Guru Tegh Bahadur Institute of Technology, New Delhi
Department of Computer Science & Engineering

QUESTION BANK

SUBJECT: DATA SCIENCE

Semester: 7th **SUB CODE: CIE 405T**

Unit I: Introduction to Data Science and Data Pre-processing

1. **What is Data Science?**
   Answer: Data Science is an interdisciplinary field that uses scientific methods, algorithms, and systems to extract insights and knowledge from structured and unstructured data.

2. **How is Data Science different from AI and Machine Learning?**
   Answer: AI is the broader concept of machines performing tasks in a way that we consider intelligent. Machine Learning is a subset of AI focusing on algorithms that learn from data. Data Science involves the collection, processing, and analysis of data using various techniques, including Machine Learning.

3. **What are some popular dataset repositories?**
   Answer: Popular dataset repositories include UCI Machine Learning Repository, Kaggle, and Google Dataset Search.

4. **What are the main steps in data pre-processing?**
   Answer: The main steps include data cleaning (handling missing values), normalization, scaling, encoding categorical variables, and data transformation (e.g., log transformations).

5. **What is the importance of data scaling in pre-processing?**
   Answer: Data scaling ensures that features have comparable ranges, which can significantly improve the performance of machine learning algorithms that rely on distance-based measures, such as KNN or SVM.

6. **What is similarity and dissimilarity in data analysis?**
   Answer: Similarity measures quantify how alike two data objects are, while dissimilarity measures how different they are. Common similarity measures include Euclidean distance and cosine similarity.

7. **What is Principal Component Analysis (PCA)?**
   Answer: PCA is a dimensionality reduction technique that transforms high-dimensional data into a lower-dimensional form while retaining the most important variance in the data.

8. **Why is data visualization important in Data Science?**
   Answer: Data visualization helps in understanding data patterns, relationships, and trends. It also aids in communicating findings clearly and effectively.

9. **What is the Chi-Square test?**
   Answer: The Chi-Square test is a statistical method used to determine if there is a significant association between categorical variables.

10. **What is data transformation, and why is it needed?**
    Answer: Data transformation involves converting data into an appropriate format or structure for analysis. It helps to make the data more interpretable and improves model performance.

Unit II: Regression Analysis

11. **What is regression analysis?**
    Answer: Regression analysis is a statistical technique used to model and analyze the relationship between a dependent variable and one or more independent variables.

12. **What is linear regression?**
    Answer: Linear regression is a method to model the relationship between a dependent variable and one or more independent variables assuming a linear relationship.

13. **What is the difference between linear and generalized regression?**
    Answer: Linear regression assumes a linear relationship between the variables, while generalized regression extends to model more complex relationships (e.g., logistic regression for binary outcomes).

14. **What is regularized regression?**
    Answer: Regularized regression adds a penalty term to the loss function to prevent overfitting. Examples include Ridge and Lasso regression.

15. **What is cross-validation, and why is it important?**
    Answer: Cross-validation is a technique to evaluate model performance by partitioning data into training and testing sets multiple times. It helps ensure that the model generalizes well to unseen data.

16. **What is the purpose of dividing data into training and testing sets?**
    Answer: Splitting data into training and testing sets allows the model to learn from one part and be evaluated on another to check its generalization ability.

17. **What is Ridge regression?**
    Answer: Ridge regression is a type of regularized regression that adds a penalty equal to the square of the magnitude of coefficients, reducing overfitting.

18. **What are latent variables in regression analysis?**
    Answer: Latent variables are variables that are not directly observed but are inferred from the model, typically affecting the observed variables.

19. **What is Structural Equation Modeling (SEM)?**
    Answer: SEM is a multivariate statistical analysis technique that models complex relationships between observed and latent variables.

20. **What is the difference between regular and Ridge regression?**
    Answer: Ridge regression adds an L2 penalty to the cost function, which helps reduce overfitting by shrinking the coefficients, whereas regular regression does not include this penalty.

Unit III: Time Series Analysis and Forecasting

21. **What is time series data?**
    Answer: Time series data consists of observations collected or recorded at regular time intervals, such as daily stock prices or monthly sales.

22. **What is stationarity in time series data?**
    Answer: Stationarity refers to a time series whose statistical properties, like mean and variance, do not change over time.

23. **What are the common techniques to test for stationarity?**
    Answer: Common techniques include the Augmented Dickey-Fuller (ADF) test, KPSS test, and visual inspection of the data.

24. **What is seasonality in time series analysis?**
    Answer: Seasonality refers to patterns that repeat at regular intervals over time, such as daily, monthly, or yearly cycles in a dataset.

25. **What is an autoregressive (AR) model?**
    Answer: An AR model is a type of time series model where the current value is regressed on its previous values.

26. **What is the difference between AR and MA models?**
    Answer: AR models use past values to predict future ones, while MA (Moving Average) models use past forecast errors.

27. **What are recurrent models in time series forecasting?**
    Answer: Recurrent models, such as RNNs and LSTMs, are neural network architectures that handle sequences and capture temporal dependencies.

28. **What is the purpose of time series decomposition?**
    Answer: Time series decomposition breaks a series into its underlying components, such as trend, seasonality, and residual, to better understand and model the data.

29. **How can we evaluate the performance of time series models?**
    Answer: Common metrics include Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and Mean Absolute Percentage Error (MAPE).

30. **What is the difference between univariate and multivariate time series?**
    Answer: A univariate time series contains a single variable observed over time, while a multivariate time series includes multiple variables.

Unit IV: Classification and Clustering

31. **What is classification in Data Science?**
    Answer: Classification is a supervised learning task where the objective is to predict a categorical label for a given input based on training data.

32. **What is Linear Discriminant Analysis (LDA)?**
    Answer: LDA is a classification technique that finds a linear combination of features that best separates two or more classes.

33. **What is a Support Vector Machine (SVM)?**
    Answer: SVM is a supervised learning algorithm used for classification and regression tasks, which finds the hyperplane that best separates classes in the feature space.

34. **What are decision trees?**
    Answer: Decision trees are tree-like models used for classification or regression, where the data is split based on feature values to make decisions.

35. **What is the difference between classification and regression?**
    Answer: Classification predicts categorical outcomes, while regression predicts continuous values.

36. **What is clustering in data analysis?**
    Answer: Clustering is an unsupervised learning technique that groups similar data points together based on certain criteria, without prior labels.

37. **What is K-Means clustering?**
    Answer: K-Means is a clustering algorithm that partitions data into K distinct clusters based on minimizing the variance within clusters.

38. **What are the strengths and weaknesses of K-Means?**
    Answer: Strengths include its simplicity and scalability. Weaknesses include sensitivity to initial centroid selection and difficulty handling non-spherical clusters.

39. **What is hierarchical clustering?**
    Answer: Hierarchical clustering builds a hierarchy of clusters, either agglomeratively (bottom-up) or divisively (top-down), without specifying the number of clusters in advance.

40. **What is DBSCAN, and how does it differ from K-Means?**
    Answer: DBSCAN is a density-based clustering algorithm that can identify arbitrarily shaped clusters and handle outliers. Unlike K-Means, it does not require specifying the number of clusters.

41. **What is the silhouette score?**
    Answer: The silhouette score is a measure of how similar an object is to its own cluster compared to other clusters. It ranges from -1 to 1, with higher values indicating better clustering.

42. **What are the key applications of classification techniques?**
    Answer: Common applications include spam detection, image recognition, and medical diagnosis.

43. **How is performance evaluated in classification models?**
    Answer: Common evaluation metrics include accuracy, precision, recall, F1-score, and AUC-ROC.

44. **What are the main differences between supervised and unsupervised learning?**
    Answer: Supervised learning uses labeled data to train models, while unsupervised learning finds patterns in unlabeled data.

45. **What is overfitting in classification models?**
    Answer: Overfitting occurs when a model learns the noise in the training data, performing well on training but poorly on unseen data.

46. **What is underfitting in machine learning?**
    Answer: Underfitting happens when a model is too simple to capture the underlying pattern in the data, leading to poor performance on both training and testing data.

47. **What are decision boundaries in classification?**
    Answer: Decision boundaries are the regions in the feature space where the decision changes from one class to another in a classification model.

48. **What is an ensemble method in classification?**
    Answer: Ensemble methods combine multiple models to improve overall prediction performance. Examples include Random Forest and Gradient Boosting.

49. **What is bagging in ensemble methods?**
    Answer: Bagging (Bootstrap Aggregating) involves training multiple models on random subsets of the data and averaging their predictions to reduce variance.

50. **What is boosting in machine learning?**
    Answer: Boosting sequentially trains models, where each model attempts to correct the errors of the previous one, to improve accuracy and reduce bias.