CS472 HW 5
Clustering

Types of Clustering Covered:

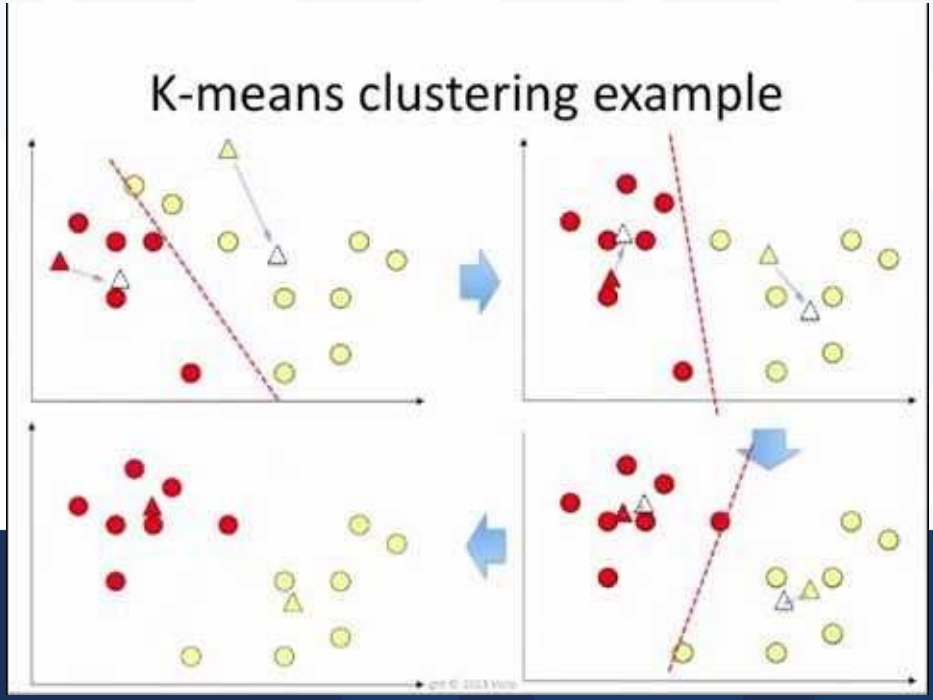
1. K-Means
2. Hierarchical
3. DBSCAN

K-Means

K-Means Clustering is a method of partitioning in which the data is divided into what are called “K clusters”. These clusters are organized based on the average/mean/’centroid’ of the data points within each cluster.

K-Means Example

An example of a good use of K-Means clustering would be using it on a store's commerce data in order to create sets of similar customers which you can use for specific targeting of ads, promotions, etc.



Hierarchical Clustering

Hierarchical Clustering works by building a tree-like structure of clusters nested within each other by successively merging or splitting data points based on their perceived similarities. It essentially iterates through the data determining if each point is alike enough to group together or split apart. Hierarchical clustering works especially well with smaller data sets.

Hierarchical Clustering

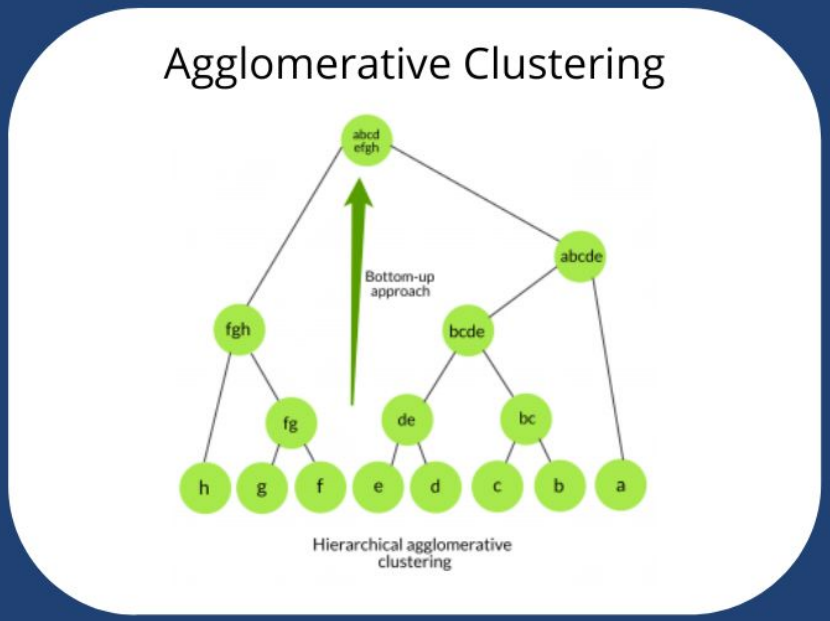
Hierarchical Clustering works by building a tree-like structure of clusters nested within each other by successively merging or splitting data points based on their perceived similarities. It essentially iterates through the data determining if each point is alike enough to group together or split apart. Hierarchical clustering works especially well with smaller data sets.

Hierarchical Clustering

Hierarchical Clustering works by building a tree-like structure of clusters nested within each other by successively merging or splitting data points based on their perceived similarities. It essentially iterates through the data determining if each point is alike enough to group together or split apart. Hierarchical clustering works especially well with smaller data sets.

Hierarchical Example

A couple examples of a good use of hierarchical clustering would be grouping animals according to their biological features to reconstruct phylogeny trees, or forming a hierarchy of employees based on salary.

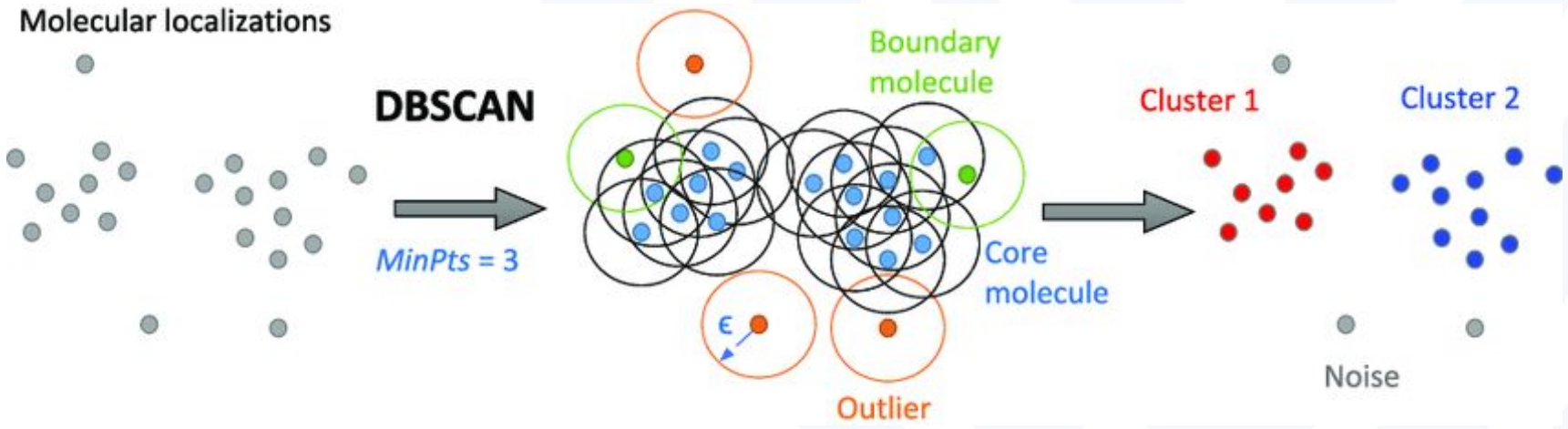


DBSCAN Clustering

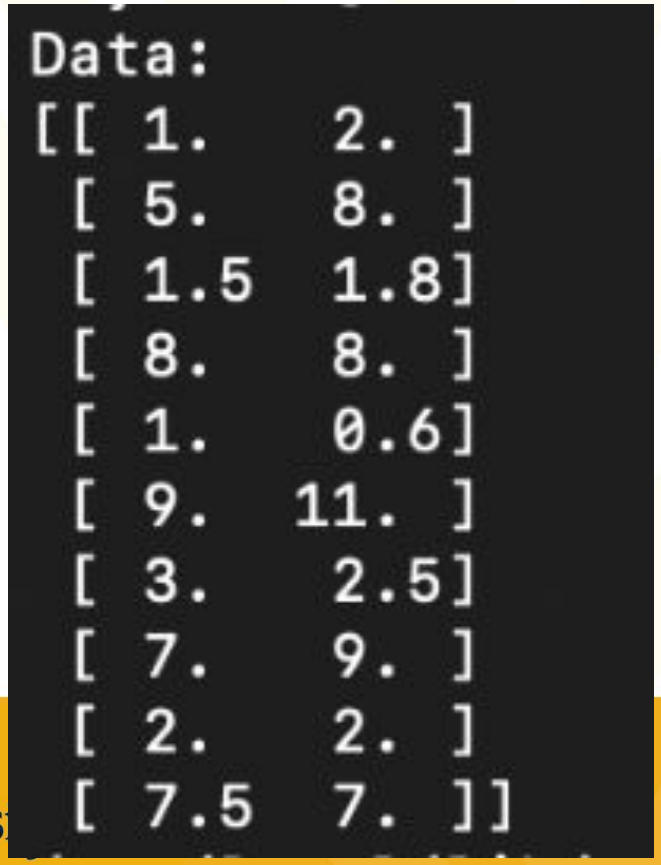
DBSCAN clustering, which stands for ‘Density-Based Spatial Clustering of Applications with Noise’, is a method of clustering which works by judging whether or not data points should be clustered together by their density. It essentially groups together data points in areas with a high density of data, and separates the data points in areas of lower density of data. The measure of high/low density is of course relative to the current dataset. Of all the common clustering algorithms out there, DBSCAN is one of the algorithms that makes the fewest assumptions about the shape of your clusters. That means that DBSCAN can be used to detect clusters that are oddly or irregularly shaped.

DBSCAN Example

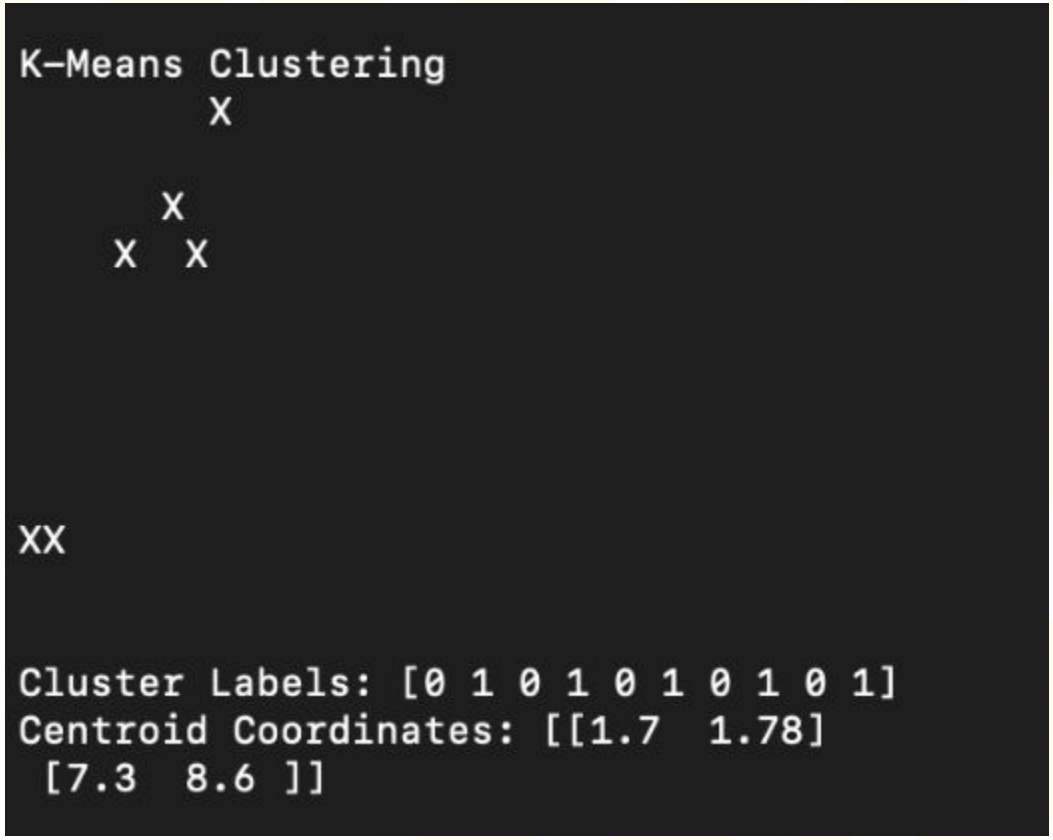
An example of a good use for DBSCAN clustering would be identifying ‘hotspots’ within datasets.



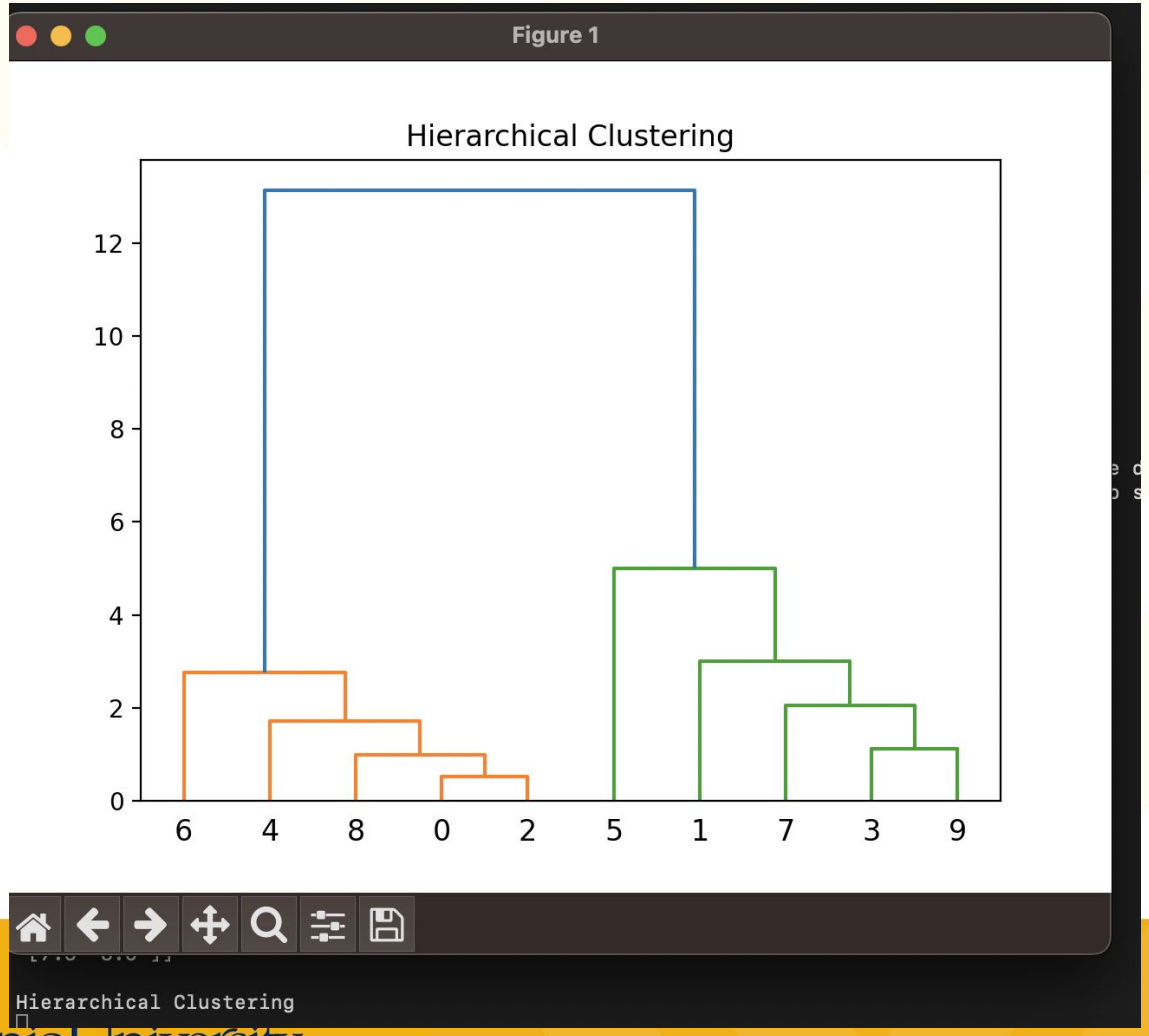
Data My Code Performed Clustering On



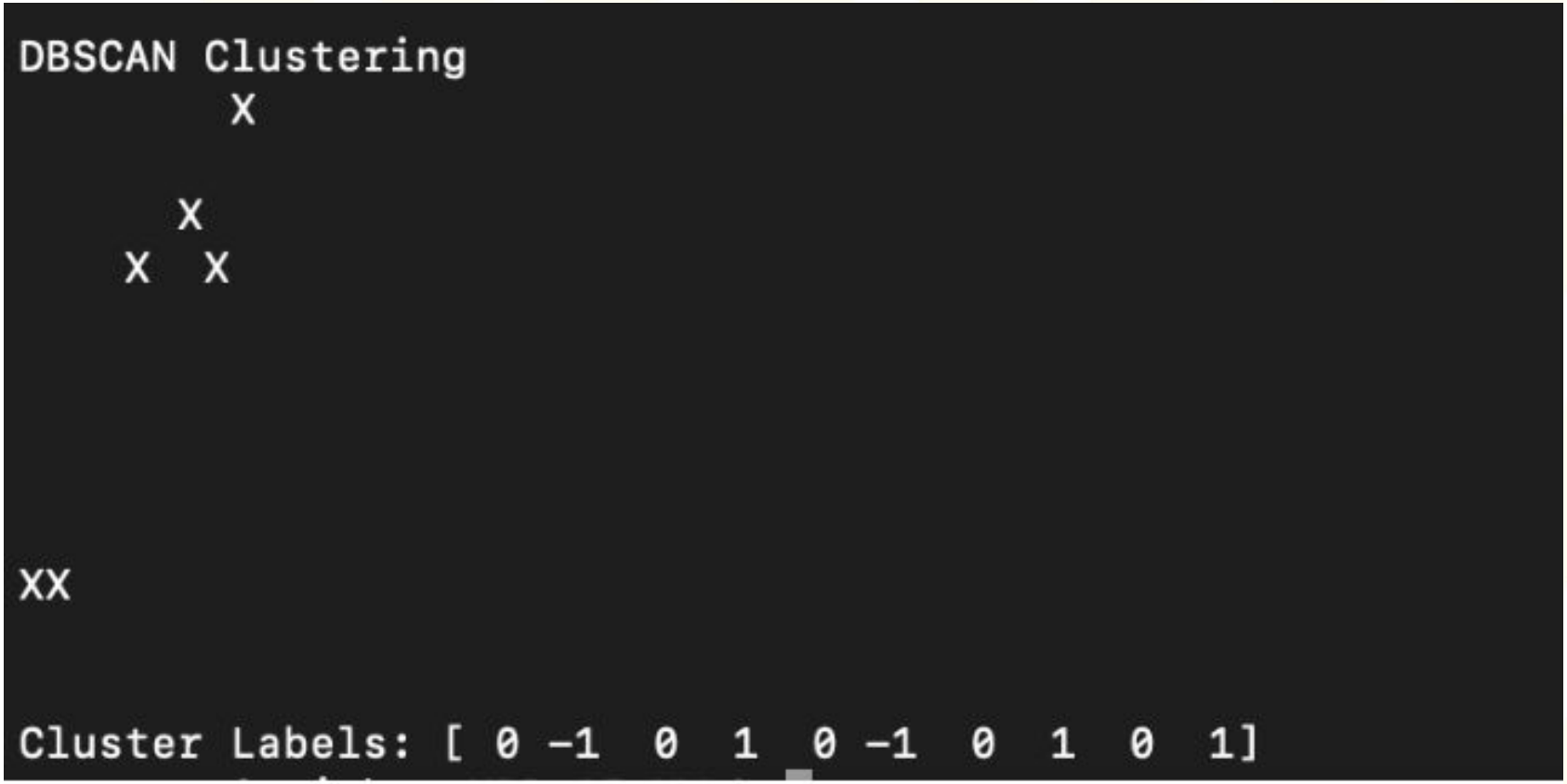
K-Means Output



Hierarchical Output



DBSCAN Output



Additional Forms of Clustering Included in Lesser Detail:

1. Agglomerative Clustering
2. Mean Shift Clustering
3. Spectral Clustering

Agglomerative

Agglomerative clustering is a type of hierarchical clustering that starts with individual data points as separate clusters and then merges the closest clusters iteratively.

Mean Shift

Mean Shift clustering is a non-parametric clustering method that shifts data points towards the mode (peak) of the density function.

Spectral

Spectral clustering works by transforming data into a lower-dimensional space and performs clustering in that space. It is effective for complex data structures.