centroids.

3. Consider a classification problem to classify the input signal s into one of N classes $\omega \in 1, 2, ..., N$ such that the action $\alpha(s) = i$ means classifying s into class i. The Bayesian decision rule is to maximize the posterior probability

$$\alpha_{Bayes}(s) = \omega^\star = \arg\max_{\omega} \quad p(\omega|s)$$

   Suppose we replace it by a randomized decision rule, which classifies s to class i following the posterior probability $p(\omega = i|s)$, *i.e.,*

$$\alpha_{rand}(s) = \omega \sim p(\omega|s)$$

   (a) What is the average risk R_{rand} for this decision rule? Derive it in terms of the posterior probability using the zero-one loss function.

   (b) Show that this risk R_{rand} is always no smaller than the Bayes risk R_{Bayes}. Thus, we cannot benefit from the randomized decision.

(c) Under what conditions on the posterior, the two decision rules are the same?

4. Derive M-step formulae for updating the covariance matrices and mixing coefficients in a Gaussian mixture model when the responsibilities are updated incrementally, analogous to the result shown in Eq. 4 for updating the means.

$$\boldsymbol{\mu}_k^{\text{new}} = \boldsymbol{\mu}_k^{\text{old}} + \left(\frac{\gamma^{\text{new}}(z_{mk}) - \gamma^{\text{old}}(z_{mk})}{N_k^{\text{new}}} \right) \left(\mathbf{x}_m - \boldsymbol{\mu}_k^{\text{old}} \right) \tag{4}$$

5. Consider a linear model of the form

$$y(\theta, w) = w_0 + \sum_{i=1}^{D} w_i, \theta_i$$

together with a sum-of-squares error function of the form

$$E_D(w) = \frac{1}{2} \sum_{n=1}^{N} (y(\theta_n, w) - t_n)^2$$

Now suppose that Gaussian noise ϵ_i with zero mean and variance σ^2 is added independently to each of the input variables θ_i.By making use of $E[\epsilon_i] = 0$ and $E[\epsilon_i \epsilon_j] = \delta_{ij}\sigma^2$. Estimate the optimal model parameters.

4 Case Study [10 marks]

Vodafone has collected data on its customers and is facing a high turnover rate. The dataset contains the following features for each customer:

(a) Monthly charges

(b) Total minutes of calls

(c) Number of customer service calls

(d) Total data usage (MB)

(e) Contract length (months)

Additionally, the target variable turnover rate indicates whether a customer has left the service (1) or not (0). The company wants to understand patterns in customer behavior and use machine learning to predict turnover, identify customer segments, and optimize retention strategies. Answer the following questions:

(a) Based on the given dataset, identify the appropriate machine learning task(s) that could help the company predict customer turnover and analyze customer segments. Explain the reasoning behind choosing the task(s) (e.g., classification, clustering, dimensionality reduction).

(b) Choose a suitable machine learning algorithm to predict whether a customer will turnover based on the features provided. Justify your choice and describe how the algorithm works in terms of how it handles the input data and makes predictions.

(c) Suppose the dataset contains many features and some of them may not be highly relevant for the task of turnover prediction. Suggest a method to reduce the dimensionality of the dataset while preserving the most important information. Explain how this method will help in improving the model's performance or interpretability.

(d) The company also wants to group its customers into distinct clusters based on their service usage patterns to target marketing strategies more effectively. Suggest a machine learning algorithm for this clustering task. Explain how it works and how the company could interpret the resulting clusters.

(e) After training the turnover prediction model, you find that the model performs well on the training set but poorly on the test set, indicating potential overfitting. Propose solutions to address this issue and explain how they could improve the model's generalization to new, unseen data.