Department of Computer Science and Engineering

Indian Institute of Technology Jodhpur

CSL 7620: Machine Learning

Minor Examination

Total marks: 40 **Set A Answers** **Time: 2 Hours**

Instructions:

- During the examination, invigilators will not address any queries. If you encounter anything unclear or incorrect in a question, make a reasonable assumption and proceed.
- Each question may have multiple correct options. You have to mark all the correct options. Otherwise, no marks will be awarded.
- Mark the correct choices with a pen on the OMR sheet. **No correction is allowed on the OMR sheet.**

Answers are marked in bold.

1. Which of the following are examples of machine learning applications? **[1 mark]**

   A. A navigation app that suggests the fastest route by analyzing real-time traffic data from other users.

   B. A weather application that displays the temperature and humidity based on data from a sensor.

   C. A video streaming service that recommends new movies and shows based on your viewing history and ratings.

   D. None of the others.

2. The hyperparameters of a machine learning model should ideally be chosen using **[1 mark]**

   A. Both training and test performances

   B. Test performance

   C. Training performance

   D. None of the others

3. Which of the following problems would generally require a data preprocessing step before training a machine learning model with the corresponding data? **[1 mark]**

   A. A dataset for house price prediction where the feature 'city' is represented as a string with over 50 unique names.

   B. A dataset of customer reviews where some entries are missing for the features 'rating' and 'review_text'.

   C. A dataset for a multiple linear regression where all features are continuous numerical values with similar ranges.

   D. None of the others

4. Which of the following statements correctly describe/ describes the objective function used in linear regression? **[1 mark]**

A. Through the minimization of the objective function, we aim to minimize the inverse of the total error between the predicted values and the actual values.

B. The objective function is not affected if some data points lie far away from other data points in the plot of independent vs dependent variables.

C. The objective function to be minimized is the sum of the squared differences between the actual and the predicted values.

D. None of the others

5. Which of the following statements about VC-Dimension is/are correct? **[1 mark]**

A. The VC-Dimension can be used to determine the exact test error for a given model.

B. A hypothesis space with a higher VC-Dimension is more likely to overfit the training data.

C. VC-Dimension provides a measure of the complexity or capacity of a hypothesis space.

D. None of the others

6. Which of the following is typically true for a model having high bias? **[1 mark]**

A. The model is too complex and fits the training data too well, including the noise.

B. The model has low error on the training data but high error on the test data.

C. The model is too simple to capture the underlying patterns in the data, leading to large errors on both the training and test sets.

D. The model performs very differently on different test sets.

7. Which of the following data transformations can be used for normalizing the scale of features?

[1 mark]

A. Log Transformation

B. Min-Max scaling

C. Z-score normalization (Standardization)

D. One-Hot Encoding

8. Consider a spaceship (S1) that observed different planets for life and water between the years 1990 and 2000. In 50% of the observed planets, the spaceship found both life and water. In 25% of the observed planets, there is life but no water. In 25% of the observed planets, there is no life and no water. All of these observations by S1 are found to be correct. Another spaceship (S2) is launched in 2024. This second spaceship can only detect the presence or absence of water on a planet. Consider one of the planets, P1, that was observed by the first spaceship, S1. Suppose the second spaceship, S2, found water on this planet, P1. What is the probability that there is life on the planet, P1? **[1 mark]**

A. 100%

B. 50%

C. 0%

D. None of the others.

9. Which statement best describes the Maximum Likelihood Estimator? **[1 mark]**

A. The parameter value that minimizes the data variance.

B. The parameter value that maximizes the likelihood of the observed data.

C. The parameter value that minimizes the posterior.

D. The parameter value that maximizes the prior.

10. Two classes $C1$ and $C2$ have the following Gaussian likelihoods: $P(x|C1) \sim N(0, 1)$, $P(x|C2) \sim N(2, 1)$. Priors for the two classes are equal. What is the Bayes decision boundary? **[1 mark]**

A. x=0.5

B. x=1.5

C. No boundary (always pick $C2$)

D. None of the others

11. Which expression is maximized by the Maximum Likelihood Estimator for samples $x^{(1)}, \ldots, x^{(n)}$?

[1 mark]

A. $\prod_{i=1}^{n} p(x^{(i)}|\theta)$

B. None of the others

C. $\sum_{i=1}^{n} p(x^{(i)}|\theta)$

D. $p(\theta|x)$

12. A machine must classify an observation x into three classes: $C1,\ C2,\ C3$

Priors: $P(C1) = 0.2,\ P(C2) = 0.5,\ P(C3) = 0.3$

Likelihoods for a particular observation x: $P(x|C1) = 0.3,\ P(x|C2) = 0.4,\ P(x|C3) = 0.2$

Which class should the minimum-risk classifier choose for observation x? **[1 mark]**

A. C1

B. C3

C. C2

D. Information is insufficient

13. Why is the log-likelihood function often used instead of the likelihood function in Maximum Likelihood Estimation (MLE)? **[1 mark]**

    A. Because the log likelihood changes the location of the maxima to make it unbiased.

    B. Because taking the log always increases the likelihood value.

    C. Because the log-likelihood eliminates the need to normalize probabilities.

    D. None of the others.

14. Given the following training samples: $(x_1 = 1, x_2 = 2, y = 5)$, $(x_1 = 2, x_2 = 3, y = 8)$, and $(x_1 = 3, x_2 = 4, y = 11)$, and a multiple linear regression model defined by the equation $\overline{y} = 2 + 1.5x_1 + 0.5x_2$, what is the Sum of Squared Errors (SSE) for this model on the given data? **[1 mark]**

    A. 1.5

    B. 8.75

    C. 11.25

    D. None of the others

15. Which of the following is/ are common reason/ reasons for performing normalization of feature values in machine learning? **[1 mark]**

    A. Increasing the dimensionality of the dataset.

    B. None of the others

    C. Converting continuous features into categorical features

    D. Preventing features with a larger range of values from dominating the learning process..

16. Which scenario best illustrates the use of reinforcement learning? **[1 mark]**

    A. Grouping genomic data to uncover genetic clusters among populations.

    B. Optimizing a supply chain's inventory levels by learning to order stock after every two days based on demand predictions and receiving feedback on the delay in delivery.

    C. Predicting employee attrition rates using a company's historical HR dataset.

    D. None of the others

17. Correctly match the type of machine learning with its primary characteristic **[1 mark]**

I. Supervised Learning	a. The model learns to find hidden patterns and groupings in unlabeled data without any predefined outcomes.
II. Reinforcement Learning	b. The model learns from trial and error, receiving rewards for correct actions and penalties for incorrect ones.
III. Semi-supervised Learning	c. The model is trained on a small amount of labeled data and a large amount of unlabeled data to improve accuracy.
IV. Unsupervised Learning	d. The model learns from a labeled dataset to predict an outcome or classify new data.

A. (I) -> (a), (II) -> (b), (III) -> (c), (IV) -> (d)

B. **(I) -> (d), (II) -> (b), (III) -> (c), (IV) -> (a)**

C. (I) -> (d), (II) -> (c), (III) -> (b), (IV) -> (a)

D. None of the others

18. It is known that 80-foot blue whales consume, on average, 3200 kg of krill per day. 100-footers consume, on average, 3600 kg of krill per day. Assume that the mean daily krill consumption varies linearly with whale length, and that the daily consumption for a given whale follows a Gaussian distribution with a standard deviation of 200 kg/day. **[1 mark]**

Which of the following is the correct form of the conditional distribution P(k|l), where k is daily krill consumption and l is whale length (in feet)?

A. $P(k|l) \ = \frac{1}{200\sqrt{2\pi}} exp\left(- \frac{(k-(40l+800))^2}{80,000}\right)$

B. $P(k|l) \ = \frac{1}{200\sqrt{2\pi}} exp\left(- \frac{(k-(10l+2400))^2}{80,000}\right)$

C. $P(k|l) \ = \frac{1}{200\sqrt{2\pi}} exp\left(- \frac{(k-(20l+1600))^2}{80,000}\right)$

D. None of the others.

19. A machine must classify incoming signals into two classes, C1 and C2.

Priors: P(C1) = 0.3, P(C2) = 0.7.

Likelihoods for a particular observation x: P(x | C1) = 0.4, P(x | C2) = 0.2

a1 is the action of predicting C1 and a2 is the action of predicting C2. The loss function (in the form L(action|true class)) is:

L(a1 | C1) = 0, L(a1 | C2) = 5 (predict C1 when true class is C2)

L(a2 | C1) = 1, L(a2 | C2) = 0 (predict C2 when true class is C1)

Which class should the minimum-risk classifier choose for observation x? **[1 mark]**

A. None of the others

B. Always classify as C2

C. For this observation x, classify as C2

D. For this observation x, classify as C1

20. A Naive Bayes model is used for **sentiment classification** (Positive vs Negative).

Prior probabilities: P(Pos) = 0.5, P(Neg) = 0.5
Likelihoods:
P("Great" | Pos) = 0.6, P("Great" | Neg) = 0.2
P("Slow" | Pos) = 0.1, P("Slow" | Neg) = 0.4

A review contains the words **"Great"** and **"Slow"**.

Which of the following are correct (posterior() stands for posterior probability)? **[1 mark]**

A. Posterior(Pos) ∝ 0.5 × 0.6 × 0.1

B. Posterior(Neg) ∝ 0.5 × 0.2 × 0.4

C. The review will be classified as Negative

D. None of the others

21. Which of the following statements about discriminant functions in classification are correct?

[1 mark]

A. A discriminant function assigns a score to each class, and the class with the highest score is chosen as the predicted class

B. In Naive Bayes, the discriminant function is nonlinear because it assumes feature independence.

C. Quadratic discriminant functions arise when class-conditional densities are Gaussian with different covariance matrices

D. None of the others

22. Given samples $x_1, \cdots, x_n \sim N(\mu, \sigma^2)$ with known σ^2, the MLE for μ is: **[1 mark]**

A. $\frac{1}{n}\sum_i x_i$

B. $\frac{1}{n}\sum_i x_i^2$

C. $\frac{1}{n-1}\sum_{i} x_{i}$

D. None of the others

23. Fill in the blanks with the best match from the options. **[1 mark]**

The **(I)** Gradient Descent algorithm is not very data efficient when the data is very similar. On the other hand, **(II)** Gradient Descent is not very computationally efficient and often produces noisy results since it uses only one data point at a time. To address these issues, many practitioners use **(III)** Gradient Descent, which balances computational efficiency and stability.

A. (I) -> Batch, (II) -> MiniBatch, (III) -> Stochastic

B. (I) -> Batch, (II) -> Stochastic, (III) -> MiniBatch

C. (I) -> MiniBatch, (II) -> Stochastic, (III) -> Batch

D. None of the others.

24. Which of the following statements correctly distinguishes Bayesian Estimation from Maximum Likelihood Estimation (MLE)? **[1 mark]**

A. Under a uniform prior, Bayesian estimation reduces to MLE.

B. MLE is not affected by the size of the dataset

C. None of the others

D. Bayesian estimation incorporates prior information, while MLE does not.

25. Consider the following values of the random variables X and Y

X={2,4,6,8,10}; Y={1,3,5,7,9}

Find out the correct statements **[1 mark]**

A. The sample variance of Y is 8.

B. The sample covariance between X and Y is 8.

C. The sample variance of X is 10.

D. None of the others

26. Which of the following statements about k-medoids clustering are true? **[1 mark]**

A. K-medoids chooses actual data points as cluster centers

B. Compared to K-means clustering, a larger number of steps in K-medoids involves randomness

C. K-medoids always has lower computational complexity than k-means

D. None of the others

27. Fill in the blanks with the best match from the options: **[1 mark]**

A data scientist is analyzing customer behavior data to create meaningful customer segments for targeted marketing. She considers using different clustering algorithms. She plans to use __(I)__ because it is efficient for large datasets, but she worries about its sensitivity to ____(II)__ and the difficulty of deciding the number of clusters. Therefore, she also considers ____(III)____, which can identify clusters based on __(IV)___ and does not require the number of clusters as input.

A. (I)->K-means, (II)-> Outliers, (III)->DBSCAN, (IV)->Centroid

B. (I)->K-medoids, (II)-> Centroid, (III)->K-means, (IV)->Density

C. (I)->K-means, (II)-> Outliers, (III)->DBSCAN, (IV)->Density

D. None of the others

28. Find out the correct statements: **[1 mark]**

A. DBSCAN is not sensitive to initialization

B. DBSCAN is guaranteed to produce final clustering results after a finite number of steps

C. K-means and K-medoids clustering techniques are guaranteed to converge

D. None of the others

29. Fill in the blanks with the best match from the options: **[1 mark]**

With an increase in the__(I)__ and __(II)__, the possibility of __(III)__ decreases. However, a reduction in ___(IV)___ may also be helpful in this context.

A. (I)->Number of training data, (II)-> Diversity in training data, (III)->Overfitting, (IV)->Model complexity

B. (I)->Model complexity, (II)-> Outliers, (III)->Overfitting, (IV)->training set size

C. (I)->number of parameters, (II)-> number of validation data, (III)->Overfitting, (IV)->Model complexity

D. None of the others

30. Consider the following 1-dimensional data points: 2,4,5,10,12,15. Suppose K-means clustering is applied with k=2 clusters and the initial cluster centroids of the first and second clusters are chosen as 4 and 12. After one iteration of assignment and centroid update, which of the following statements are true? **[1 mark]**

A. The updated centroid of the first cluster is 3.67

B. The updated centroid of the second cluster is 12.33

C. After this iteration, the sum of squared distances to centroid of the first cluster from every point of the first cluster is less than 30

D. None of the others

31. Select the correct option to fill in the blanks in the following paragraph

When preparing data for a machine learning model, choosing the right feature scaling technique is crucial. For instance, models that rely on the assumption of a normal distribution often perform better after **(I)** has been applied. This technique adjusts features so they have a mean of 0 and a standard deviation of 1. In contrast, **(II)** scales features to a fixed range, typically between 0 and 1. This method is particularly useful for algorithms that are not dependent on statistical distribution and instead require all features to be within a specific boundary. However, a major drawback is that **(III)** is highly susceptible to outliers, as they will directly impact the defined range. In such cases, **(IV)** might be a more robust choice, as it is less influenced by extreme values. Ultimately, the choice between **(V)** and **(VI)** often depends on the specific algorithm being used and the characteristics of the dataset. **[2 marks]**

A. (I) Data Standardization, (II) Min-max Normalization, (III) Min-max Normalization, (IV) Data Standardization, (V) Data Standardization, (VI) Min-max Normalization

B. (I) Data Standardization, (II) Min-max Normalization, (III) Data Standardization, (IV) Min-max Normalization, (V) Data Standardization, (VI) Min-max Normalization

C. (I) Min-max Normalization, (II) Data Standardization, (III) Data Standardization, (IV) Min-max Normalization, (V) Data Standardization, (VI) Min-max Normalization

D. None of the others

32. Consider a simple linear regression model with a single feature x and a single parameter w, defined as $\hat{y} = wx$. The objective function is the Mean Squared Error (MSE), given by $J(w) = \frac{1}{2m}\sum_{i=1}^{m}\left(y^{(i)} - \hat{y}^{(i)}\right)^2$. Given the training data points $(1, 3)$ and $(3, 5)$, an initial parameter value of $w = 1$, and a learning rate $\alpha = 0.1$, what is the updated value of w after two steps of batch gradient descent? **[2 marks]**

A. 1.6

B. 1.2

C. 1.8

D. None of the others

33. Joint Probability Table: Student Exam Performance Based on Study Habits and Sleep

	Studies Regularly	¬Studies Regularly	Studies Regularly	¬Studies Regularly
	Adequate Sleep	Adequate Sleep	¬Adequate Sleep	¬Adequate Sleep
Pass Exam	0.25	0.15	0.10	0.05

Fail Exam	0.05	0.10	0.15	0.15

Using inference by enumeration, calculate the probability that a student passes the exam given that they did not study regularly but had adequate sleep. **[2 marks]**

A. 0.40

B. 0.50

C. 0.60

D. None of the others

34. An observation x is to be classified into one of three classes: C1, C2, or C3. The model gives:

Priors: P(C1) = 0.2, P(C2) = 0.5, P(C3) = 0.3

Class-conditional likelihoods for this observation x:
P(x | C1) = 0.25, P(x | C2) = 0.10, P(x | C3) = 0.40

The following actions are possible:

I. a1 is the action of predicting C1

II. a2 is the action of predicting C2

III. a3 is the action of predicting C3

The losses for various actions are as follows (in the form L(action|true class))
L(a1|C1) =0; L(a1 |C2) = 2; L(a1 |C3) = 5
L(a2|C1) = 1; L(a2|C2)=0; L(a2|C3)=3
L(a3|C1) = 4; L(a3|C2) =1; L(a3|C3) =0

Find out the correct statements. **[2 marks]**

A. Posterior probabilities are approximately: P(C1|x)=0.327, P(C2|x)=0.227, P(C3|x)=0.445.

B. The expected risk R(a1) ≈ is 3.18.

C. The Bayes optimal action for this x (minimising expected risk under the given loss) is to predict C3.

D. None of the others

35. A tester measures lifetimes (in hours) for 5 lightbulbs from the same batch. All 5 bulbs failed during the test at times: 2, 3, 7, 5, and 8 hours. The tester models lifetimes as an Exponential with a failure rate parameter $\lambda > 0$. Consider the **Exponential function** for the lifetime of bulbs $t \geq 0$: $f(t|\lambda) = \lambda e^{-\lambda t}$. Using maximum likelihood, the estimated value of λ is **[2 marks]**

A. 0.08

B. 0.20

C. 0.25

D. None of the others