__/__/____

(38) BIRCH Algorithm (Balanced Iterative Reducing and Clustering using Hierarchies)

Motivation: K-Means and Hierarchical fail when:

- The dataset is too large (high memory/time cost)
- We can't store all pairwise distances.

BIRCH was designed for large datasets, clustering them incrementally and efficiently using a compact summary structure.

Instead of storing all datapoints, BIRCH summarizes them into smaller representations called CFs (Clustering Features) and organizes them in a CF Tree.
Then it perform Clustering on these summaries rather than the raw data.

→ Clustering Feature

Each subcluster is represented by a triple:

$CF = (N, LS, SS)$

- N: number of points
- LS: $\sum_{i=1}^{N} x_i$ = linear sum of points
- SS: $\sum_{i=1}^{N} x_i^2$ = squared sum of points

From these we can derive,

- Centeroid: $M = LS/N$
- Radius: $R = \sqrt{SS/N - M^2}$
- Diameter: $D = \sqrt{(2SS - 2LS^2/N) / N(N-1)}$

__/__/____

CF Tree structure

A height balanced tree where each node stores several CF entries.

Parameters:

Branching factor (B): max number of children per node

Threshold (T): max radius / diameter of subcluster in leaf no

- Non leaf nodes: stores CF summeries of child nodes
- Leaf nodes: store actual cluster summaries.

Algorithm:

1. Build CF Tree
   - Start with the first data point
   - Insert each point into the nearest leaf
   - If that leaf's radius exceeds threshold (T), split the leaf.
   - Continue until all data points are inserted
2. Condense CF Tree: Remove outliers / merge sparse clusters
3. Global Clustering:
   Apply a traditional clustering algorithm to the leaf entries (CFs) - much fewer than original points
4. Refinement: Reassign points to the nearest cluster centeroid for fine-tuning.

Advantages:

1. Scalability
2. Incremental, Fast
3. Handles Noise

Disadvantages:

1. Sensitive to data order
2. Not ideal for non-spherical clusters

* If you have 1M points →
  BIRCH compresses them into, 10K subclusters using CFs
  Then runs K-Means on just 10K → 100x faster, similar results.

__/__/____

(39) CURE Algorithm (Clustering Using REpresentatives)

K-Means & BIRCH assume spherical clusters and fail on arbitrary shapes or when clusters very in size and density.

CURE improves hierarchical clustering by using multiple representative points per cluster and shrinking them toward the mean to handle non-spherical shapes & outliers.

Each cluster is represented by c representative points, not just a centroid.

Steps:
1. Choose c well-scattered points from a cluster
2. Shrink them toward the cluster centeroid by a fraction α ($0 < \alpha < 1$).
3. Compute distances b/w clusters based on the closest pair of representative points.
4. Merge clusters iteratively (like hierarchical clustering)

c: no of representative points per cluster

α: shrinking factor (controls compactness)

Advantages: Handles non-spherical clusters
Robust to outliers
Better than K-Means & Hierarchical

Disadvantages: Computationally heavier than BIRCH
Requires careful choice of α and c