Roll No: | | | | | | | | | | | |

PRESIDENCY UNIVERSITY

GAIN MORE KNOWLEDGE
REACH GREATER HEIGHTS

Presidency University Act, 2013 of the Karnataka Act No. 41 of 2013 | Established under Section 2(f) of UGC Act, 1956
Approved by AICTE, New Delhi | Approved By BCI
Bengaluru

Mid Term Examinations - October 2026

Date: 05/10/2026 **Time:** 11:45 AM - 01:15 PM

Course Code: CSE3502 **Course Name:** Computational Optimization for Intelligent Systems

Semester: Seventh Semester **Max. Marks:** 50 **Weightage:** 50%

CO - Levels	CO1	CO2	CO5
Marks	92	66	22

ANSWER KEY

PART A: Answer Following Questions. 10 M

Q.N.	Questions	M	CO	BT
1	Empirical Risk Minimization (ERM) is a machine learning principle in which a model is trained by minimizing the average loss/error on the training dataset. $R_emp(f) = (1/n) * \sum_{i=1}^{n} L(y_i, f(x_i))$ Its main objective is to find a model that minimizes the training error.	2	CO1	BT1
2	VC-dimension (Vapnik–Chervonenkis dimension) is a measure of the capacity or complexity of a learning model. It represents the largest number of points that the model can correctly classify in all possible ways. A higher VC-dimension generally indicates a more complex model with greater capacity.	2	CO1	BT1
3	The sigmoid function converts the output of a linear equation into a value between 0 and 1, which can be interpreted as the probability of belonging to a class. $\sigma(z) = 1/(1 + e^{-z})$ For example, a probability greater than 0.5 may be classified as class 1.	2	CO2	BT1
4	The kernel trick is a technique used in SVM to handle non-linearly separable data by implicitly mapping data into a higher-dimensional feature space without explicitly computing the transformation. Examples of kernels include linear, polynomial, and RBF kernels.	2	CO2	BT1
5	Entropy measures the uncertainty or impurity in a dataset. For a binary variable: $H(X) = -plog_2(p) - (1 - p)log_2(1 - p)$ where p is the probability of one class and 1−p is the probability of the other class. Example: If p=0.5, entropy is maximum and equal to 1 bit.	2	CO5	BT1

PART B : Answer Any 1 Following Questions. 10 M

Q.N.	Questions	M	CO	BT

PTO (Please Turn Over)

6	Empirical Risk Minimization (ERM) is a fundamental learning principle in which a machine learning model is selected by minimizing its average loss on the available training data. The empirical risk is given by: $R_emp(f) = (1/n) * \sum_{i=1}^{n} L(y_i, f(x_i))$ where: n = number of training samples L = loss functiony i = actual output f(xi) = predicted output 1. Loss Function: A loss function measures the difference between the actual and predicted values. Examples include Mean Squared Error and Cross-Entropy Loss. 2. Empirical Risk: It is the average loss calculated over the training dataset. 3. Expected Risk: Expected risk represents the average error of the model over the entire underlying data distribution. R(f)=E[L(Y,f(X))] 4. ERM Principle: ERM chooses the model that minimizes empirical risk: $f* = argminf R_{emp}(f)$ 5. Overfitting: A highly complex model may achieve very low training error but perform poorly on unseen data. This is called overfitting. 6. Limitations: ERM primarily focuses on minimizing training error. It may lead to overfitting when: 1. Training data is limited. 2. The model is excessively complex. 3. The data contains noise. 4. Training and testing distributions differ. • • Therefore, ERM provides a basic framework for learning, but generalization and model complexity must also be considered.	10	CO1	BT2

ANSWER KEY

7	Principal Component Analysis (PCA) is a dimensionality reduction technique used to transform a dataset containing many correlated features into a smaller number of new variables called principal components. The major steps of PCA are: 1. Standardize the Data Features are generally standardized so that variables with larger numerical scales do not dominate the analysis. 2. Calculate the Covariance Matrix The covariance matrix is calculated to identify relationships between the features. 3. Calculate Eigenvalues and Eigenvectors Eigenvalues indicate the amount of variance captured by each principal component, while eigenvectors determine the direction of the principal components. 4. Rank the Principal Components Principal components are arranged in descending order according to their eigenvalues. 5. Select the Components The first few principal components that explain most of the variance are selected. 6. Transform the Data The original data is projected onto the selected principal components. For example, if a dataset has 10 original features and the first three principal components explain most of the variance, the dataset can be represented using only those three components. Thus, PCA reduces dimensionality while attempting to retain the most important information in the original dataset. It can also help reduce computational cost and make high-dimensional data easier to visualize.	10	CO1	BT2

PART C: Answer Any 1 Following Questions. 10 M				
Q.N.	**Questions**	**M**	**CO**	**BT**
8	Number of samples = 10 Class A = 6 Class B = 4 Therefore: p(A) = 6/10 = 0.6 p(B) = 4/10 = 0.4 The entropy formula is: $H(X) = -p(A)\log_2(p(A)) - p(B)\log_2(p(B))$ Substituting the values: $H(X) = -(0.6)\log_2(0.6) - (0.4)\log_2(0.4)$ Using: $\log_2(0.6) \approx -0.737$ $\log_2(0.4) \approx -1.322$ Therefore: $H(X) = -(0.6 \times -0.737) - (0.4 \times -1.322)$ $H(X) \approx 0.442 + 0.529$ $H(X) \approx 0.971$ bits Therefore, the entropy of the dataset is approximately: Entropy ≈ 0.971 bits Since the classes are relatively balanced (60% and 40%), the dataset has relatively high uncertainty.	10	CO5	BT3

ANSWER KEY

9	Given: Total samples = 20 Class A = 12 Class B = 8 **Step 1: Entropy of Original Dataset** p(A) = 12/20 = 0.6 p(B) = 8/20 = 0.4 **$H(S) = -0.6 \log_2(0.6) - 0.4 \log_2(0.4)$** Using: $\log_2(0.6) \approx -0.737$ $\log_2(0.4) \approx -1.322$ **Therefore:** $H(S) = -(0.6 \times -0.737) - (0.4 \times -1.322)$ $H(S) \approx 0.442 + 0.529$ $H(S) \approx 0.971$ bits --- **Step 2: Entropy of Subset 1** Subset 1: Class A = 8 Class B = 2 Total = 10 p(A) = 0.8 p(B) = 0.2 $H(S_1) = -0.8 \log_2(0.8) - 0.2 \log_2(0.2)$ $H(S_1) \approx 0.722$ bits --- **Step 3: Entropy of Subset 2** Subset 2: Class A = 4 Class B = 6 Total = 10 p(A) = 0.4 p(B) = 0.6 $H(S_2) = -0.4 \log_2(0.4) - 0.6 \log_2(0.6)$ $H(S_2) \approx 0.971$ bits --- **Step 4: Weighted Entropy** Both subsets contain 10 samples. Weighted Entropy = (10/20)(0.722) + (10/20)(0.971) Weighted Entropy = 0.361 + 0.486 Weighted Entropy ≈ 0.847 bits --- **Step 5: Information Gain** Information Gain = Parent Entropy − Weighted Child Entropy IG = 0.971 − 0.847 IG ≈ 0.124 bits Final Answer Information Gain ≈ 0.124 bits The split reduces the uncertainty from approximately **0.971 bits to 0.847 bits.**	10	CO5	BT3

PART D: Answer Any 1 Following Questions. 10 M				
Q.N.	**Questions**	**M**	**CO**	**BT**

ANSWER KEY

10	Support Vector Machine (SVM) is a supervised machine learning algorithm mainly used for classification. It attempts to find a decision boundary that separates different classes while maximizing the margin between them. 1. Decision Boundary: The decision boundary separates observations belonging to different classes. For a linear SVM: 2. Margin: The margin is the distance between the decision boundary and the nearest training points from each class. SVM attempts to maximize this margin. 3. Support Vectors: The training points closest to the decision boundary are called support vectors. They determine the position of the optimal decision boundary. 4. Maximum-Margin Principle: Among possible separating hyperplanes, SVM selects the one that provides the maximum separation between classes. 5. Non-linear Data: Some datasets cannot be separated using a straight line or hyperplane. 6. Kernel Trick: SVM uses kernels to implicitly transform data into a higher-dimensional feature space where the classes may become separable. Common kernels include: 1. Linear kernel 2. Polynomial kernel 3. Radial Basis Function (RBF) kernel Thus, SVM can handle both linearly and non-linearly separable classification problems while controlling the complexity of the decision boundary.	10	CO2	BT2

ANSWER KEY

11	Support Vector Machine (SVM) is a supervised machine learning algorithm mainly used for classification. It attempts to find a decision boundary that separates different classes while maximizing the margin between them. 1. Decision Boundary: The decision boundary separates observations belonging to different classes. For a linear SVM: 2. Margin: The margin is the distance between the decision boundary and the nearest training points from each class. SVM attempts to maximize this margin. 3. Support Vectors: The training points closest to the decision boundary are called support vectors. They determine the position of the optimal decision boundary. 4. Maximum-Margin Principle: Among possible separating hyperplanes, SVM selects the one that provides the maximum separation between classes. 5. Non-linear Data: Some datasets cannot be separated using a straight line or hyperplane. 6. Kernel Trick: SVM uses kernels to implicitly transform data into a higher-dimensional feature space where the classes may become separable. Common kernels include: 1. Linear kernel 2. Polynomial kernel 3. Radial Basis Function (RBF) kernel Thus, SVM can handle both linearly and non-linearly separable classification problems while controlling the complexity of the decision boundary.	10	CO2	BT2

PART E: Answer Any 1 Following Questions. 10 M				
Q.N.	**Questions**	**M**	**CO**	**BT**

ANSWER KEY

12	Entropy measures the amount of uncertainty or impurity present in a probability distribution. For a dataset with classes 1,2,...,C: $$H(X) = -sum_{i=1}^{C} p_i * log_2(p_i)$$ where pi is the probability of class i. Entropy Interpretation 1. If all samples belong to one class, entropy is 0. 2. If classes are equally distributed, entropy is higher. 3. Higher entropy indicates greater uncertainty. For binary classification: $$H(X) = -plog_2(p) - (1-p)log_2(1-p)$$ Cross-entropy measures the difference between the actual probability distribution and the predicted probability distribution. For multiclass classification: $$L = -sum_{i=1}^{C} y_i * log(y_hat_i)$$ where: yi = actual class probability y^i = predicted probability C = number of classes For a correctly classified sample, the predicted probability for the actual class should be close to 1, resulting in a low loss. If the model assigns a very low probability to the actual class, the cross-entropy loss becomes high. Importance Cross-entropy is widely used for classification because it: 1. Measures the quality of predicted probabilities. 2. Penalizes confident incorrect predictions strongly. 3. Can be optimized during model training. 4. Is commonly used with Logistic Regression and neural-network classifiers. Therefore, entropy measures uncertainty, whereas cross-entropy measures the difference between actual and predicted probability distributions.	10	CO1	BT2

ANSWER KEY

13	Structural Risk Minimization (SRM) is a learning principle that aims to minimize the risk of a model while controlling its complexity. In ERM, the model mainly attempts to minimize the training error. However, a highly complex model may fit the training data very closely and result in overfitting. SRM addresses this problem by considering both: 1. Training error 2. Model complexity The general idea can be represented as: **Total Objective = Training Error + Complexity Penalty** A simpler model generally has lower complexity, while a more complex model has greater capacity to fit the training data. Regularization is commonly used to control model complexity. For example: L1 regularization: **Objective = Loss + $\lambda \Sigma \lvert w_i \rvert$** L2 regularization: **Objective = Loss + $\lambda \Sigma w_i^2$** where λ controls the strength of regularization. If λ is too small, the model may become overly complex and overfit. If λ is too large, the model may become too simple and underfit. Thus, SRM attempts to achieve a balance between fitting the training data and maintaining a model that can generalize well to unseen data.	10	CO1	BT2

ANSWER KEY