UNIT-1

Viva Questions

1. **Q:** What is Machine Learning?
   A: Machine Learning is a field of Artificial Intelligence that enables systems to learn from data and improve performance over time without being explicitly programmed.

2. **Q:** What are the three main paradigms of Machine Learning?
   A: Supervised Learning, Unsupervised Learning, and Reinforcement Learning.

3. **Q:** What is Learning by Rote?
   A: Learning by Rote refers to memorizing data or examples without generalizing patterns; similar to storing answers without understanding.

4. **Q:** Define Reinforcement Learning.
   A: It's a learning paradigm where an agent learns to take actions in an environment to maximize a reward signal.

5. **Q:** What types of data are commonly used in ML?
   A: Numerical data, categorical data, textual data, image data, audio data, and time-series data.

6. **Q:** Explain the stages in a Machine Learning pipeline.
   A: The stages include:
   - Data Acquisition
   - Feature Engineering
   - Data Representation
   - Model Selection
   - Model Training
   - Model Evaluation
   - Prediction

7. **Q:** How does Learning by Induction differ from Learning by Rote?
   A: Learning by Induction involves generalizing from examples to find patterns, while Rote Learning memorizes specific instances without generalization.

8. **Q:** Why is Feature Engineering important in Machine Learning?
   A: It transforms raw data into meaningful features that help improve the performance and accuracy of machine learning models.

9. **Q:** What is Model Evaluation in Machine Learning?
   A: It’s the process of assessing the model's performance using metrics like accuracy, precision, recall, F1-score, and cross-validation.

10. **Q:** How do Machine Learning models make predictions?
    A: After learning patterns from training data, models use those patterns to make predictions on unseen (test) data.

Unit-2

1. **Q1: What is a proximity measure in machine learning?**
   A: A proximity measure quantifies how close or similar two data points are. It is crucial in nearest neighbor algorithms for identifying the most similar instances.
2. **Q2: Name common distance measures used in nearest neighbor models.**
   A: The most common distance measures are:
   - **Euclidean Distance**
   - **Manhattan Distance**
   - **Minkowski Distance**
   - **Cosine Similarity**
3. **Q3: What is a non-metric similarity function? Give an example.**
   A: A non-metric similarity function does not satisfy the properties of a metric (like the triangle inequality). An example is **Jaccard Similarity**, often used for binary or categorical data.
4. **Q4: How do you compute proximity between binary patterns?**
   A: Proximity between binary patterns is computed using measures like:
   - **Hamming Distance**
   - **Jaccard Index**
   - **Simple Matching Coefficient**
5. **Q5: What is the basic working principle of the K-Nearest Neighbor (KNN) algorithm?**
   A: KNN classifies a data point based on the majority class among its **k closest neighbors** in the feature space using a distance metric.

6. **Q6: What is the main difference between KNN classification and KNN regression?**
   A: In **KNN classification**, the class label is predicted based on majority voting. In **KNN regression**, the target value is predicted as the **average of the k nearest neighbors' outputs**.
7. **Q7: What is the Radius Distance Nearest Neighbor algorithm?**
   A: Instead of a fixed number of neighbors (k), RDNN includes all neighbors **within a specified radius** around a query point for classification or regression.
8. **Q8: What are the performance metrics used to evaluate classifiers?**
   A: Common metrics include:
   - **Accuracy**
   - **Precision**
   - **Recall**
   - **F1-score**
   - **Confusion Matrix**
9. **Q9: How do we evaluate the performance of regression algorithms?**
   A: Using metrics like:
   - **Mean Absolute Error (MAE)**
   - **Mean Squared Error (MSE)**
   - **Root Mean Squared Error (RMSE)**
   - **R-squared (R^2)**
10. **Q10: Why is feature scaling important in KNN models?**
    A: KNN relies on distance measures, so unscaled features can dominate distance calculations. **Standardization or normalization** ensures all features contribute equally.

Unit-3

1. **Q1: What is a Decision Tree in machine learning?**
   A: A Decision Tree is a flowchart-like model used for classification and regression. It splits data into subsets based on feature values to arrive at a decision.

2. **Q2: What is the purpose of impurity measures in decision trees?**
   A: Impurity measures determine the quality of a split. They help select the best attribute to split the data. Lower impurity means better homogeneity in child nodes.

3. **Q3: Name common impurity measures used in Decision Trees.**
   A: The most common impurity measures are:
   - **Gini Index**
   - **Entropy (Information Gain)**
   - **Classification Error**

4. **Q4: What are the key properties of Decision Trees?**
   A: Key properties include:
   - Easy to interpret and visualize
   - Handles both numerical and categorical data
   - Prone to overfitting
   - Non-parametric and recursive

5. **Q5: How are Decision Trees used for regression?**
   A: In regression, Decision Trees split the data based on features to minimize **variance** in output values. The prediction is usually the **mean of the target values** in a leaf node.

6. **Q6: What is the Bias-Variance Trade-off in Decision Trees?**
   A: A deep tree can have **low bias but high variance** (overfitting), while a shallow tree has **high bias but low variance** (underfitting). The trade-off is about finding a balance for optimal performance.

7. **Q7: What is pruning in Decision Trees, and why is it important?**
   A: Pruning removes unnecessary branches from a tree to reduce **overfitting** and improve generalization by simplifying the model.

8. **Q8: What is a Random Forest? How does it differ from a single Decision Tree?**
   A: A Random Forest is an ensemble of Decision Trees. It improves accuracy and reduces overfitting by combining multiple trees using bagging and feature randomness.

9. **Q9: How is a Random Forest used for classification and regression?**
   A: For **classification**, Random Forest uses majority voting among trees. For **regression**, it takes the average of the predicted values from all trees.

10. **Q10: What are the advantages of using Random Forest over a single Decision Tree?**
    A: Advantages include:

- Better generalization
- Higher accuracy
- Robustness to noise and outliers
- Less prone to overfitting

Unit-4

1. Q: What is a Linear Discriminant in Machine Learning? (K1)
A: A linear discriminant is a decision boundary used to separate different classes by fitting a linear function to the input features. It is commonly used for binary classification.

2. Q: What is the Perceptron Classifier? (K1)
A: The Perceptron is the simplest type of feedforward neural network and a linear classifier. It classifies data by computing a weighted sum of input features and applying an activation function.

3. Q: Explain the Perceptron Learning Algorithm. (K2)
A: The Perceptron Learning Algorithm updates the weights whenever a misclassification occurs. The update rule is:
$w = w + \eta * (\text{target} - \text{prediction}) * x$,
where η is the learning rate, and x is the input vector.

4. Q: What is the limitation of the Perceptron? (K1)
A: The Perceptron can only classify **linearly separable** data. It fails to learn when the data is not linearly separable, such as in the XOR problem.

5. Q: What is Support Vector Machine (SVM)? (K1)
A: SVM is a supervised learning algorithm that finds the optimal hyperplane which maximizes the margin between different classes in the dataset.

6. Q: What is the Kernel Trick in SVM? (K2)
A: The kernel trick transforms non-linearly separable data into a higher-dimensional space where it becomes linearly separable, without explicitly computing the transformation.

7. Q: How does Logistic Regression work as a classifier? (K2)
A: Logistic Regression models the probability of a binary outcome using a sigmoid function. It outputs values between 0 and 1, indicating the probability of a class.

8. Q: Differentiate between Linear Regression and Logistic Regression. (K2)
A:

- **Linear Regression** predicts continuous values.
- **Logistic Regression** predicts class probabilities for classification tasks.

9. Q: What is a Multi-Layer Perceptron (MLP)? (K1)
A: MLP is a feedforward neural network with one or more hidden layers. It uses non-linear activation functions and is capable of learning complex patterns.

10. Q: What is Backpropagation and how is it used in training an MLP? (K2)
A: Backpropagation is an algorithm used to train MLPs by computing gradients of the loss function and updating weights using gradient descent to minimize error.

Unit-5

1. Q: What is Clustering in Machine Learning?
A: Clustering is an unsupervised learning technique used to group similar data points into clusters. The objective is to minimize intra-cluster distances and maximize inter-cluster distances.

2. Q: Explain the concept of Partitioning of Data in Clustering.
A: Partitioning of data involves dividing the data set into distinct clusters based on some criteria. Each data point belongs to exactly one cluster. An example of a partitioning algorithm is **K-Means**.

3. Q: What is Matrix Factorization in Clustering?
A: Matrix factorization decomposes a large matrix into multiple smaller matrices. In clustering, it's used to reduce the dimensionality of data, making it easier to identify hidden patterns or clusters.

---

4. Q: Describe the difference between Divisive and Agglomerative Clustering.
A:

- **Divisive Clustering** is a top-down approach where all data points start in one cluster, and the cluster is recursively split into smaller clusters.
- **Agglomerative Clustering** is a bottom-up approach where each data point starts as a separate cluster, and pairs of clusters are merged iteratively.

5. Q: What is Partitional Clustering?
A: Partitional clustering divides the data into a pre-defined number of clusters. A classic example is **K-Means clustering**, where the number of clusters is specified beforehand.

6. Q: Explain the K-Means Clustering Algorithm.
A: K-Means is an iterative algorithm that divides a dataset into K clusters by minimizing the variance within each cluster. The process involves randomly initializing K centroids, assigning points to the nearest centroid, and then recalculating the centroids until convergence.

7. Q: What is Soft Partitioning in Clustering?
A: Soft partitioning allows data points to belong to multiple clusters with different membership levels (i.e., probabilistic assignment). This contrasts with hard partitioning, where each point belongs to exactly one cluster.

8. Q: Explain the Fuzzy C-Means (FCM) Clustering.
A: Fuzzy C-Means (FCM) is a soft clustering algorithm where each data point can belong to multiple clusters with a membership value between 0 and 1. It aims to minimize the weighted within-cluster sum of squared distances.

9. Q: What is Rough Clustering? (K1)
A: Rough clustering is based on the concept of rough set theory, where data points may belong to a set partially. It provides a way to handle uncertainty and imprecision in cluster assignments.

10. Q: Describe the Expectation Maximization (EM) Algorithm for Clustering. (K2)
A: Expectation Maximization (EM) is an iterative algorithm for clustering that alternates between assigning data points to clusters (Expectation step) and updating the parameters of the clusters (Maximization step). It's commonly used for Gaussian Mixture Models (GMMs).