UNIVERSITY OF RUHUNA

Faculty of Engineering

End-Semester 7 Examination in Engineering: August 2024

Module Number: EE7209 **Module Name: Machine Learning (C-18)**

Part-B

[2 Hours and 15 Minutes]

[Answer **all questions**, each question carries **10 marks**]

Q1 a) In Table Q1.a, a part of the "Annual Healthcare Cost" dataset has been provided. Here the Age, Gender, Height (cm), Weight (kg), Smoker? and, Exercising? act as the input features and the Healthcare Cost (LKR) acts as the target variable. Based on the information provided in the Table Q1.a formulate a linear regression function/hypothesis $h(\theta)$ for the given dataset using appropriate notations. Specify all the symbols that you have used.

Table Q1.a

Age	Gender	Height (cm)	Weight (kg)	Smoker?	Exercising?	Healthcare Cost (LKR)
35	Male	158	65	No	Yes	50000
25	Female	150	45	Yes	No	80000
18	Female	148	50	No	No	45000
45	Male	180	75	Yes	No	90000
55	Male	165	80	Yes	Yes	75000

[0.5 Marks]

b) Assume that, under the Exploratory Data Analysis (EDA) you have scatter plotted the height against the weight as shown in Figure Q1. b.

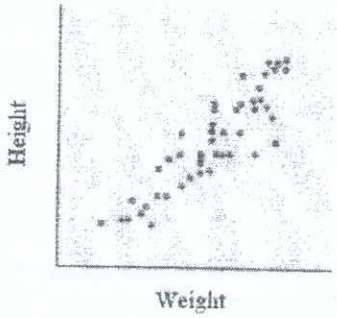

Figure Q1.b

i) What are the assumption(s)/fact(s) you can derive? Modify the linear regression model that you have defined in Q1.a based on the derived assumptions/facts.

ii) Explain the purpose of any changes that you have made Q1.b.i.

[1.5 Marks]

c) In Machine Learning we are using data to learn the best possible functions/hypothesis that accurately predicts the relationship between dependent and independent variables. In regression problems like the one formulated in Q1.a, and Q1.b, we need to minimize a cost function using some optimization algorithm like Gradient Descent.

i) Define the cost function ($J(\boldsymbol{\theta})$) for a linear regression problem based on the Mean Squared Error (MSE). Assume $\boldsymbol{n}$ training examples.

ii) Formulate the gradient descent update rule for a linear regression problem assuming there is a single datapoint/example in the dataset and consider only one parameter ($\boldsymbol{\theta_j}$).

iii) Extend the previously derived update rule when there are $\boldsymbol{n}$ number datapoints/examples in the dataset and more than one parameter.

iv) What is the type of gradient descent formula you have derived in the Q1.c.iii? Mention the other types of gradient descent algorithms and differentiate them from the derived version.

[4.0 Marks]

d) A Machine Learning engineer has decided to design the above regression problem as a classification problem and, in the target variable column (Healthcare Cost) the values less than or equal to 50000 LKR have been mapped to Class 0 and rest has been mapped to Class 1. Furthermore, the data engineer has selected Logistic Regression as the classification algorithm.

i) Transform the hypothesis defined in Q1.b to meet the new requirements.

ii) The cost function of Logistic regression can be mentioned as follows:

$$J(\boldsymbol{\theta}) = \frac{1}{N}\sum_{i=1}^{N}\left(y^{(i)}\log\left(h(x^{(i)})\right) + \left(1 - y^{(i)}\right)\log\left(1 - h(x^{(i)})\right)\right)$$

Formulate the above equation starting from appropriate assumptions on the hypothesis. Assume $\boldsymbol{n}$ training examples in the dataset and mention all the important steps in the derivation.

iii) Explain the behavior of the above cost function under correct classifications and possible misclassifications.

[4.0 Marks]

Q2 a) In Machine Learning, the ultimate goal is to generalize the model to unseen data. Bias and Variance are two types of errors that can influence the performance of Machine Learning models. Therefore, in model generalization, the Bias Variance trade-off plays a vital role.

i) Describe Bias and Variance errors in the context of Machine Learning.

ii) Under what learning conditions do Bias and Variance errors arise? Please specify the conditions for each type.

iii) Explain the bias-variance trade-off in a Machine Learning model using the relationship between model complexity and the total test error.
Hint: Utilize a plot and show the decomposition of the total error in Bias and Variance components.

[2.0 Marks]

b) Decision trees are one of the most popular classification algorithms (also a regression algorithm) used in Machine Learning. The decision trees can be constructed using various algorithms.

i) ID3 (Iterative Dichotomiser 3) is one of the algorithms that can be used to construct a decision tree. State two other decision tree construction algorithms along with the metrics used in each algorithm (including the metric for ID3 algorithm) to perform splits at each node.

ii) Assume that you are a Machine Learning engineer attached to a hospital and, you have been given the dataset in Table Q2.b and you are supposed to construct a decision tree for health risk classification for new patients using ID3 algorithm. Treat the missing values in the given dataset (Here (X), (Y), (Z_ID) are NULL values) and mention your preprocessing strategy for each case).

Table Q2.b

ID	Gender	Blood Sugar	Blood Pressure	Smoker?	Taking Medicine	Risk
1	Male	High	High	Yes	(Z_1)	High
2	Male	High	(Y)	Yes	(Z_2)	High
3	Male	Low	High	No	Yes	Low
4	Female	High	High	Yes	(Z_4)	High
5	Male	High	High	Yes	(Z_5)	High
6	Male	High	High	Yes	(Z_6)	High
7	(X)	Low	Low	No	Yes	Low
8	Male	Low	High	No	No	Low
9	Male	Low	High	Yes	(Z_9)	High
10	Female	Low	High	Yes	(Z_10)	Low

iii) Construct the decision tree for the cleaned dataset obtained in Q2.b.ii using ID3 algorithm.

iv) Suppose the management of the hospital decided to include an additional numerical feature named "Age" which varies from 23 years to 83 years. Briefly explain the strategy you will be followed to reconstruct the decision tree using ID3 algorithm.

[6.0 Marks]

c) Suppose you are using a linear Support Vector Classifier for a 2-class classification problem. You have been given the data shown in Figure Q2.c, in which some points are circled to represent support vectors.

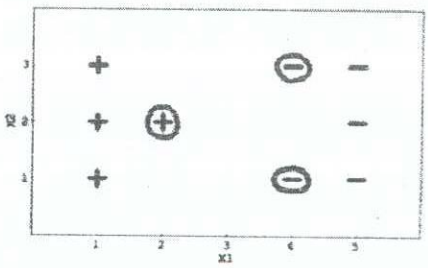

Figure Q2.c

i) If you remove the one of the non-circled points from the data points shown in Figure Q2.c, will the decision boundary change for the remaining 8 data points? Justify your answer.

ii) If you remove one circled point from the above data points shown in Figure Q2.c, will the decision boundary change for the remaining 8 data points? Justify your answer.

[1.0 Marks]

d) i) Explain how you apply Support Vector Machine (SVM) classifier for the dataset given in Figure Q2.d. X1 and X2 are separate two features respectively.

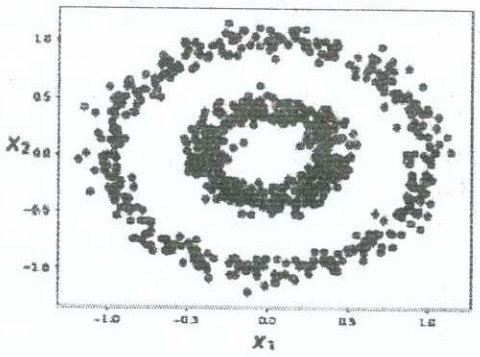

Figure Q2.d

ii) Describe the terms "Hard Margin" and "Soft Margin" in SVM using suitable illustrations.

iii) Describe the term "Kernel trick" in SVM classifier.

[1.0 Mark]

Q3) a) i) Write down the equation of Bayes theorem while explaining the terms.

ii) Starting from the Bayes theorem, derive the equation of the Naïve Bayes classifier with explaining relevant steps.

iii) State two assumptions you made when applying the Naïve Bayes classifier to solve a particular problem.

iv) Consider a scenario where an electronics retailer wants to predict whether an employee will purchase the iPhone 14 Pro Max. Table Q3.a shows whether an employee will purchase the iPhone 14 Pro Max based on four features (Age Group, Monthly Income, Previous iPhone Ownership, and Preferred Smartphone feature).

Table Q3.a

Employee ID	Age Group	Monthly Income	Previous iPhone Ownership	Preferred Smartphone feature	Purchase iPhone 14 Pro Max
1	Low	Low	Yes	Camera	No
2	Low	Low	Yes	Design	No
3	Medium	Low	Yes	Camera	Yes
4	High	Medium	Yes	Camera	Yes
5	High	High	No	Camera	Yes
6	High	High	No	Design	No
7	Medium	High	No	Design	Yes
8	Low	Medium	Yes	Camera	No
9	Low	High	No	Camera	Yes
10	High	Medium	No	Camera	Yes
11	Low	Medium	No	Design	Yes
12	Medium	Medium	Yes	Design	Yes
13	Medium	Low	No	Camera	Yes
14	High	Medium	Yes	Design	No

Predict whether employee number 15 will purchase iPhone 14 Pro Max using the Naïve Bayes Classifier. State all the following steps.

- Create frequency tables for each attribute against the target.
- Create Likelihood tables.
- Calculate the posterior probability for each class.
- Predict the outcome.

Employee ID	Age Group	Monthly Income	Previous iPhone ownership	Preferred Smartphone feature	Purchase iPhone 14 Pro Max
15	Low	High	Yes	Design	?

[4.0 Marks]

b) Consider the following data points to be clustering using the K-Means Clustering algorithm.

P1(1,3), P2(2,2), P3(5,8), P4(8,5), P5(3,9), P6(10,7), P7(3,3), P8(9,4), P9(3,7)

The initial centroids for K = 3 are C1 = (3,3), C2 = (3,7) and C3 = (9,4) respectively.

i) Calculate the distance between data points and cluster centroids using Euclidean distance formula and fill the following Table Q3.b.i. Sate all of your workings.

Important: Cluster assignment should be done in "Cluster" column after the first iteration.

Table Q3.b.i

Data Points	C1 (3,3)	C2 (3,7)	C3 (9,4)	Cluster
P1(1,3)				
P2(2,2)				
P3(5,8)				
P4(8,5)				
P5(3,9)				
P6(10,7)				
P7(3,3)				
P8(9,4)				
P9(3,7)				

ii) Re compute the new cluster centroids for Cluster 1, Cluster 2, and Cluster 3.

iii) According to the new cluster centroids calculate the distance between data points and cluster centroids again using the Euclidean distance formula and fill the following Table Q3.b.iii. State all of your workings.

Important: Cluster assignment should be done in "Cluster" column after the second iteration.

Table Q3.b.iii

Data Points	New C1	New C2	New C3	Cluster
P1(1,3)				
P2(2,2)				
P3(5,8)				
P4(8,5)				
P5(3,9)				
P6(10,7)				
P7(3,3)				
P8(9,4)				
P9(3,7)				

iv) Re compute the new cluster centroids for Cluster 1, Cluster 2, and Cluster 3 again.

v) According to the answer of part b(iv), should you calculate the distance between data points and the new cluster centroids again? Justify your answer?

[6.0 Marks]