DATE: / /
PAGE:

K-Means Clustering

- **What is K-means**
K-means is an unsupervised clustering algorithm that partitions data into K clusters, where each data point belongs to the cluster with the nearest centroid.

- Purpose (When)
To group similar data points together based on distance, minimizing intra-cluster variance.

- **Why it is used**
  - Simple and fast
  - Scales well to large datasets
  - Easy to implement and interpret
  - Effective when clusters are spherical

- **How it works (Algo)**
  - Choose number of clusters K
  - Randomly initialize K centroids.
  - Assign each data point to nearest centroid
  - Recompute centroids as cluster mean
  - Repeat until convergence.

- Formula
  - Within cluster sum of squares: $J = \sum_{i=1}^{K} \sum_{x \in C_i} \|x - \mu_i\|^2$

C_i = cluster
μ_i = centroid

DATE: \ \
PAGE:

- Technical Details
  - Distance metric : Euclidean
  - Convergence when centroids stop moving
  - Sensitive to initialization

- Parameters
  - n_clusters (K) : no. of clusters
  - init : centroid initialization
  - max_iter: maximum iterations
  - tol: convergence threshold

- Pros
  - Fast and scalable
  - Easy to understand
  - Efficient for large datasets

- Cons
  - Must choose K beforehand
  - Sensitive to outliers
  - Fails on non-spherical clusters
  - Sensitive to initialization

- Real-world Applications
  - Customer segmentation
  - Image compression
  - Market basket analysis
  - Document clustering

→ **Hierarchical clustering**

- What is hierarchical clustering
  Hierarchical clustering builds a tree-like structure (dendogram) that represents nested clusters without requiring K beforehand.

- Purpose (When)
  To discover natural grouping structures and cluster hierarchy in data.

- Why it is used
  - No need to predefine number of clusters.
  - Produces interpretable dendogram
  - Works well for small datasets

- How it works
  - Agglomerative (Bottom up)
    - Start with each point as its own cluster
    - Merge closest clusters
    - Repeat until one cluster remains
    - Cut dendogram at desired level
  - Divise (Top-Down)
    - Start with all points as one cluster
    - Divide most apart clusters

- Distance Linkage methods
  - Single : Min distance
  - Complete : Max distance
  - Average : Avg distance
  - Ward : Variance minimization.

- Formula (Ward's method)

$$\Delta ESS = \frac{n_1 n_2}{n_1 + n_2} \| \mu_1 - \mu_2 \|^2$$

- Technical Details
  - Time complexity : $O(n^3)$ (naive)
  - Memory intensive
  - Distance matrix required

- Parameters

  linkage : ~~mage~~ merge criterion

  metric : distance measure

  n_clusters : optional

- Pros
  - No K required initially
  - Interpretable hierarchy
  - Deterministic results

- Cons
  - Not scalable
  - Sensitive to noise
  - Cannot undo merges.

- Real World Applications
  - Biological Taxonomy
  - Document classification
  - Gene expression analysis.

DATE:
PAGE:

→ DBSCAN (Density-based Spatial Clustering)

- What is DBSCAN
DBSCAN is a density-based clustering algorithm that groups points in high-density regions and labels sparse points as outliers.

- Purpose (When)
To identify arbitrarily shaped clusters and detect noise automatically.

- Why it is used
  - No need to specify number of clusters
  - Handles noise naturally
  - Detects non-spherical clusters

- How it works
  - Select parameters ε (epsilon) and MinPts
  - Identify core points
  - Expand clusters from core points
  - Label non-reachable points as noise

- Definitions
  - Core Point: $\geq$ MinPts within ε radius
  - Border Point: reachable but not core
  - Noise Point: not reachable.

- Technical Details
  - Distance based density estimation
  - Sensitive to ε choice
  - Struggles with varying densities

- Parameters
eps (ε) : Neighborhood radius
min_samples : min points for core
metric : distance measure

- Pros
  - No K required
  - Detects noise / outliers
  - Finds complex shapes

- Cons
  - Poor with varying densities
  - Parameter tuning is hard
  - Struggles in high dimensions

- Real-world applications
  - Anomaly detection
  - GPS location clustering
  - Image segmentation
  - Fraud detection.