Lecture 04

Clustering

A clustering is a grouping / assignment of objects (data points) such that objects in the same group / cluster are:

- similar to one another
- dissimilar to objects in other groups

Applications:

- Outlier detection / anomaly detection
    - Data Cleaning / Processing
    - Credit card fraud, spam filter etc.
- Feature Extraction
- Filling Gaps in your data
    - Using the same marketing strategy for similar people
    - Infer probable values for gaps in the data (similar users could have similar hobbies, likes / dislikes etc.)

K-means Clustering

- **Type**: Unsupervised machine learning (no labels).
- **Goal**: Partition data into **K groups (clusters)** such that:
    - Points in the same cluster are similar (low intra-cluster distance).
    - Points in different clusters are dissimilar (high inter-cluster distance).
- **"Means"** comes from the algorithm using the **mean (centroid)** of the points in a cluster as its representative.

Steps in the K-Means Algorithm

1. **Choose K** (the number of clusters).
2. **Initialize Centroids**: Randomly pick K points from the dataset as starting centroids.
3. **Assign Points**: Each data point is assigned to the nearest centroid (usually using Euclidean distance).
4. **Update Centroids**: Recalculate the centroid of each cluster as the mean of all points assigned to it.
5. **Repeat**: Steps 3 and 4 until centroids stop moving significantly or max iterations are reached.

K – means – Lloyd's Algorithm

1. Randomly pick k centers $\{\mu 1, \dots, \mu k\}$
2. Assign each point in the dataset to its closest center
3. Compute the new centers as the means of each cluster
4. Repeat 2 & 3 until convergence

Properties:

- **Convergence:** Algorithm always converges in a finite number of steps (since cost function decreases each iteration).
- **Local minimum**: Not guaranteed to find global optimum (sensitive to initialization).
- **Time complexity**: $O(n \cdot k \cdot d \cdot I)$
    - n = number of points
    - k = number of clusters
    - d = number of dimensions
    - I = number of iterations