Hierarchical

Tuesday, April 30, 2019 2:04 PM

Grouping of objects/samples into sets.
Unsupervised: Everything is unlabeled.
No knowledge about the true number of clusters.

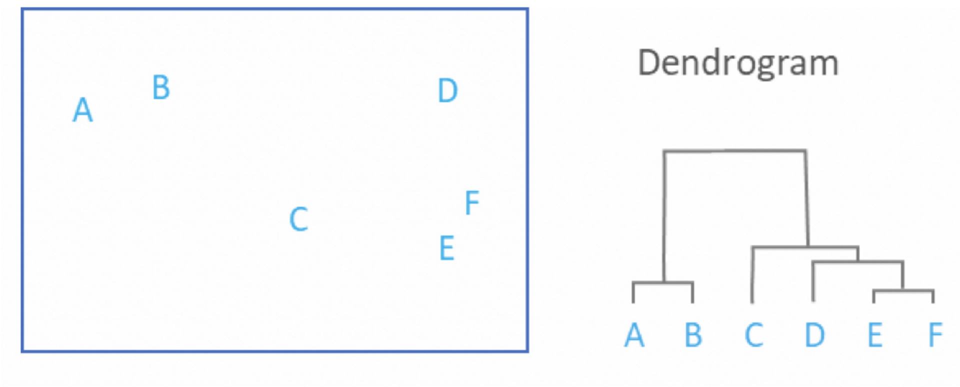
K-means: good for convex and isotropic clusters.
not great for irregular or elongated clusters.

Hierarchical: builds nested clusters through merging or splitting

Top Down: divisive (all objects start as one big cluster and you split it up until every point is its own cluster)

Bottom Up: agglomerative (ahg glohm er aye tive ?) is the opposite. You start with each point being its own cluster and group them into bigger clusters. (pebbles to boulders)

Steps



First step: most obvious choices. A and B are obviously close. E and F and super close.
Second step: It's more difficult to decide what comes next we use the "linkage criteria."

Linkage criteria :

Maximum - Max distance between elements of the cluster.
Example: If considering distance between c and the cluster AB you'd choose the distance between the farthest point, A

Average - Average distance between elements of the cluster.
Example: If considering distance between c and the cluster AB you'd choose the average point between point A and point B

Minimum - Minimum distance between elements of the cluster.
Example: If considering distance between c and the cluster AB you'd choose the distance between the closest point, B

Ward - Minimizes the inertia of the clusters.

Maximum: $d_{max}\ (C_k, C_{k'}) = {}_{x\epsilon\ c_k}max_{x'\ \epsilon\ c_k}\ ||x' - x||$
Minimum: $d_{min}\ (C_k, C_{k'}) = {}_{x\epsilon\ c_k}min_{x'\ \epsilon\ c_k}\ ||x' - x||$
Average: $d_{avg}(C_k, C_{k'}) = \frac{1}{N_k N_{k'}} \sum_{x\epsilon c_k} \sum_{x\epsilon c_{k'}} ||x - x'||$

Clustering_class = AgglomerativeClustering()

- N_clusters
- Linkage
  - Average
  - Complete (maximum)
  - Ward

Nothing went well with anisotropic data.
With unevenly sized blobs average does pretty well.

K_neighbors graph
Connectivity = kneighbora_graph(include_self = False) Does well for the moon data with average and complete
In AgglomerativeClustering you set connectivity=connectivity