K-Means Clustering

1. Model Overview

K-Means Clustering is an **unsupervised learning** algorithm used to group data points into **K distinct, non-overlapping clusters** based on feature similarity. It minimizes intra-cluster distance and maximizes inter-cluster distance using an iterative optimization process.

---

2. Key Aspects to Analyze for K-Means

A. Assumptions

1. **Clusters are spherical** (points in a cluster are closer to the centroid than to other clusters).
2. **Clusters are of similar size and density**.
3. **Features contribute equally —** hence, scaling is required.
4. Each data point **belongs to exactly one cluster** (hard clustering).
5. The number of clusters **K is known or can be estimated** beforehand.
6. Data has **low noise and few outliers**, as they can distort centroids.

B. Limitations

1. Requires pre-specifying the number of clusters **(K)**.
2. **Sensitive to outliers** — they can pull cluster centroids away from the true center.
3. **Assumes convex and isotropic clusters**, struggles with irregular shapes.
4. **Initialization impacts results —** poor initial centroids can cause suboptimal clustering.
5. **Scales poorly** for high-dimensional or very large datasets.

6. Only supports **numeric (continuous)** data.
7. **Does not guarantee a global optimum**, only a local one due to random initialization.

C. Attributes / Input Features

1. Works best with **continuous, numerical** variables.
2. Categorical data requires **encoding (e.g., One-Hot Encoding)** or use of **K-Modes/K-Prototypes**.
3. **Feature scaling** (standardization or normalization) is mandatory.
4. Outliers should be removed or treated before clustering.
5. Features should have **comparable influence** to avoid bias in distance computation.

D. Internal Model Variations / Subtypes

1. **K-Means++** — improved centroid initialization to enhance convergence and accuracy.
2. **Mini-Batch K-Means** — uses small random batches for faster clustering on large datasets.
3. **Bisecting K-Means** — hierarchical approach that splits clusters recursively.
4. **Fuzzy K-Means (Soft Clustering)** — allows data points to belong to multiple clusters with probabilities.
5. **K-Prototypes** — hybrid algorithm for mixed numerical and categorical data.

E. Hyperparameters

1. **K (Number of Clusters):**
    - Chosen using the **Elbow Method**, **Silhouette Score**, or **Gap Statistic**.

2. **Initialization Method:**
    - Random or **K-Means++** (recommended).
3. **Number of Initializations (n_init):**
    - Number of times the algorithm runs with different centroid seeds.
4. **Max Iterations (max_iter):**
    - Upper limit on the number of optimization steps.
5. **Distance Metric:**
    - Default: **Euclidean distance** (can be customized in extensions).

F. Performance Evaluation Metrics

Since K-Means is **unsupervised**, evaluation is often internal:

- **Inertia (Within-Cluster Sum of Squares)**
- **Silhouette Score**
- **Davies–Bouldin Index**
- **Calinski–Harabasz Index**

If labels are available (for validation):

- **Adjusted Rand Index (ARI)**
- **Normalized Mutual Information (NMI)**

G. Use Cases

- Customer segmentation
- Image compression and color quantization
- Market basket analysis

- Document or text clustering
- Anomaly detection (via distance from centroids)

H. Optimization Tips

1. Use **K-Means++ initialization** to reduce poor clustering outcomes.
2. **Standardize features** to equalize influence.
3. Use **PCA** for dimensionality reduction before clustering high-dimensional data.
4. Evaluate multiple K values to find optimal cluster count.
5. Remove **outliers** before fitting the model.
6. Use **Mini-Batch K-Means** for large datasets.

I. Model Interpretation

- Each cluster is represented by its **centroid** (mean of all points in the cluster).
- Distances from centroids indicate **similarity or dissimilarity**.
- Visualize clusters using **2D/3D plots** or **PCA-reduced features**.
- The algorithm does not provide explainability like feature importance — interpretation is geometric.

J. Summary Table

Aspect	Description
Type	Unsupervised Clustering
Objective	Minimize intra-cluster variance
Key Parameter	K (number of clusters)
Data Type	Numerical (continuous)

Distance Metric	Euclidean (default)
Cluster Type	Hard, spherical
Requires Scaling	Yes
Handles Outliers	Poorly
Handles Categorical Data	No (use K-Modes/K-Prototypes)
Ideal Dataset Size	Small to medium
Initialization Sensitivity	High