PART A

Unit 1

1) Define Machine Learning?

Machine learning is a subset of artificial intelligence (AI) that focuses on building systems that can learn from and make decisions based on data. Instead of being explicitly programmed to perform a task, these systems use algorithms to identify patterns in data and make predictions or decisions based on new data.

2) List the types of machine learning.

- **Supervised Learning**
- **Unsupervised Learning**
- **Semi-Supervised Learning**
- **Reinforcement Learning**
- **Self-Supervised Learning**
- **Transfer Learning**
- **Multi-Task Learning**
- **Active Learning**

3) State the goals of PAC learning?

The goals of **Probably Approximately Correct (PAC) learning** are:

- **Quantify the concept of learnability** — to formally define when and how a concept can be learned by an algorithm.
- **Provide performance guarantees** — to ensure that the learning algorithm, with high probability (confidence 1−δ1 - \delta1−δ), outputs a hypothesis whose error is less than or equal to a small value ϵ\epsilonϵ.
- **Determine sample complexity** — to find out how many training examples are needed for a learning algorithm to achieve the desired error and confidence levels.

4) Give an example of a standard learning task.

Classification tasks involve predicting a categorical label for a given input. For example, **spam detection** classifies emails as *spam* or *not spam*. This task uses algorithms such as Logistic Regression, Decision Trees, Support Vector Machines (SVM), k-Nearest Neighbors (k-NN), or Neural Networks (e.g., CNNs for image classification).

5) Compare supervised learning with unsupervised Learning.

Aspect	Supervised Learning	Unsupervised Learning
Data	Uses labeled data.	Uses unlabeled data.
Goal	Predict outputs from inputs.	Find hidden patterns or groupings.
Types	Classification, Regression.	Clustering, Dimensionality Reduction.
Examples	Spam detection, house price prediction.	Customer segmentation, market basket analysis.

6) Define a learning problem in the context of ML.

In the context of machine learning, a **learning problem** involves designing a model that can learn from data to perform a specific task, such as classification, regression, clustering, or prediction.

It is defined by:

- **The type of data** available (labeled or unlabeled).
- **The goal** of learning (predict outputs, find patterns, or make decisions).
- **The scenario** in which the model will be applied (e.g., spam detection, customer segmentation)

7) State the goal of a machine learning algorithm.

The goal of a machine learning algorithm is to learn patterns from data and produce a model that can make accurate predictions or decisions on new, unseen data, ensuring good generalization rather than just memorizing the training data.

8) Name any one application of machine learning.

Application – Fraud Detection

In the finance sector, machine learning is used to analyze transaction data in real time to identify unusual patterns or anomalies that may indicate fraudulent activity. Algorithms can learn from historical fraud cases to detect and prevent future fraudulent transactions.

9) Explain the key idea behind statistical learning.

The key idea of **statistical learning** is to build models that capture the relationship between inputs and outputs using statistical methods, optimize them to minimize errors, and evaluate their performance for making accurate predictions.

10) Differentiate between classification and regression in one sentence.

Classification is a supervised learning task where the goal is to predict a discrete label or category for an input (e.g., classifying emails as spam or not spam), whereas **Regression** is a supervised learning task aimed at predicting a continuous numerical value based on input features (e.g., estimating house prices from size and location).

Unit 2

1) State the key difference between linear and non-linear classification.

Aspect	Linear Classification	Non-Linear Classification
Decision Boundary	Straight line or hyperplane.	Curved or complex boundary.
Data Separation	Works when data is linearly separable.	Works for data that is not linearly separable.
Example Algorithms	Logistic Regression, Linear SVM.	Kernel SVM, Decision Trees, k-Nearest Neighbors, Neural Networks.

2) Define multi-label classification.

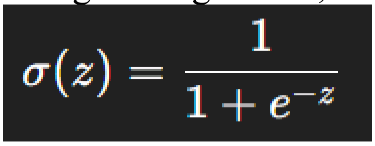
Multi-label classification is a type of classification problem where each instance can be assigned **multiple labels simultaneously** instead of just one.
For example, in image tagging, a single image might be labeled as *"beach"*, *"sunset"*, and *"vacation"* at the same time.

3) Expand the acronym ID3 as used in decision tree algorithms.

In decision tree algorithms, **ID3** stands for **Iterative Dichotomiser 3**.

4) Identify the function used in logistic regression to model probability.

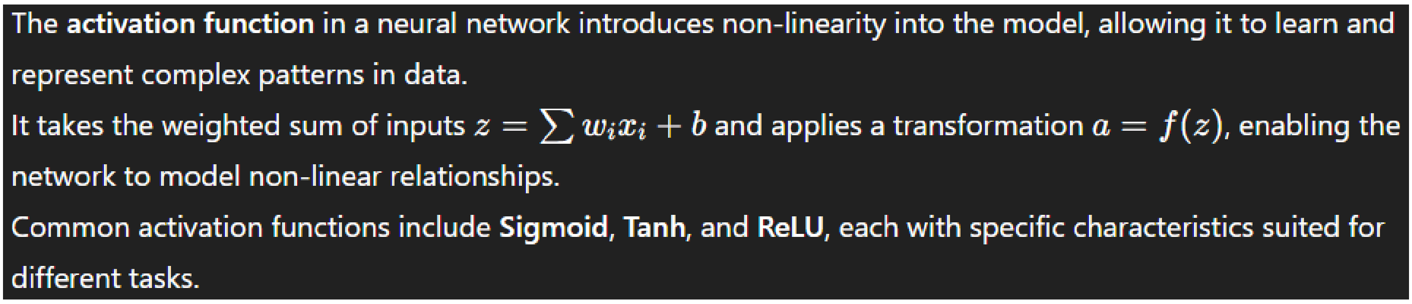
In logistic regression, the **sigmoid function**



$$\sigma(z) = \frac{1}{1 + e^{-z}}$$

is used to model probability, mapping any real value zzz to a range between 0 and 1, representing the likelihood of class membership.

5) Explain the role of an activation function in neural networks.



The **activation function** in a neural network introduces non-linearity into the model, allowing it to learn and represent complex patterns in data.

It takes the weighted sum of inputs $z = \sum w_i x_i + b$ and applies a transformation $a = f(z)$, enabling the network to model non-linear relationships.

Common activation functions include **Sigmoid**, **Tanh**, and **ReLU**, each with specific characteristics suited for different tasks.

6) How many layers does a perceptron consist of?

A **perceptron** consists of **two layers**:

1. **Input Layer** – Receives the feature values from the dataset.
2. **Output Layer** – Produces the final prediction after applying weights, bias, and an activation function.

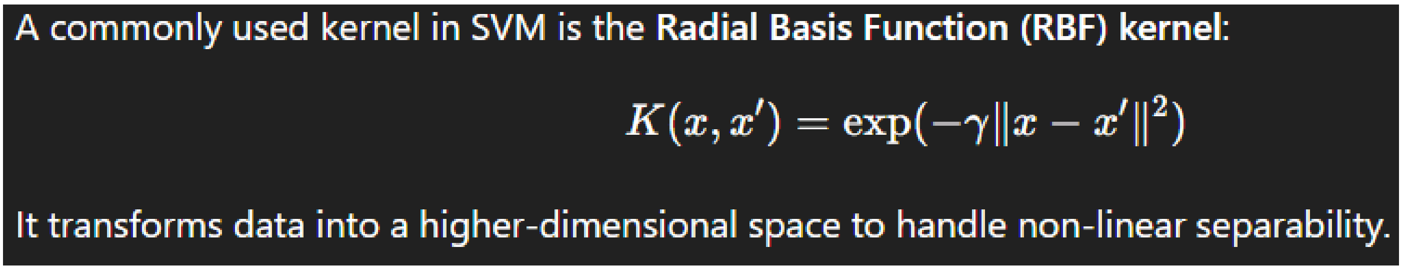
It does not have any hidden layers, which is why it is considered a **single-layer neural network**.

7) Name the algorithm that uses information gain for splitting data.

ID3 (Iterative Dichotomiser 3) algorithm uses **information gain** as a criterion to split data at each node in a decision tree.

It selects the attribute that provides the highest information gain, meaning it best reduces uncertainty or entropy in the dataset, resulting in more homogeneous child nodes and improved classification accuracy.

8) Name a kernel function used in SVM.



A commonly used kernel in SVM is the **Radial Basis Function (RBF) kernel**:

$$K(x, x') = \exp(-\gamma \|x - x'\|^2)$$

It transforms data into a higher-dimensional space to handle non-linear separability.

9) Define the meaning of 'K' in K-Nearest Neighbors.

In **K-Nearest Neighbors (KNN)**, **'K'** represents the number of nearest data points (neighbors) considered when determining the class or value of a new data point.
The algorithm assigns the label or value based on the majority vote (classification) or average (regression) of these **K** neighbors.

10) Identify whether linear regression is suitable for classification problems?

No, **linear regression** is not suitable for classification problems because it predicts continuous values, whereas classification requires discrete class labels.
For classification tasks, algorithms like **logistic regression**, **decision trees**, or **SVM** are more appropriate.

Unit 3

1) Define clustering in machine learning?

   In machine learning, **clustering** is an **unsupervised learning technique** used to group a set of data points into clusters, where points in the same cluster are more similar to each other than to those in other clusters.

   It helps in discovering hidden patterns or structures in data without using predefined labels.

2) Identify an algorithm that follows the partitional clustering approach.