

Question Bank

Chapter 1

Short Questions

1. Why is machine learning needed for business organizations?
2. List out the factors that drive the popularity of machine learning.
3. What is a model?
4. Distinguish between the terms: Data, Information, Knowledge, and Intelligence.
5. How is machine learning linked to AI, Data Science, and Statistics?
6. List out the types of machine learning.
7. List out the differences between a model and a pattern. Patterns are local, whereas a model is global for an entire dataset—Justify.
8. Are classification and clustering the same or different? Justify.
9. List out the differences between labeled and unlabeled data.
10. Point out the differences between supervised and unsupervised learning.
11. What are the differences between classification and regression?
12. What is semi-supervised learning?
13. List out the differences between reinforced learning and supervised learning.
14. List important classification and clustering algorithms.
15. List at least five major applications of machine learning.

Long Questions

1. Explain in detail the machine learning process model.
2. List and briefly explain classification algorithms.
3. List and briefly explain unsupervised algorithms.

Numerical Problems and Activities

1. Let us assume a regression algorithm generates a model: $y=0.254+0.06x$ for data pertaining to week-wise sales data of a product. Here, x is the week, and y is the product sales. Find the prediction for the 5th and 8th weeks.
2. Give two examples of patterns and models.
3. Survey and find out at least five latest applications of machine learning.
4. Survey and list out at least five products that use machine learning.

Chapter 2

Short Questions

1. What is univariate data?
2. What are the types of data?
3. Distinguish between 'good' and 'bad' data.
4. What are the problems of data collection?
5. Explain missing data analysis.
6. What are the measures of central tendencies?
7. Why are central tendency and dispersion measures important for data miners?
8. What are the measures of skewness and kurtosis?
9. How is interquartile range useful in eliminating outliers?
10. List the visualization aids available for exploratory data analysis.
11. What is the use of the correlation coefficient for data mining?
12. What are the advantages of LU decomposition?
13. What is the difference between correlation and covariance?
14. What is the need for PCA analysis?
15. Justify the use of scree plots in PCA.
16. Explain the role of probability distributions in machine learning.
17. Explain the basics of the binomial distribution.
18. Distinguish between sample error and actual error.
19. List out the types of sampling techniques.
20. What is a hypothesis?

Long Questions

1. Explain the stages of the data management life cycle.
2. Explain the types of Big Data with an example.
3. Explain in detail data cleaning processes.
4. Explain in detail univariate data analysis.
5. Explain at least five charts in detail that help data visualization.
6. Explain the procedure for SVD.
7. Explain the process of obtaining principal components and its relevance in feature reduction.
8. Explain the procedure for hypothesis testing.
9. Explain the procedure for pair-t tests and the Chi-Square goodness-of-fit test.

Numerical Problems

1. For a given univariate dataset $S = \{5, 10, 15, 20, 25, 30\}$ of marks, find the mean, median, mode, standard deviation, and variance.
2. For a given univariate dataset $S = \{5, 10, 15, 20, 25, 30\}$ of marks, find the arithmetic mean and geometric mean.
3. For a given univariate dataset $S = \{5, 10, 15, 20, 25, 30\}$ of marks, find the five-point summary and plot the box chart.
4. For the given tables below, perform the descriptive analysis of data:

Table 2.6: Sample Data

Age	Weight
1	4.2
2	4.5
3	4.7
4	5.2
5	6
6	6.2
7	7
8	7.2
9	7.5
10	8.5

Table 2.7: Students Marks Table

Sid	English	Hindi	Maths	Science
1	45	70.5	90	40
2	60	72.5	80	45
3	60	80	90	50
4	80	80	90	80
5	85	72	70	60

Tasks:

- (a) Find min and max marks scored in each subject.
 - (b) Find details of the student who scored the highest marks in Maths.
 - (c) Find the students with English marks > 60 and Maths > 70 .
5. For univariate attributes such as Weight, English, and Maths marks (consider Tables 2.6 and 2.7), find:
- (a) Mean, Median, Mode
 - (b) Weighted Mean, Geometric Mean, and Harmonic Mean
 - (c) Variance and Standard Deviation
 - (d) Absolute Deviation, Mean Absolute Deviation, and Median Absolute Deviation
 - (e) Coefficient of Variation
 - (f) Skewness and Kurtosis
 - (g) Five-point summary, IQR, and Semi-Quartile
6. For the bivariate data such as English and Maths (consider Tables 2.6 and 2.7), find:
- a) Covariance and Correlation between two variables
 - b) Covariance between English and Hindi Marks

Machine Learning Tasks

7. Use appropriate Data Visualization to plot the above Table 2.6 using the following charts:

- a) Bar Plot and Pie Chart
 - b) Histogram, Box Plot, and QQPLOT
 - c) Dot Plot, Line Chart, Scatterplot
 - d) Stem and Leaf Plot
8. Solve the following set of equations using the Gaussian elimination method:
- a) $2x_1 + 5x_2 = 7$
 - b) $6x_1 + 12x_2 = 18$
9. Solve the following set of equations using the LU decomposition method:
- a) $2x_1 + 5x_2 = 7$
 - b) $x_1 + 12x_2 = 18$
10. Apply PCA for the following matrix and prove that it works:

$$\begin{pmatrix} 4 & 3 \\ 1 & 2 \end{pmatrix}$$

11. Apply SVD for the following matrix:

$$\begin{pmatrix} 4 & 3 \\ 1 & 2 \end{pmatrix}$$

Perform matrix decompositions and prove that SVD works.

12. Find covariance and correlation coefficients for the following two sets of data:

X: 1, 2, 6, 12

Y: 8, 12, 18, 22

Chapter 3

Review Questions

1. What are the various methods through which learning occurs?
2. Define the terms "hypothesis" and "hypothesis space."
3. Explain predictor variables and target variables.
4. What is meant by a target function?
5. Provide a definition of concept learning.
6. How would you describe a "Concept" in learning?
7. What are the three essential components required for concept learning?
8. How is a hypothesis typically represented?
9. Analyze the training dataset provided in Table 3.5, which consists of six instances.

Table 3.5: Training Dataset

S.No.	Horns	Tail	Tusks	Paws	Fur	Color	Hooves	Size	Elephant
1	No	Short	Yes	No	No	Black	No	Big	Yes
2	No	Short	No	No	No	Brown	No	Medium	Yes
3	Yes	Short	No	No	No	Brown	Yes	Medium	No
4	No	Short	Yes	No	No	Black	No	Medium	Yes
5	No	Long	No	Yes	Yes	White	No	Medium	No
6	No	Short	Yes	Yes	Yes	Black	No	Big	Yes

Generate a consistent set of hypotheses using: (a) The Find-S algorithm (b) The Candidate Elimination algorithm

10. Consider the sample training instances presented in Table 3.6, which describe symptoms of individuals along with their Covid-19 test results. Apply the hypothesis search space to derive the consistent set of hypotheses.

Table 3.6: Sample Training Instances

S.No.	Fever	Cough	Throat Pain	Body Pain	Covid-19
11	N	Y	Y	Y	Positive
12	Y	Y	Y	Y	Positive
13	Y	N	Y	Y	Positive
14	N	N	Y	N	Negative

15	Y	Y	N	N	Positive
16	N	N	N	N	Negative
17	N	N	N	N	Negative

11. Define induction and inductive bias.
12. Explain the three types of prediction errors.
13. Discuss the relationship between bias and variance.
14. What do the terms "overfitting" and "underfitting" mean?
15. Differentiate between model parameters and hyperparameters.
16. Explain the various resampling methods used in machine learning.
17. Provide an overview of different learning frameworks.

Chapter 4

Review Questions

1. What do you understand by similarity-based learning?
2. Compare and contrast instance-based learning and model-based learning.
3. Why are instance-based learners called lazy learners?
4. Differentiate between lazy learning and eager learning.
5. Why is the k-NN method referred to as a memory-based method?
6. Why is data normalization/standardization required in k-NN?
7. What are the benefits and limitations of the k-NN algorithm?
8. Consider the following training dataset consisting of 10 data instances, as shown in Table 4.12. The dataset describes the award performance of students based on GPA and the number of projects completed. The target variable is 'Award', which is a discrete-valued variable taking values 'Yes' or 'No'.

Table 4.12: Training Dataset

S.No.	GPA	No. of Projects Done	Award
1	9.5	5	Yes
2	8.0	4	Yes
3	7.2	1	No
4	6.5	5	Yes
5	9.5	4	Yes
6	3.2	1	No
7	6.6	1	No
8	5.4	1	No
9	8.9	3	Yes
10	7.2	4	Yes

Given a test instance (GPA = 7.8, No. of projects done = 4), classify the test instance using the training set with $k = 3$ by applying:

- k-Nearest Neighbor classifier
 - Weighted k-Nearest Neighbor classifier
 - Nearest Centroid Classifier
9. A COVID care center aims to develop a case-based reasoning system to predict whether a person will test positive or negative based on symptoms. The table below presents a sample set of training instances.

Table 4.13: Sample Set of Instances

S.No.	Fever	Dry Cough	Tiredness	Sore Throat	Diarrhea	Headache	Loss of Taste	Shortness of Breath	Chest Pain	Result
-------	-------	-----------	-----------	-------------	----------	----------	---------------	---------------------	------------	--------

							or Smell			
1	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Positive
2	Yes	No	Yes	No	No	Yes	No	No	No	Negative
3	No	No	No	No	No	No	No	No	No	Negative
4	Yes	No	No	No	No	No	No	No	No	Negative
5	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Positive
6	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Positive
7	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Positive
8	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Positive
9	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Positive
10	No	No	No	No	No	No	No	No	No	Negative

- Determine an optimal k value for better prediction accuracy.
- Examine the impact of increasing the k value on prediction accuracy. Is it beneficial to have a larger or smaller k value?
- Apply an appropriate similarity measure (asymmetric binary features) to predict the result for an instance with the following symptoms:
 - Fever = Yes
 - Dry Cough = Yes
 - Tiredness = Yes
 - Sore Throat = Yes
 - Diarrhea = No
 - Headache = No
 - Loss of Taste or Smell = No
 - Shortness of Breath = No
 - Chest Pain = No

10. What is meant by locally weighted regression?

Chapter 5

Review Questions

1. What is the role of the regression model in exploratory data analysis?
2. Distinguish between the terms: Classification, Regression, and Estimation.
3. What is the difference between Classification and Regression models?
4. What is the principle of ordinary least squares in linear regression? What are the pros and cons of regression models?
5. Consider the following dataset in **Table 5.11**, where the week and the number of working hours per week spent by a research scholar in a library are tabulated. Based on the dataset, predict the number of hours that will be spent by the research scholar in the 7th and 9th weeks. Apply the linear regression model.

Table 5.11: Sample Data

x (Week)	1	2	3	4	5
y (Hours Spent)	12	18	22	28	35

6. The height details of boys and girls are given in **Table 5.12**.

Table 5.12: Sample Data

Height of Boys (x)	63	70	75	78
Height of Girls (y)	63	67	70	73

Fit a suitable line of best fit for the above data.

7. Using multiple regression, fit a line for the following dataset shown in **Table 5.13**. Here, z is the equity, x is the net sales, and y is the asset. z and x are independent variables, and y is the dependent variable. All the data is in million dollars.

Table 5.13: Sample Data

z	x	y
4	12	8
7	22	16
8	22	36
11	35	42

8. How does the polynomial regression model work?
9. How is logistic regression different from linear regression?

Chapter 6

Review Questions

1. How does the structure of a decision tree help in classifying a data instance?
2. What are the different metrics used in deciding the splitting attribute?
3. Define Entropy.
4. Relate Entropy and Information Gain.
5. How does a C4.5 algorithm perform better than ID3? What metric is used in the algorithm?
6. What is CART?
7. How does CART solve regression problems?
8. What is meant by pre-pruning and post-pruning? Compare both the methods.
9. How are continuous attributes discretized?

Table 6.42: Training Dataset

S.No.	Percentage	Award
1	95	Yes
2	80	Yes
3	72	No
4	85	Yes
5	95	Yes
6	32	No
7	96	Yes
8	64	No
9	89	Yes
10	72	Yes

10. Consider the training dataset shown in Table 6.42. Discretize the continuous attribute 'Percentage'.
11. Consider the training dataset in Table 6.43. Construct decision trees using ID3, C4.5, and CART.

Table 6.43: Training Dataset

S.No.	Assessment	Assignment	Project	Seminar	Result
1	Good	Yes	Good	Pass	
2	Average	Yes	Poor	Fail	
3	Good	No	Yes	Pass	

4	Good	No	Good	Pass	
5	Good	Yes	Good	Pass	
6	Average	No	Fair	Pass	
7	Poor	Yes	Poor	Fail	
8	Poor	No	Good	Fail	
9	Good	Yes	Fair	Pass	
10	Good	Yes	Fair	Pass	

Chapter8

Review Questions

1. What is meant by probabilistic-based learning?
2. Differentiate between probabilistic models and deterministic models.
3. What is Bayesian learning?
4. Define the following terms:
 - Conditional probability
 - Joint probability
 - Bayesian probability
 - Marginal probability
5. What is belief measure?
6. What is marginalization?
7. What is the difference between prior, posterior, and likelihood probabilities?
8. State Bayes' theorem.
9. Define Maximum A Posteriori (MAP) Hypothesis, , and Maximum Likelihood (ML) Hypothesis, .
10. Verify Bayes' theorem with an example.
11. Consider three baskets, Basket I, Basket II, and Basket III, each containing rings of red and green colors. The distribution of rings is as follows:
 - Basket I: 6 red rings and 5 green rings.
 - Basket II: 3 green rings and 2 red rings.
 - Basket III: 6 red rings. A person picks a ring randomly from a basket. If the ring is red, determine the probability that it was picked from Basket II.
12. Given the following probabilities:
 - Probability of a person having Malaria: 0.02%.
 - Probability of a positive test result given the person has Malaria: 98%.
 - Probability of a negative test result given the person does not have Malaria: 95%.Find the probability that a person has Malaria, given that the test result is positive.

Naïve Bayes and Gaussian Naïve Bayes Applications

13. Predict the result of a student using the Naïve Bayes algorithm with the following training dataset:

Table 8.17: Training Dataset

S.No	Assessment	Assignment	Project	Seminar	Result
1	Good	Yes	Yes	Good	Pass
2	Average	Yes	No	Poor	Fail
3	Good	No	Yes	Good	Pass
4	Average	No	No	Poor	Fail

5	Average	No	Yes	Good	Pass
6	Good	No	No	Poor	Pass
7	Average	Yes	Yes	Good	Fail
8	Good	Yes	Yes	Poor	Pass

Given a test instance with (Assessment = Average, Assignment = Yes, Project = No, Seminar = Good), predict the result of the student using Naïve Bayes, applying Laplace Correction if necessary.

14. Predict a student's result using the Gaussian Naïve Bayes algorithm with the following training dataset:

Table 8.18: Training Dataset

S.No	Assessment Marks	Assignment Marks	Seminar Done	Result
1	95	8	Good	Pass
2	71	5	Poor	Fail
3	93	9	Good	Pass
4	62	4	Poor	Fail
5	81	9	Good	Pass
6	93	8.5	Poor	Pass
7	65	9	Good	Pass
8	45	3	Poor	Fail
9	78	8.5	Good	Pass
10	56	4	Poor	Fail

Given a test instance (Assessment Marks = 75, Assignment Marks = 6, Seminar Done = Poor), predict the result of the student.

Chapter10

Review Questions

1. Compare biological neurons and artificial neurons.
2. What is the drawback of the McCulloch & Pitts mathematical model of an artificial neuron?
3. What are activation functions?
4. List some examples of linear and non-linear activation functions.
5. Why can't a simple perceptron solve the XOR problem?
6. Design a two-layer perceptron network to implement NAND gates. Assume weights and biases in the range of $[-0.5, 0.5]$ with a learning rate of $\alpha = 0.4$.
7. Discuss the different types of artificial neural networks.
8. Explain how a Multi-Layer Perceptron (MLP) solves the XOR problem. Design an MLP with backpropagation to implement the XOR Boolean function.
9. Compare a Radial Basis Function Neural Network (RBFNN) with a Multi-Layer Perceptron (MLP).
10. Consider a network with 4 input units and 2 output units. Four training samples are given, each represented as a vector of length 4:
 - o i1: (1, 1, 1, 0)
 - o i2: (0, 0, 1, 1)
 - o i3: (1, 0, 0, 1)
 - o i4: (0, 0, 1, 0)

Output Units: Unit 1, Unit 2

Learning Rate: $\eta(t) = 0.6$

Initial Weight Matrix:

- o **Unit 1:** [0.2, 0.8, 0.5, 0.1]
- o **Unit 2:** [0.3, 0.5, 0.4, 0.6]

Identify an algorithm to learn without supervision. How do you cluster them as expected?

Chapter13

Review Questions

1. Define and distinguish between classification and clustering.
 2. What are the advantages and disadvantages of clustering schemes?
 3. List the applications of clustering algorithms.
 4. Identify the challenges associated with clustering algorithms.
 5. What are the problems associated with clustering large datasets?
 6. Write the procedure for the agglomerative algorithm.
 7. What is k in the k -means algorithm? How is it selected?
 8. Explain why different initializations of the k -means algorithm yield different results.
 9. Distinguish between core points, border points, and noise points in clustering.
 10. Define density and explain how it is measured in the DBSCAN algorithm.
 11. What is a grid in clustering?
 12. State and explain the monotonicity property.
 13. What is a subspace, and why is it important in clustering?
 14. Define a model in the context of clustering.
 15. List the advantages and disadvantages of the Fuzzy C-Means (FCM) algorithm.
 16. What is a mixture model in clustering?
 17. Explain the concept of silhouette coefficient.
 18. List internal and external measures used for cluster validation.
-

Long Questions

1. Explain the k -means algorithm in detail.
 2. Describe the working of the DBSCAN algorithm.
 3. Provide a detailed explanation of the CLIQUE algorithm.
 4. What is fuzzy logic? How does the FCM algorithm help in cluster formation?
 5. Explain the Expectation-Maximization (EM) algorithm in detail.
-

Numerical Problems

1. Calculate the Euclidean, Manhattan, and Chebyshev distances for the given data points:
 - (2, 3, 4) and (1, 5, 6)
 - (2, 2, 9) and (7, 8, 9)
2. Compute cosine similarity, SMC, and Jaccard coefficients for the following binary data:
 - (1, 0, 1, 1) and (1, 1, 0, 0)
 - (1, 0, 0, 0, 1) and (1, 0, 0, 0, 1) with (1, 1, 0, 0, 0)
3. Find the Hamming distance between the following binary data pairs:
 - (1, 1, 1) and (1, 0, 0)
 - (1, 1, 1, 0, 0) and (0, 1, 1, 1, 1)
4. Determine the distance between:

- Employee ID: 1000 and 1001
 - Employee names: "John & John" and "John & Joan"
5. Find the distance between:
 - (Yellow, red, green) and (red, green, yellow)
 - (bread, butter, milk) and (milk, sandwich, tea)
 6. Apply hierarchical clustering on the dataset given in Table 13.14 using seed points (7, 8) and (16, 9). Generate a dendrogram.

Table 13.14: Sample Data

S.No.	X	Y
1	3	5
2	7	8
3	12	5
4	16	9
5	20	8

7. Apply the k-means algorithm with $k = 2$ to the dataset in Table 13.15 and present the results.

Table 13.15: Sample Data

S.No.	X	Y
1	3	5
2	7	8
3	12	5
4	16	9

Chapter14

Short Questions

1. How does reinforced learning differ from supervised and unsupervised learning methods?
2. What are the key components of reinforced learning?
3. Differentiate between total reward and total future reward.
4. Define a Markovian assumption.
5. Why is reinforced learning suitable for dynamic programming?
6. Explain temporal difference learning.
7. How does the Monte-Carlo method differ from temporal differencing?
8. What is Q-learning, and how is it different from SARSA?
9. Compare online and offline learning methods.

Long Questions

1. Explain how reinforcement learning problems can be modeled and solved in a conventional way.
2. Describe the Q-learning algorithm in detail.
3. Explain the SARSA algorithm.

Numerical Problems

1. Consider three universities: X, Y, and Z.
 - o 60% of students at university X prefer to do a master's at X itself, while 30% move to university Y and 10% to university Z.
 - o At university Y, 50% stay, 30% move to university X, and 20% move to university Z.
 - o All students at university Z stay for a master's.
 - o Construct the transition matrix and predict for two years given the initial distribution (0.5, 0.3, 0.2).
2. Consider the grid game where a robot can move:
 - o Identify empty grids and write states, actions, and episodes.
3. If the reward is 10, what is the expected reward for the 10th move with a discount factor of 0.3?
4. In the grid game, assume the agent can move UP, DOWN, RIGHT, and LEFT with probabilities (0.4, 0.2, 0.2, 0.2). Write the Bellman equation for the agent's position.
5. Using dynamic programming, determine the number of paths from school to home in the given grid where allowed actions are RIGHT and DOWN.
6. Compute utility values using a delayed reward factor ($\gamma = 0.7$) for navigating a robot from the start position to the goal state with a reward of +100. All other states are initialized with a reward of 0. Allowed operations: UP, RIGHT, LEFT, and DOWN. Diagonal movements are not allowed.
7. Using dynamic programming, find the optimal path from start to destination, given initial rewards and a grid where only RIGHT and DOWN movements are allowed.

#####