Machine Learning Practice Questions Part-I

Dr. Rajesh K. Maurya

2025

Instructions: All questions carry 5 marks each.

Part A: Theory-Based Questions

1. Explain the relationship and key distinctions between Artificial Intelligence (AI), Machine Learning (ML), and Deep Learning (DL). In your answer, describe the primary goal of machine learning.

2. Outline the machine learning process. Begin by explaining the four fundamental components of how machines learn (Data Storage, Abstraction, Generalization, and Evaluation) and then detail the practical stages from data collection to model optimization.

3. Compare and contrast supervised and unsupervised learning, highlighting the key difference in their training data. Describe the main applications for unsupervised learning, such as clustering, visualization, and anomaly detection.

4. Define supervised learning, explaining the role of "labels" in the training data. Describe its two primary tasks, classification and regression, and list at least three common algorithms for supervised learning.

5. Differentiate between Batch Learning and Online Learning. Discuss the advantages of online learning for systems with a continuous data flow and mention a key challenge associated with it.

6. Explain the difference between Instance-Based (Lazy) Learning and Model-Based (Eager) Learning. How does each approach utilize the training data to make predictions on new instances?

7. Discuss why data preparation is a critical step in the machine learning workflow. List and briefly describe the main steps involved, from data collection and loading to data splitting.

8. Explain the significance of (a) handling missing data and (b) feature engineering in preparing a dataset. For each, describe two techniques or strategies that can be employed.

9. Describe the purpose of feature scaling and its importance for certain ML algorithms. Explain the difference between Min-Max Scaling (Normalization) and Z-Score Scaling (Standardization).

10. Define Accuracy, Precision, Recall, and F1-Score as evaluation metrics for classification models. Explain a scenario where the F1-Score would be a more appropriate metric than accuracy.

11. What is a confusion matrix? Draw a 2x2 confusion matrix, labeling the axes and the four outcomes: True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN).

12. Describe the working principle of the K-Nearest Neighbors (KNN) algorithm for classification. Explain the role of the hyperparameter 'k' and discuss how the choice of a small vs. a large 'k' value can affect the model's performance.

13. What are the primary advantages and disadvantages of the KNN algorithm? Explain why KNN is considered a "lazy learner".

14. Explain the foundational principle of the Naive Bayes classifier, including its "naive" assumption of feature independence. State Bayes' Theorem and define its components (posterior probability, prior probability, likelihood).

15. Discuss the pros and cons of using the Naive Bayes classifier. Explain the "Zero Frequency" problem and how it can be addressed using a technique like Laplace smoothing.

16. Describe the role of distance metrics in the K-Nearest Neighbors (KNN) algorithm. Explain the mathematical formulation for both Euclidean and Manhattan distance. In what scenario would Hamming distance be a more appropriate choice?

17. Explain the concept of the "Curse of Dimensionality" and how it specifically affects the performance of the K-Nearest Neighbors (KNN) algorithm. What are some strategies to mitigate this issue?

18. Discuss the importance of feature scaling for the K-Nearest Neighbors (KNN) algorithm. Explain why distance-based algorithms like KNN are sensitive to the scale of features and what problems might arise if the data is not scaled.

19. How is the K-Nearest Neighbors (KNN) algorithm adapted for regression tasks as opposed to classification tasks? Describe how the final prediction is calculated in each case.

20. Discuss the methods for selecting an optimal value for the hyperparameter 'k' in the KNN algorithm. Explain the trade-off between model complexity and generalization by comparing the effects of a small 'k' (e.g., k=1) versus a large 'k'.

Part B: Numerical-Based Questions

16. **Naive Bayes with Laplace Smoothing (Customer Churn):**
A company has the following customer data:

Customer	Age Group	Plan Type	Tenure (yrs)	Churn
C1	Young	Basic	Short	Yes
C2	Young	Premium	Long	No
C3	Middle	Basic	Short	Yes
C4	Senior	Premium	Long	No
C5	Senior	Basic	Short	Yes
C6	Middle	Premium	Long	No

Using Naive Bayes with Laplace (Add-1) smoothing, compute the posterior probabilities and predict the churn for a customer with the following attributes: **Age Group = Senior, Plan Type = Premium, Tenure = Long.** Show all steps, including the calculation of prior and smoothed conditional probabilities.

17. **Naive Bayes with Laplace Smoothing (Disease Diagnosis):**
A hospital has collected the following data on a disease:

Patient	Cough	Fever	Fatigue	Disease
P1	Yes	Yes	Yes	Positive
P2	No	Yes	No	Negative
P3	Yes	No	Yes	Positive
P4	No	No	No	Negative
P5	Yes	Yes	No	Positive
P6	No	Yes	Yes	Negative

Predict whether a new patient with symptoms **Cough = Yes, Fever = Yes, and Fatigue = Yes** is "Positive" or "Negative" using Naive Bayes with Laplace smoothing. Show your complete calculations for the posterior probabilities for both outcomes.

18. **Naive Bayes with Laplace Smoothing (Spam Classification):**
Given the following email data:

Email	'Offer'	'Free'	Length > 200	Spam
E1	Yes	Yes	No	Yes
E2	No	Yes	Yes	No
E3	Yes	No	Yes	No
E4	Yes	Yes	Yes	Yes
E5	No	No	No	No
E6	Yes	Yes	No	Yes

Predict if an email with **'Offer' = Yes, 'Free' = Yes, and Length > 200 = No** is "Spam" using Naive Bayes with Laplace smoothing. Show your calculations for the posterior probabilities and state the final prediction.

19. **Naive Bayes with Laplace Smoothing (Loan Default):**
A bank uses the following features for predicting loan default:

Applicant	Job Type	Credit History	Homeowner	Default
A1	Salaried	Good	Yes	No
A2	Self-Employed	Poor	No	Yes
A3	Salaried	Poor	Yes	No
A4	Salaried	Good	No	No
A5	Self-Employed	Good	Yes	No
A6	Self-Employed	Poor	No	Yes

Using Naive Bayes with Laplace smoothing, determine whether a new applicant with **Job Type = Self-Employed, Credit History = Poor, and Homeowner = No** is likely to "Default." Show all calculations for prior probabilities, conditional probabilities, and posterior probabilities.

20. **KNN with Euclidean Distance:**
You are given four training samples with two attributes (X1: Acid Durability, X2: Strength) and a classification of "Good" or "Bad".

X1	X2	Classification
7	7	Bad
7	4	Bad
3	4	Good
1	4	Good

A new paper tissue has **X1 = 3 and X2 = 7**. Using the K-Nearest Neighbors algorithm with K=3 and the squared Euclidean distance, predict the classification of this new tissue. Show your calculations for the distance to all training samples and explain your final prediction based on the 3 nearest neighbors.

21. **KNN with Hamming Distance:**
A restaurant records customer preferences for burger flavors.

Burger	Chilly	Ginger	Pepper	Liked
A	true	true	true	false
B	true	false	false	true
C	false	true	true	false
D	false	false	true	true
E	true	false	false	true

Using the Hamming distance and a 3-NN classifier, predict whether a new burger with attributes **{pepper = false, ginger = true, chilly = true}** will be "liked". Show the calculated distance from the new burger to each burger in the training set and determine the final classification via majority vote.

22. **Evaluation Metrics from a Confusion Matrix:**
A binary classification model for spam email detection was tested on a dataset of 200 emails. The resulting confusion matrix is as follows:

	Predicted Negative (Non-Spam)	**Predicted Positive (Spam)**
Actual Negative (Non-Spam)	140 (True Negative)	10 (False Positive)
Actual Positive (Spam)	5 (False Negative)	45 (True Positive)

Based on this matrix, calculate the **Accuracy, Precision, Recall, and F1-Score** for the model. Show the formula and calculation for each metric.

23. **Naive Bayes (Weather Prediction):**
Given the weather dataset, first create a frequency table and then a likelihood table for the 'Weather' and 'Play' attributes.

Weather	Play
Sunny	No
Overcast	Yes
Rainy	Yes
Sunny	Yes
Sunny	Yes
Overcast	Yes
Rainy	No
Rainy	No
Sunny	Yes
Rainy	Yes
Sunny	No
Overcast	Yes
Overcast	Yes
Rainy	No

Using the likelihoods and prior probabilities, calculate the posterior probability P(Yes — Sunny) to determine if players will play when the weather is sunny. Show your work.

24. **Evaluation Metrics Calculation:**
A model is designed to predict whether a customer will churn (leave a service). After testing on 200 customers, the model produces the following results:

- True Positives (Correctly predicted churn): 25
- True Negatives (Correctly predicted no churn): 150
- False Positives (Incorrectly predicted churn): 10
- False Negatives (Incorrectly predicted no churn): 15

Construct a confusion matrix from these results. Then, calculate the model's **Accuracy, Precision, and Recall**.

25. **Naive Bayes with Laplace Smoothing (Text Classification):**
Using the text classification dataset below, calculate the probability that the sentence "A very close game" belongs to the "Sports" tag and the "Not Sports" tag. Use Naive Bayes with Laplace smoothing (add-1). There are 11 total words in the "Sports" category, 9 in "Not Sports," and a total vocabulary of 14 unique words. State the final classification based on your results.

Text	Tag
"A great game"	Sports
"The election was over"	Not Sports
"Very clean match"	Sports
"A clean but forgettable game"	Sports
"It was a close election"	Not Sports

26. **KNN for Regression (Car Price Prediction):**
A used car dealership has the following data for its inventory. Note that this is a regression problem.

Car ID	Mileage (1000s km)	Age (years)	Price ($1000s)
1	70	6	8
2	20	2	25
3	55	5	12
4	30	3	20
5	45	4	18

Using the KNN algorithm with K=3 and Euclidean distance, predict the price of a used car with **Mileage = 50 (in 1000s km) and Age = 4 years**. Show your distance calculations and explain how you arrived at the final predicted price.

27. **Naive Bayes (University Admission Prediction):**
A university has the following data on student admissions.

Applicant	GPA > 3.5	Entrance Score	Recommendation	Admitted
A1	Yes	High	Strong	Yes
A2	No	Medium	Weak	No
A3	Yes	Medium	Strong	Yes
A4	Yes	High	Weak	Yes
A5	No	Low	Weak	No
A6	Yes	Low	Strong	No
A7	No	High	Strong	Yes

Using the Naive Bayes classifier with Laplace (add-1) smoothing, predict whether a new applicant will be admitted with the following profile: **GPA > 3.5 = Yes, Entrance Score = High, Recommendation = Weak**. Show all steps, including prior and conditional probability calculations.

28. **Naive Bayes with Laplace Smoothing:**
An online marketing team tracks user data to predict if a user will click on an ad.

User	Age Group	Device	Visited Before	Clicked Ad
U1	Young	Mobile	Yes	Yes
U2	Adult	Desktop	No	No
U3	Young	Desktop	No	Yes
U4	Senior	Mobile	Yes	No
U5	Adult	Mobile	Yes	Yes
U6	Young	Desktop	No	Yes
U7	Senior	Desktop	Yes	No

Using Naive Bayes with Laplace (Add-1) smoothing, predict if a new user with the profile **Age Group = Adult, Device = Mobile, Visited Before = No** will "Click Ad". Show your complete calculations for the posterior probabilities.

29. **KNN with Hamming Distance (Movie Success):**
A film studio analyzes the success of its recent movies based on three categorical features.

Movie	Genre	Director	Lead Actor	Result
M1	Action	X	A	Hit
M2	Comedy	Y	A	Flop
M3	Action	Y	B	Hit
M4	Comedy	X	B	Hit
M5	Action	X	B	Flop

Using the Hamming distance and a 3-NN classifier, predict the result for a new movie with the following attributes: **Genre = Comedy, Director = X, Lead Actor = A**. Show the calculated distance to each movie in the training set and determine the final classification.

30. **Naive Bayes (E-commerce Purchase Prediction):**
An e-commerce site tracks user behavior to predict if a user will purchase a specific high-value item.

User	Age Group	Time on Site > 10m	Prior Purchase	Purchased Item
U1	Teen	Yes	No	No
U2	Adult	Yes	Yes	Yes
U3	Senior	No	Yes	No
U4	Adult	Yes	No	Yes
U5	Teen	No	No	No
U6	Senior	Yes	Yes	Yes
U7	Adult	No	Yes	Yes

With Naive Bayes and Laplace smoothing, predict if a new user will purchase the item given their profile: **Age Group = Adult, Time on Site > 10m = Yes, Prior Purchase = No**. Show your calculations for the posterior probabilities for both outcomes.

31. **KNN with Euclidean Distance (Fruit Classification):**
A system is trained to classify fruits based on their weight and a texture score.

Fruit	Weight (grams)	Texture (1-10)	Type
F1	150	9	Apple
F2	130	3	Orange
F3	180	8	Apple
F4	165	4	Orange
F5	190	5	Orange

Using the K-Nearest Neighbors algorithm with K=3 and standard Euclidean distance, classify a new fruit with **Weight = 160g and Texture = 7**. Show your distance calculations for each fruit in the dataset and determine the final classification using a majority vote.