Lecture 9/27/2021

Hierarchical

A set of nested clusters organized in a tree

Merge clusters in order to produce a dendrogram

Choose certain levels of the dendrogram in order to decide the clusters

Agglomerative:

1. Start with every point in its own cluster
2. At each step, merge the two closest clusters
3. Stop when every point is in the same cluster

Divisive

1. Start with all points in same cluster
2. Divide into two clusters until every point is in its own cluster

Single link distance:

Can handle clusters of different sizes

Sensitive to noise points

Tends to create elongated clusters

Complete link distance:

Maximum of all pairwise distances between points in two clusters

Less vulnerable to noise

More balanced clusters

Tends to split up large clusters

Average link distance:

Average of all link distances

Less vulnerable to noise and outliers

Tends to be biased toward globular clusters

Centroid distance:

Distance between cluster centroids

Ward's distance:

Difference between the variance of points in the merged cluster and unmerged clusters

Density-Based

Defined based on the local density of points

Define a radius **epsilon** around each point

Define a region as dense when a point has a minimum number of points around it

Epsilon neighborhood - points within the radius epsilon around a single point

Core point - center of a dense region

Noise point - Neither core nor border point

Border point - within an epsilon neighborhood but not a core point

Create clusters by connecting core points

Lecture 9/27/2021

DBScan Algorithm

Epsilon and **min_pts** defined/given:

1. Find epsilon neighborhood of each point
2. Label point as core if contains at least **min_pts**
3. Label points in neighborhood that are not core as border
4. Label as noise if neither core nor border
5. For each core point, assign to the same cluster all core points in its neighborhood
6. Assign border points to nearby clusters

Generating the clusters, by labeling core points and examining border points, can be implemented with Breadth-First Search (**BFS**) algorithm

Benefits:

Works with many shapes and sizes for clusters
Resistant to noise

Disadvantages:

Fixed density can drastically change the number of points considered as noise
Creates clusters of same density
Notion of density in problematic in high-dimensional spaces