Sessional one

Q1: Multiple Choice Questions

(c) both inter-cluster distance and intra-cluster distance equally accounted

(d) lower the metric value, the better it is

(c) overfitting

(a) Low

(g) A

(b) A & C are correct. B is wrong

(d) False

(a) always quadratic

(c) A & C but not B

(b) Only A

Q2: Objective Type Questions

(b) Same variance for 2 features but the proportion of records from the class corresponding to $\mu 1$ is more in the dataset. Justification: The decision boundary is closer to the distribution with higher class proportion.

(c) B & C are correct, A is wrong. Justification: Both I and II can be plotted in Rd, and I is a regression equation, not a decision boundary for classification.

(b) Anjali's method is flawed. Feature engineering a new feature as the product of two z-features and averaging them as if it is the correlation coefficient has no mathematical basis.

(a) Covariance matrix with all variances = 1 and covariances = 0

(d) It cannot be said that eating at either Boda sheera or Suresh mess is causing students to fall sick.

1. K-Means Initialization:

- **Initial Centroids:**
  - $\mu_1=(-1,3)$
  - $\mu_2=(2,5)$
  - $\mu_3=(4,1)$
- **Manual Sampling:**
  - For each data point, calculate the distances to the centroids and assign the point to the nearest centroid.
  - **Assignments:**
    - (-2,3): $\blacksquare 1\mu_1$
    - (-1,3): $\blacksquare 1\mu_1$
    - (1,4): $\blacksquare 2\mu_2$
    - (2,5): $\blacksquare 2\mu_2$
    - (3,1): $\blacksquare 3\mu_3$
    - (4,1): $\blacksquare 3\mu_3$
    - (4,2): $\blacksquare 3\mu_3$

2. K-Means Clustering:

- **Iteration 1:**
  - **Assignments:**
    - (-2,3): $\blacksquare 1\mu_1$
    - (-1,3): $\blacksquare 1\mu_1$
    - (1,4): $\blacksquare 2\mu_2$
    - (2,5): $\blacksquare 2\mu_2$
    - (3,1): $\blacksquare 3\mu_3$
    - (4,1): $\blacksquare 3\mu_3$
    - (4,2): $\blacksquare 3\mu_3$
  - **New Centroids:**
    - $\blacksquare 1=(-3/2,3)\mu_1=(-3/2,3)$
    - $\blacksquare 2=(3/2,9/2)\mu_2=(3/2,9/2)$
    - $\blacksquare 3=(11/3,4/3)\mu_3=(11/3,4/3)$
- **Iteration 2:**
  - **Assignments:**

- (-2,3): μ_1
- (-1,3): μ_1
- (1,4): μ_2
- (2,5): μ_2
- (3,1): μ_3
- (4,1): μ_3
- (4,2): μ_3

- **Centroids remain unchanged.**

Q4

(a) Classifications:

- Nearest Centroid: **Supervised**
- kNN: **Supervised**
- KMeans Clustering: **Unsupervised**

(b) Table Filling:

Algorithm	**Nearest Centroid**	**kNN**	**Gaussian Model**
Eager vs Lazy	Eager (Compute centroids during training)	Lazy (Compute distances during prediction)	Lazy (Compute parameters during training)
Batch vs Instance	Batch (Single pass for centroid computation)	Instance (Considers individual data points)	Batch (Single pass for parameter computation)
Parametric vs Non-parametric	Parametric (Stores centroids)	Non-parametric (Stores data points)	Parametric (Stores parameters)
Discriminative vs Generative	Discriminative (Focuses on decision boundary)	Discriminative (Focuses on nearest neighbors)	Generative (Models class distributions)

Q2: Model Storage Space

(a) Nearest Centroid Model:

- **Storage Space:** $n \times 4$ bytes (for floating-point features)
- **Explanation:** In the Nearest Centroid model, only centroids are stored, and each feature value requires 4 bytes.

(b) kNN Model:

- **Storage Space:** $m \times n \times 4$ bytes (for floating-point features) + $2 \times 12 \times 1$ bytes (for k and weighting mechanism)
- **Explanation:** For kNN, all data points (m records with n features) need to be stored, along with two integer hyperparameters (k and weighting mechanism).

Q5: Feature Transformation

(a) Feature A (Lognormal Distribution):

- **Appropriate Transformation:** LogTransformer
- **Reasoning:** Feature A follows a lognormal distribution, and applying the LogTransformer is appropriate for such distributions.

(b) Feature B (Moderately Right-skewed Distribution):

- **Appropriate Transformation:** Square-root Transformer
- **Reasoning:** A moderately right-skewed distribution can benefit from a square-root transformation to make it more symmetric.

(c) Box-Cox Transformation:

- **Feature A (Lognormal):** Use $\lambda=0$ (no transformation)
- **Feature B (Right-skewed):** Use $\lambda=0.5$ (square root transformation)
- **Reasoning:** Box-Cox transformation is applicable when features are greater than 0, and the choice of lambda is data-dependent. $\lambda=0$ corresponds to a log transformation, and $\lambda=0.5$ corresponds to a square root transformation.

Sessional 2

Q1: Agglomerative Clustering and Linkage Types

Which linkage method results in long chains?

Answer: (d) Single linkage

Reason: Single linkage tends to create long chains because it joins clusters based on the minimum distance between individual data points in the clusters.

Top-Down Hierarchical Clustering:

Answer: Divisive clustering

Explanation: Divisive clustering is top-down, as it starts with the entire dataset and recursively divides it into smaller clusters.

Matching Clustering Types to Algorithms:

a - ii (KMeans)

b - iv (GMM clustering)

c - i (DBSCAN)

d - iii (Divisive clustering)

Matching Linkage Types to Formulas:

a - ii

b - iii

c - i

d - iv

Regularization for Sparsity in Linear Regression:

Answer: (a) L1 regularization

Reason: L1 regularization (Lasso) induces sparsity by adding the absolute values of the coefficients to the loss function, encouraging some coefficients to become exactly zero.

Q2: Decision Trees

Decision Tree for Given Split:

Drawing a decision tree based on the split shown is not possible without specific details about the split boundaries.

Gini Impurity for First Split:

Selected Feature: Age

Calculation Steps: (Details provided by dataset are needed for calculation.)

Pruned Decision Trees:

Answer: Decision tree (c)

Reason: Decision tree (c) has higher accuracy post pruning, as pruning can prevent overfitting to the training dataset.

Matching Decision Tree Algorithms to Formulas:

a - i

b - ii

c - iv

d - iii

Q3: Ensemble Methods and Other Concepts

Bagging Ensemble and Random Forest:

Answer: False

Reason: A bagging ensemble is a broader concept; a Random Forest is a specific type of bagging ensemble that uses decision trees as base learners.

SMOTE Oversampling:

Answer: False

Reason: SMOTE oversamples the minority class by generating synthetic examples rather than under sampling the majority class.

Random Forest Impact on Bias and Variance:

Answer: (d) B and D are correct

Reason: Random Forest tends to decrease variance but may increase bias slightly.

SMOTE-Tomek Links Algorithm:

Explanation: SMOTE-Tomek Links combines SMOTE oversampling with Tomek Links undersampling to create a balanced dataset.

Mutual Information Calculation:

Answer: (a) Feature is numerical and target is categorical

Reason: Mutual information is often used for measuring the dependency between numerical and categorical variables.

Feature Selection Methods:

Answer: (b) Embedded methods

Reason: Embedded methods include feature selection as part of the machine learning training process.

False Statements:

Answer: (b) B, C and E

Explanation: Dependence between features does not imply correlation, and correlation does not imply dependence. Correlated features can be independent.

Q1: Metrics for Rare Disease Detection

1. **Choice of Metrics:**
    - **Answer:** (a) Precision and False Positive Rate (FPR)
    - **Reasons:**
        - **Precision:** Emphasizes the accuracy of positive predictions, crucial for a rare disease to avoid false alarms.
        - **False Positive Rate (FPR):** Important for understanding the rate of false alarms, crucial in a scenario where the disease is rare.

Q2: Summary Metrics in ML Classification

2. **Summary Metrics:**
    - **Answer:** Precision and Recall
    - **Reason:** Precision and Recall provide a balanced view, especially in the context of a rare disease, where both false positives and false negatives need careful consideration.

Q3: F-1 Score Design

3. **F-1 Score Design:**
    - **Explanation:** The F-1 score uses the harmonic mean because it balances precision and recall. The harmonic mean is less sensitive to extreme values, making it suitable for cases where one metric should not dominate the other.

Q4: Threshold Adjustment for Chocolate Muffins

4. **Approach to Decrease False Positives:**
    - **Approach:** Adjust the probability threshold to classify instances as chocolate muffins. Increase the threshold to decrease false positives.

Q6: Linear Regression and Related Concepts

6. **Linear Regression Objective Function:**
    - **Answer:** (c) $J(x; \theta) = m1/(m) \sum(y(i) - \theta^T x(i))^2 + \beta \Sigma(\theta j)^2$
    - **Reason:** Represents linear regression with L2 regularization, balancing the fit to data and regularization.

Mechanism for Detecting Multicollinearity:

- **Answer:** (b) Eigen decomposition
- **Reason:** Eigen decomposition is not a mechanism for detecting multicollinearity; it is a method for diagonalizing the covariance matrix.

7. **Heteroskedasticity:**
    - **Explanation:** Heteroskedasticity is the phenomenon where the variability of a variable is unequal across different values of another variable. *(Diagram not provided, but it would illustrate the concept.)*
8. **Linear Regression Type:**
    - **Answer:** (c) Overdetermined system of equations
    - **Reason:** Linear regression involves solving an overdetermined system where the number of equations (data points) is greater than the number of unknowns (coefficients).
9. **Linear Regression Objective Function (with Intercept):**
    - **Answer:** (a) $J\theta(x) = m^{-1} \Sigma_i(y - (\theta^T x(i) + z))^2$
    - **Reason:** Represents linear regression with an intercept z.