Clustering

What Is Clustering?

- An **unsupervised learning** method
- No class labels provided
- Goal: group instances into **natural clusters** based on similarity
- Assumes underlying structure exists in the data
- Cluster types:
  - Exclusive (hard)
  - Overlapping (soft)
  - Probabilistic
  - Hierarchical

Why Clustering?

- Reveals hidden data structure
- Useful for:
  - Customer segmentation
  - Outlier detection
  - Data compression
  - Exploratory analysis
  - Preprocessing for other ML tasks

Classic Algorithm: k-Means

- Steps:
    1. Choose **k** clusters
    2. Randomly initialize **k centroids**
    3. Assign points to nearest centroid
    4. Recompute centroid of each cluster
    5. Repeat until convergence
- Produces **hard** clusters.

Objective Function

- k-Means minimizes within-cluster sum of squares (WCSS):

$$\sum_{i=1}^{k} \sum_{x \in S_i} \|x - \mu_i\|^2$$

- μ_i = centroid of cluster i
- Finds **local**, not global, minimum

Sensitivity to Initialization

- k-Means can:
  - Converge to poor solutions
  - Produce different results depending on initial seeds
  - Fail on non-spherical or unequal-sized clusters
- Example: rectangle clusters where long-side grouping is incorrect.

k-Means++ Initialization

- Improved seeding technique:
  1. Pick first centroid randomly
  2. Pick others with probability proportional to distance^2 from existing centroid
- Produces better clusters and faster convergence.

Numeric Attributes Requirements

- k-Means assumes Euclidean space
- Works best with numeric attributes
- Preprocessing:
  - Normalize data
  - Remove outliers
  - Optional: dimensionality reduction

Stopping Criteria

- Stop when:
    - Assignments no longer change
    - Centroid shift $<$ tolerance
    - Max iterations reached
- Most datasets converge rapidly.

Issues with k-Means

- Sensitive to initialization
- Assumes spherical clusters
- Poor for:
  - Different-sized clusters
  - Varying densities
  - Non-convex shapes

Measuring Cluster Quality

- Metrics:
    - WCSS (lower is better)
    - Silhouette score
    - Inter-cluster vs. intra-cluster distance
- Clustering lacks ground truth—quality is often subjective.

Choosing k

- Methods:
  - Elbow method
  - Silhouette coefficient
  - Gap statistic
  - Domain knowledge

Speeding Up k-Means

- Distance computations dominate cost.
- Optimization:
  - Use kD-trees
  - Use ball trees
- Entire nodes can sometimes be assigned at once.

Hierarchical Clustering (Overview)

- Two types:
  - Agglomerative: bottom-up merging
  - Divisive: top-down splitting
- Produces a **dendrogram** (tree of clusters).

Density-Based Clustering

- Finds arbitrarily-shaped clusters
- Identifies noise/outliers
- Does not need k
- Mentioned for context beyond k-Means.

Clustering Applications

- Customer segmentation
- Image compression
- Document clustering
- Bioinformatics
- Anomaly detection

Practical Tips

- Always normalize data
- Try multiple seeds
- Use k-Means++
- Examine multiple values of k
- Compare with non-centroid methods

Summary

- Clustering:
  - Discovers patterns without labels
  - k-Means = most widely used method
  - Sensitive to initialization -> k-Means++ helps
  - Tree structures accelerate distance calculations
  - Choosing k requires experimentation