Exam Date & Time: 26-Apr-2023 (01:15 PM - 04:30 PM)

CHAROTAR UNIVERSITY OF SCIENCE AND TECHNOLOGY

University Examination April 2023
B.Tech (CS)- VI
Time:- 01:15 pm to 04:30 pm
MACHINE LEARNING [CS360]

Marks: 70 **Duration: 195 mins.**

Section-I

Answer all the questions. Section Duration: 40 mins

1 Which of the following is NOT hyper-parameter in neural network? (1)

1) Learning rate 2) Number of hidden layers 3) Weights 4) Batch size

2 Which of the following is an application of Generative AdversarialNetworks (GANs)? (1)

1) Face classification and detection 2) Data compression 3) Speech recognition 4) Face Aging

3 In which algorithm, more computation time is required to testunseen samples? (1)

1) K-Nearest Neighbours 2) Neural Network 3) Logistic Regression 4) Support Vector Machines

4 Which algorithm is used for dimensionality reduction? (1)

1) K-Nearest Neighbors 2) Random Forest 3) Support Vector Machines 4) Principal Component Analysis

5 Which evaluation metric is used for regression problems? (1)

1) Mean Squared Error 2) Recall 3) F1 Score 4) Precision

6 Which of the following is NOT a popular machine learning library inPython? (1)

1) TensorFlow 2) Scikit-learn 3) Keras 4) PyCharm

7 In leave-one-out cross-validation, how many samples are used fortraining in each fold if there are N samples in a dataset? (1)

1) N-1 2) N/2 3) N/3 4) 1

8 Cluster quality depends on _____ intra-class distance and_____ inter-class distance. (1)

1) average, minimum 2) minimum, maximum 3) maximum, minimum 4) minimum, average

9 In density-based clustering, what is the core point? (1)

1) A point that has a certain number of other points within a specified distance 2) A point that is not close to any other point 3) A point that is close to all other points 4) None of the above

10 What is the minimum no. of variables/ features required to performclustering? (1)

1) 0 2) 1 3) 2 4) 3

11 In SVM, what is the role of the kernel function? (1)

1) To transform the data into a higher-dimensional space 2) To reduce the dimensionality of the data 3) To normalize the data 4) None of the above

12 Which of the following statements is true about the K-meansclustering algorithm? (1)

1) The K-means algorithm is not sensitive to outliers. 2) The K-means algorithm is sensitive to outliers. 3) The algorithm removes outliers from the data before clustering. 4) None of the above

13 Which of the following techniques can be used to handle missingdata? (1)

1) Deletion 2) Imputation 3) Regression 4) All of the above

14 Which of the following is a common technique used in datavisualization to show the distribution of a continuous variable? (1)

1) Bar chart 2) Box plot 3) Pie chart 4) Scatter plot

15 In which of the following situations is the most appropriate forchi-squared test? (1)

1) Comparing the means of two groups 2) Testing the correlation between two variables 3) Comparing the variances of two groups 4) Testing the independence of two categorical variables

16 What is a common activation function used in deep learning neuralnetworks? (1)

1) Linear activation function 2) Logistic activation function 3) Rectified Linear Unit (ReLU) activation function 4) Exponential activation function

17 Suppose you are given the following data for the age and weight (inkg) of 5 individuals (Age, Weight): (3,7), (4,6), (8,8), (6,7), and(2,4). What is the correlation coefficient between age and weight?Which kind of relationship is shown by these attributes? (2)

1) -0.729, Strong negative correlation 2) 0.807, Strong positive correlation 3) 0.623, Moderate positive correlation 4) 0.123, No correlation

18 The company collected data on the salaries of 100 employees. Thesample data is given in thousands (Rs): 22, 13, 28, 5, 45, 13, 25,48, 56. What is the range and mode of given data? (2)

1) 51, 13 2) 51, 56 3) 34, 13 4) 34, 56

Section-II

Answer 5 out of 6 questions.

- Figures to the right indicate ***full*** marks.
- Make suitable assumptions and draw neat figures wherever if required.

1 AtoZ-mart company wants to analyze the sales data of its fivebranches located in different areas. They have collected data onthe number of items sold and the total revenue generated by eachbranch in a particular month. The company wants to segment thebranches into two groups based on their sales. (5)

Branch	Items Sold	Total Revenue
B1	100	5000
B2	300	12000
B3	500	20000

Branch	Items Sold	Total Revenue
B4	200	10000
B5	400	15000

Find the two groups of branches using k-means clustering withEuclidean distance. Initialize the centroids with the followingpoints: (100, 5000) and (500, 20000). Show the steps of thealgorithm, including the calculation of distance, assignment ofpoints to clusters, and update of centroids.

2 Consider a dataset of the ages of 6 people: 2, 8, 3, 10, 5, and 15.Perform agglomerative hierarchical clustering and plot dendrogram.Use complete-linkage method. Find the two clusters. (5)

3 Manav Corporation is a clothing manufacturer that sells itsproducts in different regions of the world. They have collecteddata on the age, expenditure, and income of their customers andwhether they purchased a particular item or not. The data has beenrecorded for 8 customers as shown below:

Age	Expenditure (in K)	Income (in K)	Purchased
25	10	40	Yes
35	15	50	Yes
45	12	60	Yes
22	8.5	35	No
27	11	45	Yes
33	14.5	55	No
50	18	70	Yes
40	15.5	65	No

(5)

Manav Corporation wants to use k-Nearest Neighbors (kNN)classification to predict whether a customer will purchase aparticular item or not based on their age, expenditure, and income.Using k=3 and the Euclidean distance, classify whether the customerwith the following characteristics will purchase the item ornot:

Age	Expenditure (in K)	Income (in K)	Purchased
30	12.5	55	?

Perform all the necessary steps of the kNN classificationalgorithm.

4 ALL-IS-WELL Organization is a pharmaceutical company that producesa drug to treat a certain disease. The drug is considered effectiveif it cures the disease in the patient. The ML system of thecompany has conducted a clinical trial to test the drug'seffectiveness on 100 patients. The results of the trial are asfollows:

- 70 patients were correctly cured by the drug.
- 20 patients were mistakenly diagnosed as not cured.
- 5 patients were mistakenly diagnosed as cured.
- 5 patients were correctly diagnosed as not cured.

(5)

Using the above information, construct a 2x2 confusion matrix andcalculate the accuracy, precision, recall, and F1-score.

5 Case Study: "Bright-Future Investments (BFI)" is a financialcompany that provides loans to individuals. The company hascollected data on previous loan applications and their outcomes,including whether the loan was approved or not, the applicant'sincome, credit score, and other factors. They want to use this datato build a machine learning model that can predict whether a newloan application will be approved or not. (5)

Design a machine learning pipeline that would help BFI company toachieve its goal. Provide a step-by-step explanation of thepipeline and justify your choices.

6 Match the following:

1)	Supervised Learning	A)	A machine learning approach that learns patterns by matching the characteristics from unlabeled data.
2)	Machine Learning	B)	A subfield of machine learning that uses neural networks with many layers to learn complex representations of data.
3)	Deep Learning	C)	A subfield of artificial intelligence which has ability to learn without being explicitly programmed.
4)	Unsupervised Learning	D)	A type of neural network designed to recognize patterns in sequential data, such as time series or text.
5)	Recurrent neural network	E)	Machine learning algorithms that uses labeled data to make predictions or classifications.

(5)

Section-III

Answer 5 out of 6 questions.

- Figures to the right indicate ***full*** marks.
- Make suitable assumptions and draw neat figures wherever if required.

1 You are appointed as ML engineer in the AI-based company and wantsto use a convolutional neural network (CNN) to classify images ofcats and dogs. You have a dataset of 10,000 images, with 5,000images of cats and 5,000 images of dogs. The images are of varyingsizes, with the largest being 1000x1000 pixels. Design a CNNarchitecture to perform classification task. Also, mention thehyper-parameters values for this architecture. (5)

2 Suppose a supermarket wants to analyze its sales data to find outwhich items are frequently purchased together by customers. Thestore has collected data on the transactions of 5 customers. Thedata is stored in a transactional database in the followingformat:

Transaction ID	**Items Purchased**
T1	{Bread, Milk, Eggs}
T2	{Milk, Eggs, Cheese}
T3	{Bread, Milk, Cheese}
T4	{Bread, Milk, Eggs, Cheese}
T5	{Bread, Eggs}

(5)

Using the Apriori algorithm, identify all the frequent itemsetswith a minimum support of 50%.

3 Suppose we want to predict the selling price of used cars based ontheir mileage. We collect the following data on 6 cars:

Mileage (in thousands)	**Selling Price (in thousands)**
50	100
100	70

(5)

80	80
120	50
60	90
70	95

Build a model using Ordinary Least Square (OLS) linear regressionmethod. Predict the selling price of a car which has 75,000mileage.
Hint: One of the variable can be calculated as

$$m = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}$$

x = independent variables
$\bar{x}$ = average of independent variables
y = dependent variables
$\bar{y}$ = average of dependent variables

4 Consider the following dataset of candidates who appeared for theinterview. The first three features of dataset are used to predictwhether the candidate is considered as qualified for a job or not.Organization wants to build a decision tree using information gainfor classification of the candidate . Find the root node of thedecision tree. (5)

Education	Experience	Certification	Qualified
Bachelor's	Low	No	No
Bachelor's	High	No	No
Master's	Medium	Yes	Yes
PhD	High	Yes	Yes
Bachelor's	Low	Yes	No
Master's	High	Yes	Yes
PhD	Medium	No	Yes

5 Solve: Use the following methods to normalize the below group ofdata.
25, 35, 40, 60, 100
(a) min-max normalization by setting min = 0 and max = 1 (5)
(b) z-score normalization

6 Define Bayesian Belief Networks. What are the components of aBayesian Belief Network? Explain with example. (5)

-----End-----