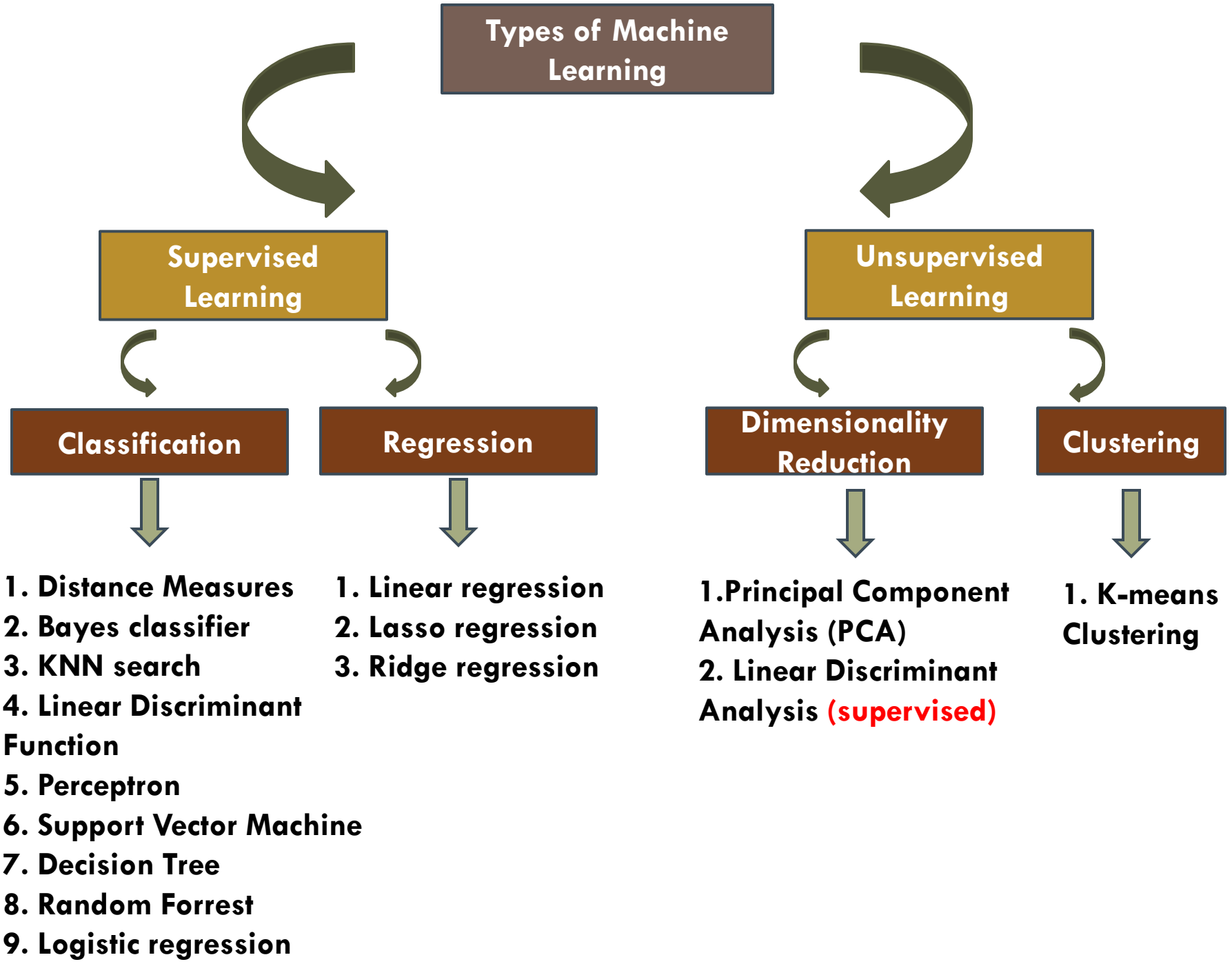
K-NEAREST NEIGHBOUR SEARCH ALGORITHM

Dr. Umarani Jayaraman
Assistant Professor

INDIAN INSTITUTE OF INFORMATION TECHNOLOGY,
DESIGN AND MANUFACTURING,
KANCHEEPURAM


Types of Machine Learning
Supervised Learning
Unsupervised Learning
Classification
Regression
Dimensionality Reduction
Clustering
1. Distance Measures
2. Bayes classifier
3. KNN search
4. Linear Discriminant Function
5. Perceptron
6. Support Vector Machine
7. Decision Tree
8. Random Forrest
9. Logistic regression
1. Linear regression
2. Lasso regression
3. Ridge regression
1.Principal Component Analysis (PCA)
2. Linear Discriminant Analysis (supervised)
1. K-means Clustering

K-NN Search

- The **K-Nearest Neighbors (KNN) algorithm** is a supervised machine learning method employed to tackle classification and regression problems.
- It is non-parametric, lazy learning method meaning it does not make any underlying assumptions about the distribution of data
- As opposed to other algorithms such as GMM, which assume a Gaussian distribution of the given data).

Key Steps of k-NN Search Algorithm

- **1. Data Storage:**
  - k-NN stores the entire dataset, which will be used for prediction.
- **2. Distance Calculation:**
  - When given a new input (query point), k-NN calculates the distance between this input and all the points in the dataset.
  - The most commonly used distance metrics are:
    - **Euclidean distance** (for continuous data)
    - **Manhattan distance** (for grid-like data)
    - **Hamming distance** (for categorical data)

Key Steps of k-NN Search Algorithm

- **3. Choosing Neighbors:**
  - After calculating the distances, the algorithm selects the **k** points from the dataset that are closest to the query point. These are the "k-nearest neighbors."
- **4. Majority Voting (for classification):**
  - The class labels of the k-nearest neighbors are examined.
  - The most common label (i.e., the mode) among them is assigned to the query point.

Key Steps of k-NN Search Algorithm

- **5. Averaging (for regression):**
  - In regression tasks, instead of voting, the algorithm computes the average of the output values of the k-nearest neighbors to make a prediction.

K-NN Search

- **Computational Complexity:**
  - One of the main disadvantages of k-NN is that it is computationally expensive, especially with large datasets, since it requires calculating the distance to every point in the dataset for every query.
  - The time complexity for a single query is $O(n \cdot d)$, where ‘n’ is the number of data points and ‘d’ is the number of dimensions.

K-NN Search

- **Applications:**
  - **Classification**: Handwritten digit recognition, image classification, and document categorization.
  - **Regression**: Predicting housing prices or other continuous variables.
  - **Recommendation Systems**: Finding items that are similar to user preferences based on their k-nearest neighbors.

THANK YOU