MUSA (MUMBAI UNIVERSITY STUDENTS ASSOCIATION)

VIVA QUESTION AND ANSWER OF MACHINE LEARNING

B.E SEM-VII **SCHEME: R-19 C SCHEME** **FOR WINTER SESSION 2025**

BRANCH: COMPS

Module 1 – Introduction to Machine Learning

Q1) What is Machine Learning (ML)?
Ans: ML is a field of AI that enables systems to learn from data, identify patterns, and make decisions without explicit programming.

Q2) What are the types of ML?
Ans: Three main types: Supervised (with labeled data), Unsupervised (finding patterns without labels), and Reinforcement Learning (learning via rewards and penalties).

Q3) What are the common issues in ML?
Ans: Issues include overfitting, underfitting, insufficient data, noisy data, high dimensionality, and model bias or variance.

Q4) What is overfitting and underfitting?
Ans: Overfitting occurs when a model fits training data too well but fails on new data. Underfitting occurs when a model is too simple to capture patterns in data.

Q5) Explain Bias-Variance trade-off.
Ans: Bias is error due to oversimplification; variance is error due to sensitivity to data fluctuations. Trade-off balances model complexity and generalization.

Q6) What is training error vs generalization error?
Ans: Training error is the error on training data, while generalization error is the model's error on unseen data.

Q7) What are the steps to develop a ML application?
Ans: Steps include problem definition, data collection, preprocessing, model selection, training, evaluation, and deployment.

Q8) Name some applications of ML.
Ans: Applications include speech recognition, image classification, recommendation systems, fraud detection, and autonomous vehicles.

Q9) How to prevent overfitting?
Ans: Use regularization, cross-validation, more data, simpler models, or ensemble techniques.

Q10) What is the difference between ML and traditional programming?
Ans: Traditional programming requires explicit rules; ML learns patterns and rules from data automatically.

Module 2 – Learning with Regression and Trees

Q1) What is Linear Regression?
Ans: A regression technique modeling the relationship between input variables and output by fitting a straight line using least squares.

Q2) What is Multivariate Linear Regression?
Ans: Extension of linear regression using multiple input variables to predict a single output variable.

Q3) Explain Logistic Regression.
Ans: Used for binary classification; models the probability of the output using the sigmoid function to map predictions between 0 and 1.

Q4) What is a Decision Tree?
Ans: Tree-structured model for classification or regression; splits data based on feature values to reach a decision at leaf nodes.

Q5) What is Gini Index?
Ans: Metric to measure impurity of a dataset split; lower Gini indicates better homogeneity in nodes.

Q6) What is CART?
Ans: Classification and Regression Tree; algorithm for constructing decision trees using splitting criteria like Gini Index or variance reduction.

Q7) What is a Confusion Matrix?
Ans: Table summarizing model performance by showing true positives, true negatives, false positives, and false negatives.

Q8) Define Precision and Recall.
Ans: Precision is the ratio of true positives to all predicted positives; Recall is the ratio of true positives to all actual positives.

Q9) What is F-measure?
Ans: Harmonic mean of precision and recall; combines both metrics for model performance evaluation.

Q10) What is ROC curve?
Ans: Receiver Operating Characteristic curve plots true positive rate vs. false positive rate, useful for evaluating classifier thresholds.

Module 3 – Ensemble Learning

Q1) What is ensemble learning?
Ans: Combines multiple models to improve accuracy, reduce variance, and prevent overfitting.

Q2) Explain K-fold cross-validation.
Ans: Data is split into K subsets; model trains on K-1 folds and validates on the remaining fold, repeated K times to evaluate performance.

Q3) What is Boosting?
Ans: Sequentially trains weak learners, giving higher weight to misclassified samples to improve accuracy.

Q4) What is Bagging?
Ans: Bootstrap Aggregating; trains multiple models on random subsets of data and averages predictions to reduce variance.

Q5) What is Random Forest?
Ans: Ensemble of decision trees using bagging and random feature selection for robust classification or regression.

Q6) What is Stumping?
Ans: Using very simple decision trees (stumps) as base learners in ensemble methods like boosting.

Q7) What is XGBoost?
Ans: Extreme Gradient Boosting; optimized boosting algorithm using gradient descent and regularization to improve accuracy.

Q8) What is Subagging?
Ans: Subsample Aggregation; variant of bagging using smaller data subsets for training base models.

Q9) Difference between Bagging and Boosting?
Ans: Bagging reduces variance by averaging multiple independent models; boosting reduces bias by sequentially improving weak models.

Q10) How are ensemble predictions combined?
Ans: Using majority voting for classification, averaging for regression, or weighted combination of models.

Module 4 – Learning with Classification

Q1) What is Support Vector Machine (SVM)?
Ans: SVM is a supervised learning algorithm that finds the optimal hyperplane separating classes with maximum margin.

Q2) What is a support vector?
Ans: Data points closest to the hyperplane that determine its position and margin in SVM.

Q3) Explain SVM as constrained optimization.
Ans: SVM maximizes margin while minimizing classification error, formulated as a quadratic optimization problem with constraints.

Q4) What is kernel trick?
Ans: Technique to map input data to higher dimensions to make it linearly separable without explicitly computing the transformation.

Q5) What is Support Vector Regression (SVR)?
Ans: Extension of SVM for regression; predicts continuous outputs by fitting within a margin of tolerance around data points.

Q6) How does SVM handle nonlinear classification?
Ans: Uses kernel functions (linear, polynomial, RBF) to project data into higher-dimensional space for linear separation.

Q7) What is multiclass classification?
Ans: Problem of classifying data into more than two categories using strategies like one-vs-rest or one-vs-one SVM.

Q8) Define margin in SVM.
Ans: Distance between the hyperplane and nearest data points from each class; larger margin gives better generalization.

Q9) What are common applications of SVM?
Ans: Text classification, image recognition, bioinformatics, and fraud detection.

Q10) Difference between SVM and logistic regression?
Ans: Logistic regression predicts probabilities; SVM finds the optimal separating hyperplane and focuses on support vectors.

Module 5 – Learning with Clustering

Q1) What is clustering?
Ans: Unsupervised technique to group similar data points into clusters based on distance or similarity metrics.

Q2) Name common distance metrics.
Ans: Euclidean, Manhattan, Cosine similarity, and Mahalanobis distance.

Q3) What is Graph-based clustering?
Ans: Forms clusters using graph structures like minimum spanning tree to find connectivity-based groups.

Q4) What is Model-based clustering?
Ans: Assumes data is generated from probabilistic models; Expectation-Maximization (EM) algorithm estimates parameters and assigns clusters.

Q5) Explain DBSCAN.
Ans: Density-based clustering groups points densely packed and marks sparse points as outliers; handles arbitrary shapes well.

Q6) Advantages of clustering?
Ans: Discover hidden patterns, reduce dimensionality, detect anomalies, and aid in exploratory data analysis.

Q7) Difference between hierarchical and partition clustering?
Ans: Hierarchical builds nested clusters; partition clustering (like K-means) divides data into non-overlapping clusters directly.

Q8) What is cluster centroid?
Ans: Representative point (usually mean) of a cluster used to assign and evaluate points in clustering algorithms.

Q9) What is silhouette score?
Ans: Measures how similar a point is to its own cluster compared to other clusters; evaluates clustering quality.

Q10) Applications of clustering?

Ans: Market segmentation, image segmentation, anomaly detection, document classification, and bioinformatics.

Module 6 – Dimensionality Reduction

Q1) Why use dimensionality reduction?

Ans: Reduces features, avoids overfitting, improves computation, and helps visualization while retaining essential information.

Q2) What is PCA?

Ans: Principal Component Analysis transforms data into orthogonal components capturing maximum variance.

Q3) What is LDA?

Ans: Linear Discriminant Analysis finds a projection that maximizes class separability for supervised dimensionality reduction.

Q4) What is SVD?

Ans: Singular Value Decomposition decomposes matrix into U, Σ, V matrices; used in data compression, PCA, and latent feature extraction.

Q5) Difference between PCA and LDA?

Ans: PCA is unsupervised, focuses on variance; LDA is supervised, focuses on maximizing class separability.

Q6) Applications of dimensionality reduction?

Ans: Feature extraction, noise reduction, visualization, image compression, and speedup in ML algorithms.

Q7) What is eigenvalue in PCA?

Ans: Indicates variance captured by each principal component; higher eigenvalue means more importance.

Q8) How many components to select in PCA?

Ans: Select top components capturing 90–95% of total variance to retain most information.

Q9) How does dimensionality reduction help ML models?

Ans: Reduces overfitting, improves training speed, lowers memory usage, and simplifies models.

Q10) Difference between linear and nonlinear dimensionality reduction?

Ans: Linear techniques (PCA, LDA) assume linear relationships; nonlinear (t-SNE, Isomap) capture complex manifold structures.

*******ALL THE BEST*******

MUSA (MUMBAI UNIVERSITY STUDENTS ASSOCIATION)

VIVA QUESTION AND ANSWER OF BIG DATA ALALYSIS

B.E SEM-VII **SCHEME: R-19 C SCHEME** **FOR WINTER SESSION 2025**

BRANCH: COMPS

Module 1 – Introduction to Big Data and Hadoop

Q1) What is Big Data?
Ans: Big Data refers to extremely large datasets that cannot be processed using traditional methods, characterized by volume, velocity, variety, veracity, and value.

Q2) What are the types of Big Data?
Ans: Structured (tables), Semi-structured (XML, JSON), and Unstructured (text, images, videos).

Q3) How does Big Data differ from traditional business data?
Ans: Big Data involves large, fast-changing, and diverse datasets, requiring distributed storage and parallel processing, unlike traditional static relational databases.

Q4) What is Hadoop?
Ans: Hadoop is an open-source framework for distributed storage and processing of big data using clusters of commodity hardware.

Q5) What are the core components of Hadoop?
Ans: Hadoop Distributed File System (HDFS) and MapReduce framework; ecosystem includes YARN, Hive, Pig, HBase, and Spark.

Q6) What is HDFS?
Ans: HDFS is Hadoop's distributed file system that stores large files across multiple nodes, providing high throughput and fault tolerance.

Q7) What is the Hadoop ecosystem?
Ans: Collection of tools and frameworks like Hive, Pig, HBase, Flume, Sqoop, and Spark to process, query, and manage Big Data efficiently.

Q8) Give an example of a Big Data application.
Ans: Social media analytics, e-commerce recommendations, fraud detection in banking, and sensor data from IoT devices.

Q9) What is the advantage of Hadoop over traditional systems?
Ans: Handles huge data efficiently, provides fault tolerance, scalable, cost-effective using commodity hardware.

Q10) Name some challenges in Big Data.
Ans: Data privacy, storage management, processing speed, integration of heterogeneous data, and real-time analytics.

Module 2 – Hadoop HDFS and MapReduce

Q1) What is a distributed file system?

Ans: System that stores data across multiple nodes for parallel access, redundancy, and fault tolerance.

Q2) Explain MapReduce.

Ans: Programming model for processing large datasets using two functions: Map (process and produce key-value pairs) and Reduce (aggregate results).

Q3) What is a combiner in MapReduce?

Ans: Optional function that performs local aggregation of Map outputs to reduce data transfer to the Reduce phase.

Q4) How does MapReduce handle node failures?

Ans: By re-executing failed tasks on other nodes, leveraging data replication in HDFS.

Q5) Give an example algorithm using MapReduce.

Ans: Word count, matrix-vector multiplication, computing projections, union, intersection, and difference operations.

Q6) What is the role of HDFS in MapReduce?

Ans: HDFS stores large datasets in blocks across nodes, enabling parallel MapReduce processing.

Q7) What are the limitations of Hadoop?

Ans: High latency, not ideal for real-time processing, complex programming, limited iterative processing efficiency.

Q8) How are Map and Reduce tasks scheduled?

Ans: Hadoop schedules tasks on nodes holding the data (data locality) to minimize network transfer and improve efficiency.

Q9) What is the purpose of grouping by key?

Ans: Ensures that all values for a given key are sent to the same reducer for aggregation.

Q10) Difference between HDFS and traditional file system?

Ans: HDFS is distributed, fault-tolerant, stores large files, optimized for high throughput; traditional FS is local and not fault-tolerant.

Module 3 – NoSQL

Q1) What is NoSQL?

Ans: NoSQL refers to non-relational databases designed for scalable, distributed, and flexible data storage for Big Data.

Q2) Name NoSQL data models.

Ans: Key-value stores, Document stores, Column-family stores, Graph databases.

Q3) Why use NoSQL for Big Data?

Ans: Handles large volumes of unstructured or semi-structured data, provides horizontal scaling, and flexible schema.

Q4) What are NoSQL business drivers?

Ans: Scalability, flexibility, faster data access, distributed architecture, and handling diverse data types.

Q5) Explain key-value store.

Ans: Stores data as a collection of key-value pairs; simple, fast, and suitable for caching or session storage.

Q6) What is a column-family store?

Ans: Stores data in columns rather than rows, suitable for analytical queries on large datasets (example: HBase, Bigtable).

Q7) What is a document store?

Ans: Stores semi-structured data like JSON or XML documents, supports queries on document attributes (example: MongoDB).

Q8) Explain graph databases.

Ans: Stores entities as nodes and relationships as edges; useful for social networks and recommendation systems.

Q9) Difference between master-slave and peer-to-peer distribution.

Ans: Master-slave has centralized control, peer-to-peer is fully distributed; peer-to-peer is more fault-tolerant.

Q10) Give a case study of NoSQL in Big Data.

Ans: Facebook uses Cassandra (column-family store) for handling large-scale messaging and social graph data.

Module 4 – Mining Data Streams

Q1) What is a data stream?

Ans: Continuous, rapid, and time-varying sequence of data generated by sources like sensors, social media, or network logs.

Q2) What is a Data-Stream-Management System?

Ans: System designed to query and process continuous data streams in real-time or near real-time.

Q3) What is a Bloom Filter?

Ans: Probabilistic data structure to test set membership efficiently; may have false positives but no false negatives.

Q4) Explain Flajolet-Martin algorithm.

Ans: Estimates the number of distinct elements in a data stream using probabilistic counting and hash functions.

Q5) What is DGIM algorithm?

Ans: Datar-Gionis-Indyk-Motwani algorithm approximates counts of 1's in a sliding window efficiently using limited memory.

Q6) What is sampling in streams?

Ans: Technique to reduce data volume by selecting representative data points for processing.

Q7) What are the issues in stream processing?

Ans: Limited memory, high velocity, approximate computation, windowing, and query efficiency.

Q8) How to count distinct elements in a stream?

Ans: Using Flajolet-Martin algorithm or other probabilistic methods to estimate unique items with low memory.

Q9) What is a decaying window?
Ans: Window where older data gradually loses weight, emphasizing recent data in streaming analytics.

Q10) Give an example of stream query.
Ans: Counting the number of clicks on a website in the last 10 minutes.

Module 5 – Real-Time Big Data Models

Q1) What is a recommendation system?
Ans: System predicting user preferences for products or services using past interactions or behavior patterns.

Q2) Explain content-based recommendations.
Ans: Recommends items similar to what the user liked previously based on item features.

Q3) Explain collaborative filtering.
Ans: Predicts user preferences by analyzing patterns from other users with similar behavior.

Q4) Give a case study of a product recommendation system.
Ans: Amazon recommends products by combining collaborative filtering and content-based methods using user activity.

Q5) How is a social network represented as a graph?
Ans: Users are nodes, relationships/friendships are edges; enables analysis of communities and influence.

Q6) What is clustering in social networks?
Ans: Groups users with dense connections; used for community detection or targeted marketing.

Q7) How to detect communities in a social graph?
Ans: Using graph algorithms like modularity maximization, spectral clustering, or label propagation.

Q8) What are the benefits of real-time Big Data analytics?
Ans: Immediate insights, timely recommendations, fraud detection, and adaptive decision-making.

Q9) Difference between batch and real-time analytics?
Ans: Batch processes large datasets periodically; real-time processes continuous streams instantly.

Q10) Give an example of real-time Big Data model.
Ans: Twitter trending topics analysis or Netflix content recommendation.

Module 6 – Data Analytics with R

Q1) What is R?
Ans: R is a programming language and environment for statistical computing, data manipulation, and visualization.

Q2) What are RGUI and RStudio?
Ans: RGUI is the basic interface; RStudio is an IDE with features for code editing, debugging, and visualization.

Q3) How to handle variables in R?

Ans: Variables store data; created dynamically and can be numeric, character, logical, or vector types.

Q4) How to read datasets in R?

Ans: Using functions like read.csv(), read.table(), or packages like readr.

Q5) How to export data from R?

Ans: Using write.csv(), write.table(), or saveRDS() functions for persistence.

Q6) How to visualize data in R?

Ans: Using functions like plot(), hist(), boxplot(), or libraries like ggplot2 for advanced visualization.

Q7) Difference between vectors and objects in R?

Ans: Vectors are basic data structures; objects can be lists, data frames, or user-defined complex structures.

Q8) How to create a function in R?

Ans: Using function() keyword; encapsulates code for reuse and modular programming.

Q9) What are data visualization applications?

Ans: Trend analysis, anomaly detection, reporting, dashboarding, and decision support.

Q10) How to manipulate data in R?

Ans: Using functions like subset(), merge(), transform(), apply(), and dplyr package for filtering and aggregation.

*****ALL THE BEST*****