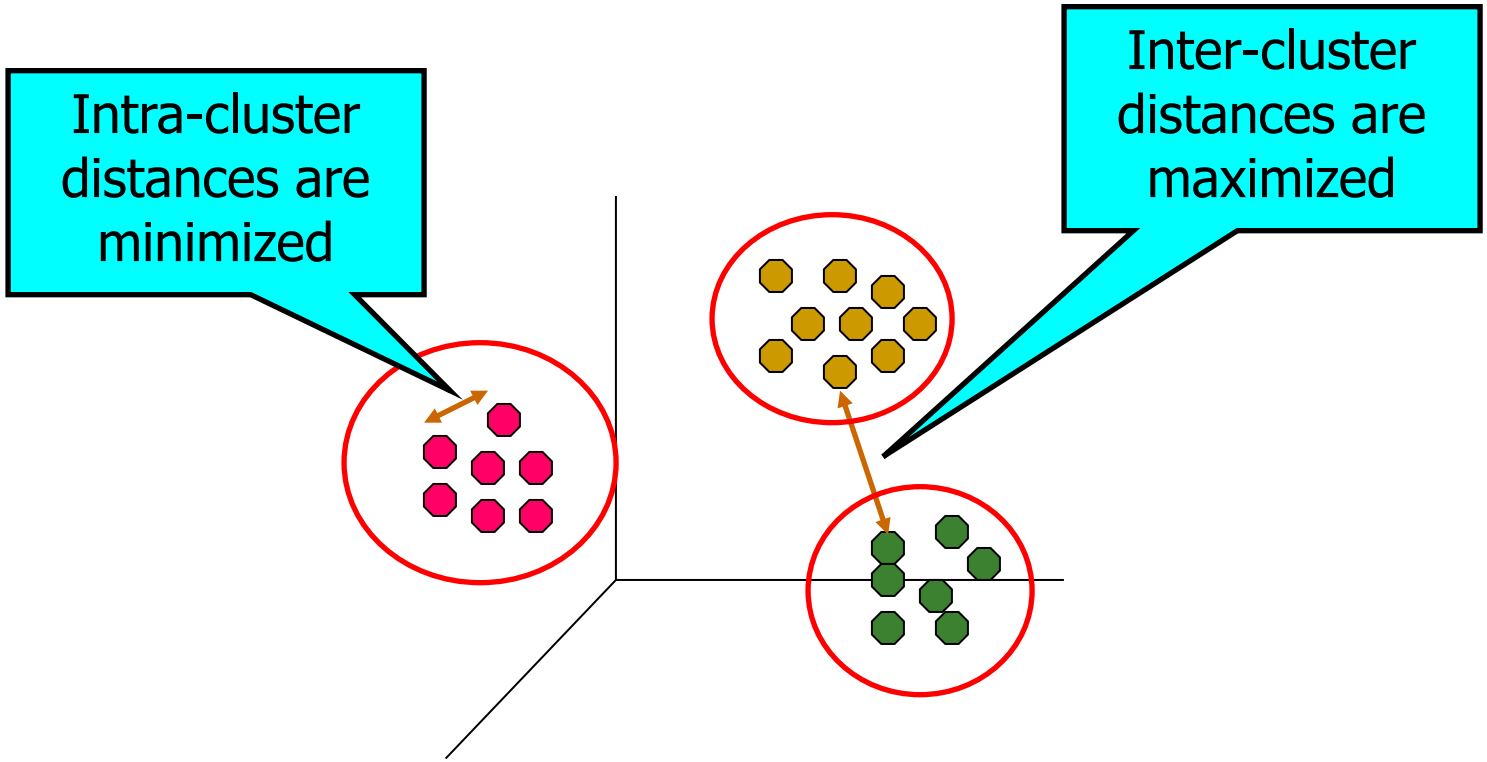
Lecture 4
Machine Learning: Unsupervised Learning

Phayung Meesad, Ph.D.
King Mongkut's University of Technology
North Bangkok (KMUTNB)
Bangkok Thailand

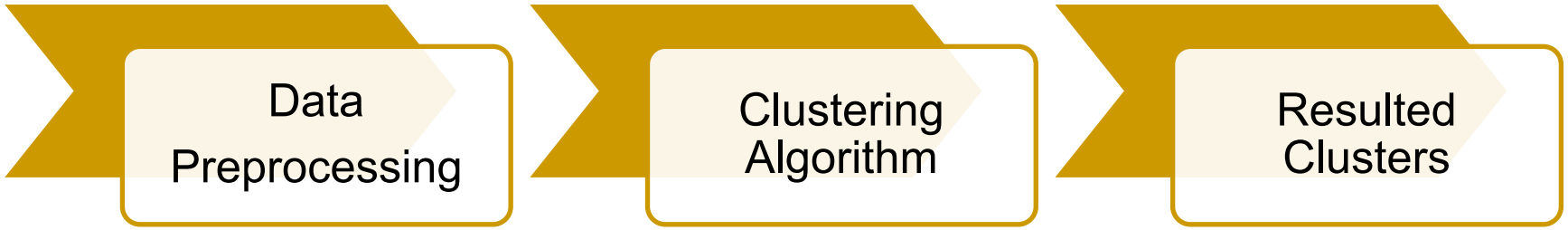
Unsupervised Learning

- Unsupervised Learning is a machine learning algorithm that learns data without targets.
- Objects in the same cluster are more similar to each other than to those in other clusters.

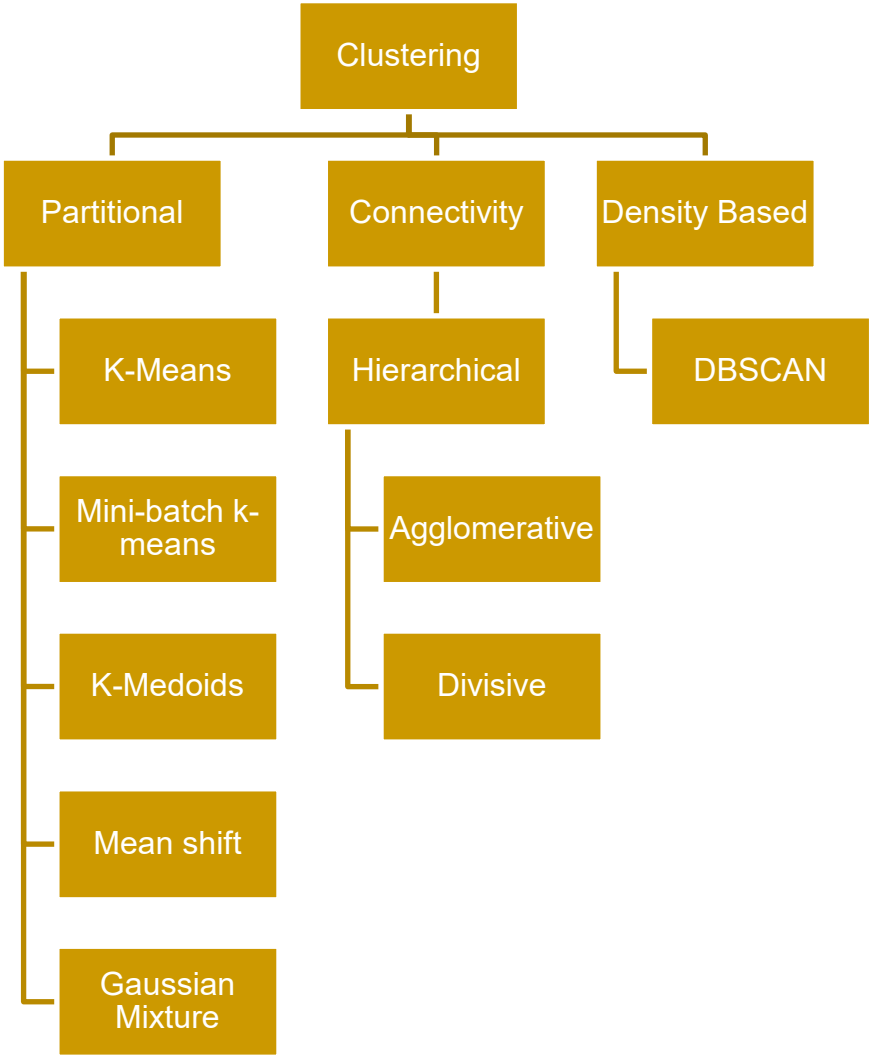
Intra-Cluster & Inter-Cluster Distance



Clustering Process

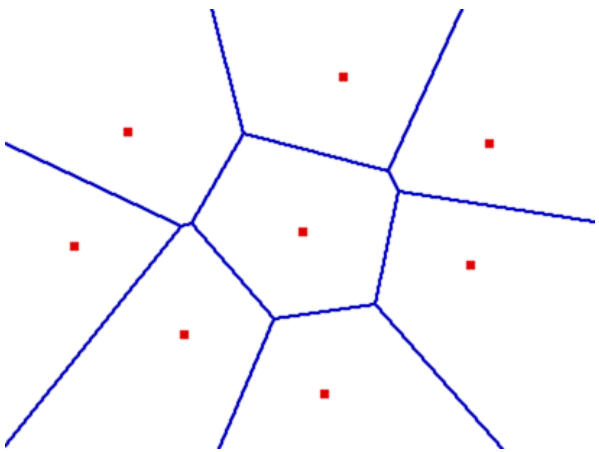


Clustering Methods



Centroid models

- Centroids are centers or means of the clusters. Each data point is assigned to the closest centroids.

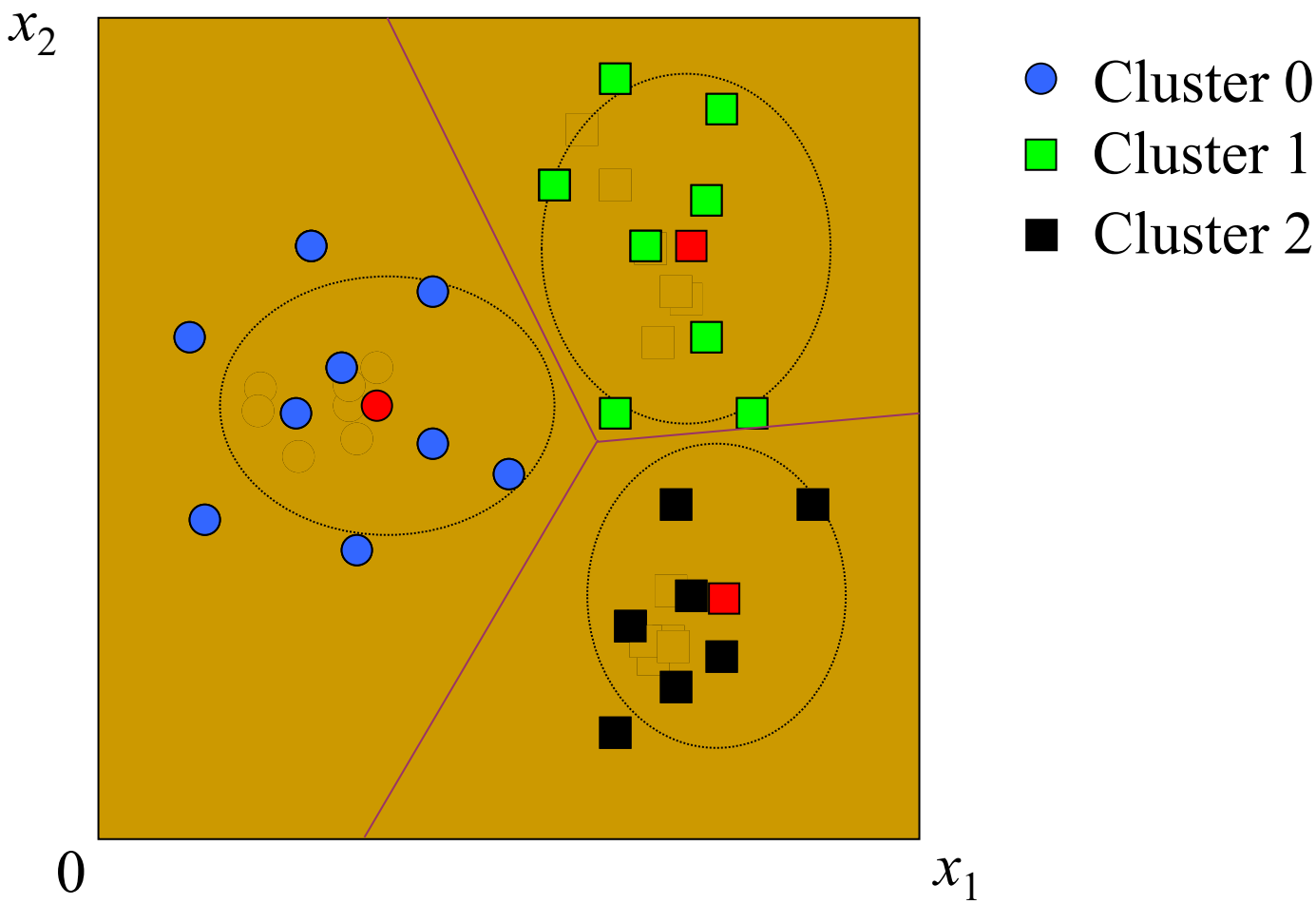


Partitional method

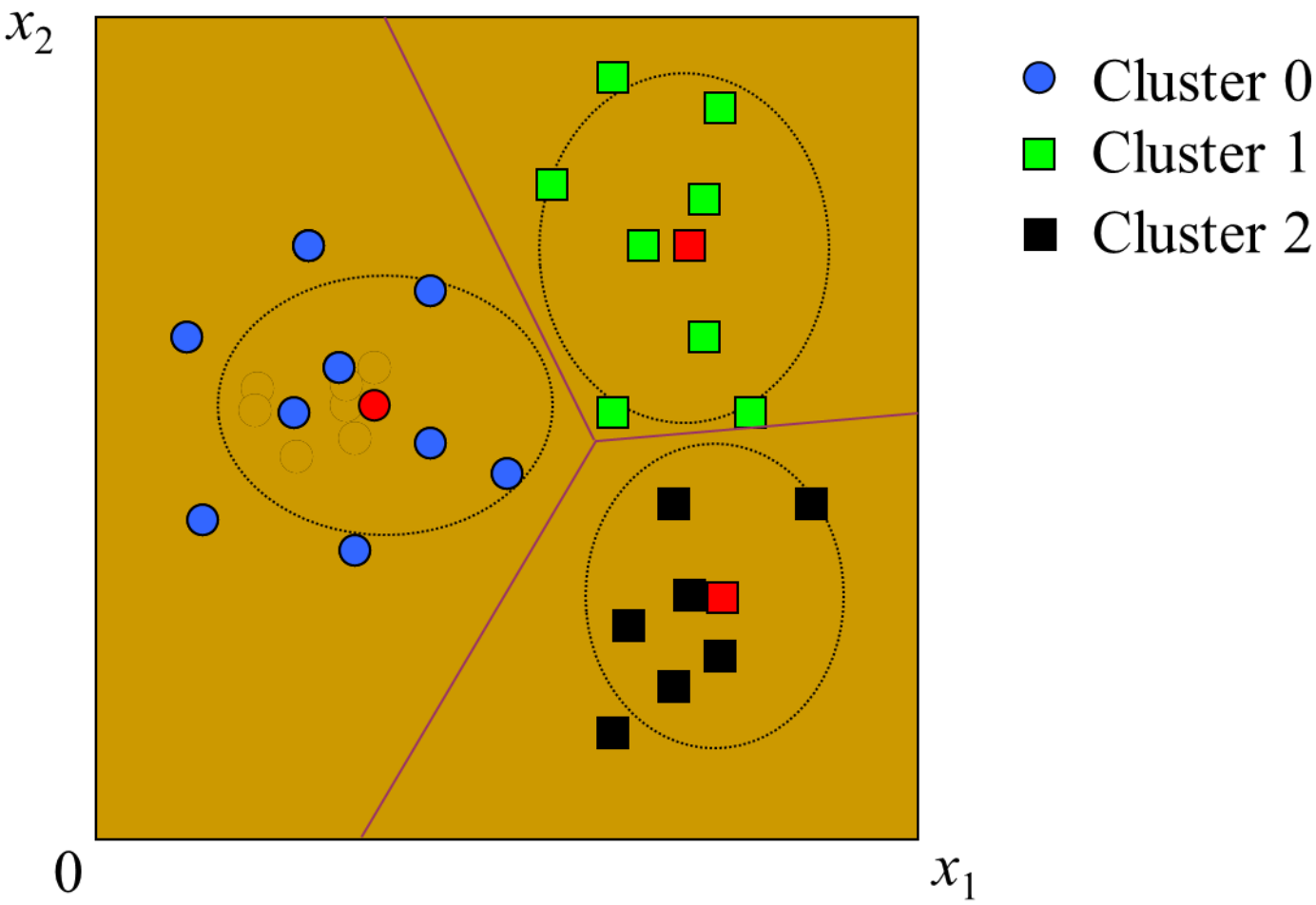
$$similarity = \cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}}$$

$$\mathrm{d}(\mathbf{p},\mathbf{q}) = \mathrm{d}(\mathbf{q},\mathbf{p}) = \sqrt{(q_1-p_1)^2+(q_2-p_2)^2+\cdots+(q_n-p_n)^2} = \sqrt{\sum_{i=1}^{n}(q_i-p_i)^2}.$$

Incremental Update

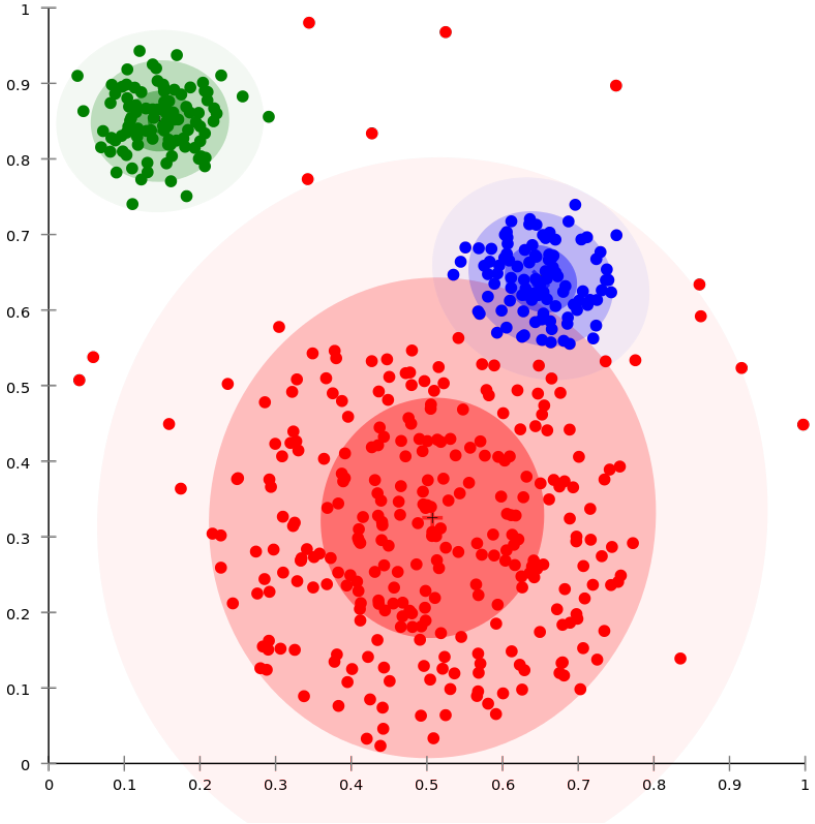


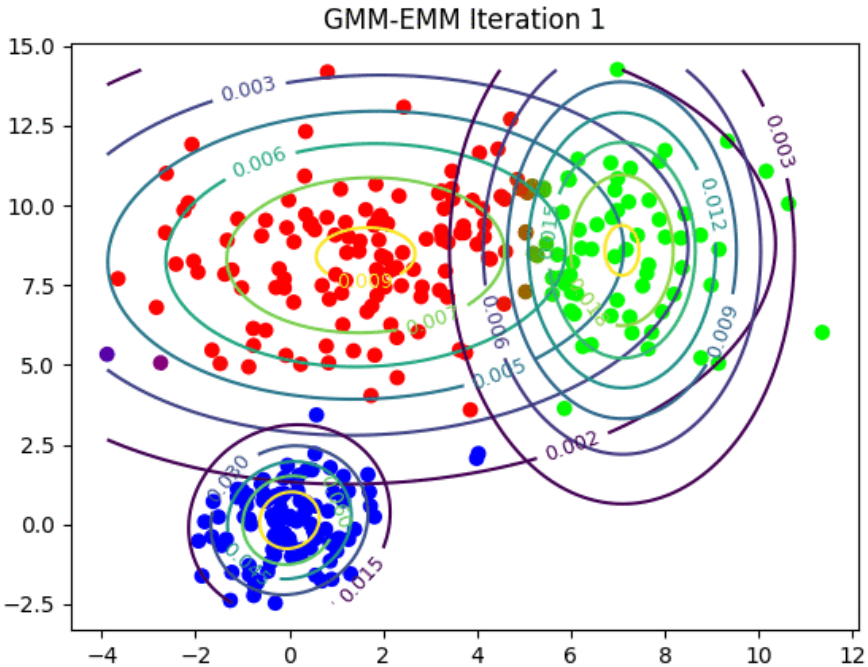
Batch Update



Distribution models

- Gaussian Mixture Models
- Expectation Maximization Algorithm

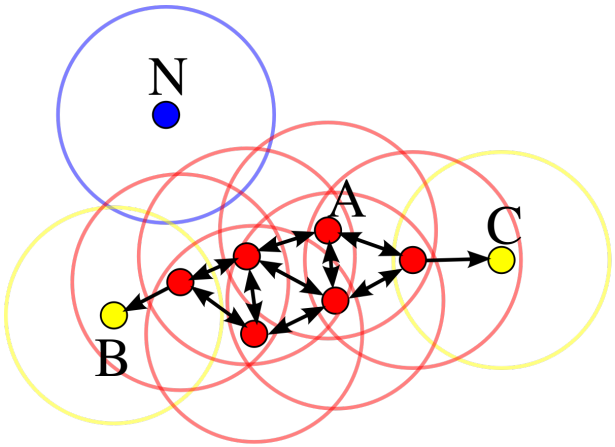


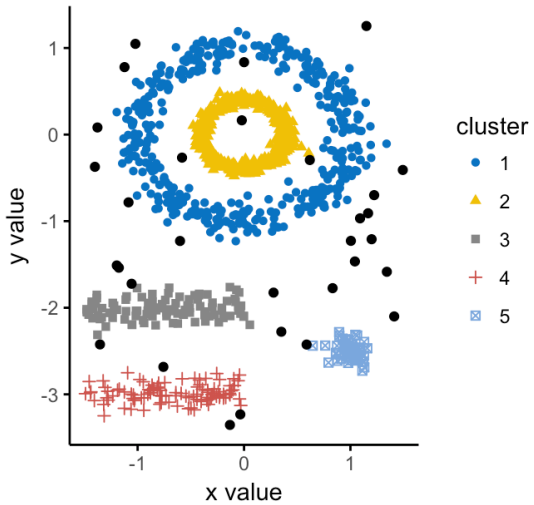


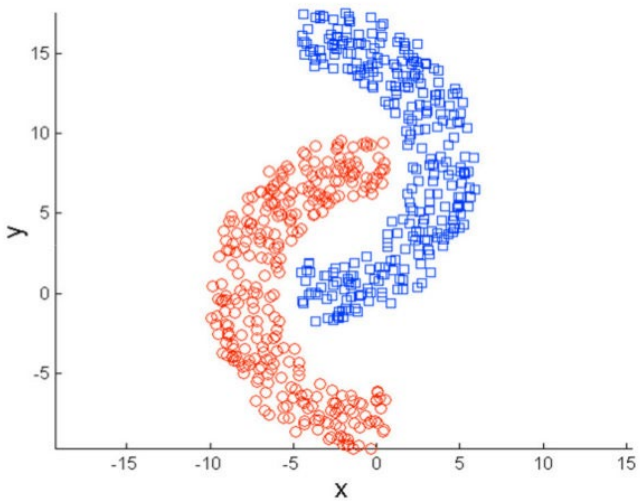
Connectivity models

- The data points closer in data space exhibit more similarity to each other than the data points lying farther away.
- Two approaches:
  - 1) starting with classifying all data points into separate clusters & then aggregating them as the distance decreases;
  - 2) all data points are classified as a single cluster and then partitioned as the distance increases.

Density Models







Clustering Algorithms

- K-Means
- Affinity propagation
- Mean-shift
- Spectral clustering
- Gaussian Mixture Models (GMM)
- Hierarchical clustering
- BIRCH
- DBSCAN

K-Means

- This algorithm requires the number of clusters to be specified.
- It scales well to large number of samples and has been used across a large range of application areas in many different fields.
- The KMeans algorithm clusters data by trying to separate samples in n groups of equal variance, minimizing a criterion known as the inertia or within-cluster sum-of-squares.

KMeans

- The K-means algorithm divides a set of N samples X into K disjoint clusters C, each described by the mean μ_j of the samples in the cluster.
- The means are commonly called the cluster "centroids".
- The K-means algorithm aims to choose centroids that minimise the inertia, or within-cluster sum of squared criterion:
- $\sum_{i=0}^{n} \min_{\mu_i \in C} \left(\left| \left| x_j - \mu_i \right| \right|^2 \right)$

KMeans Clustering

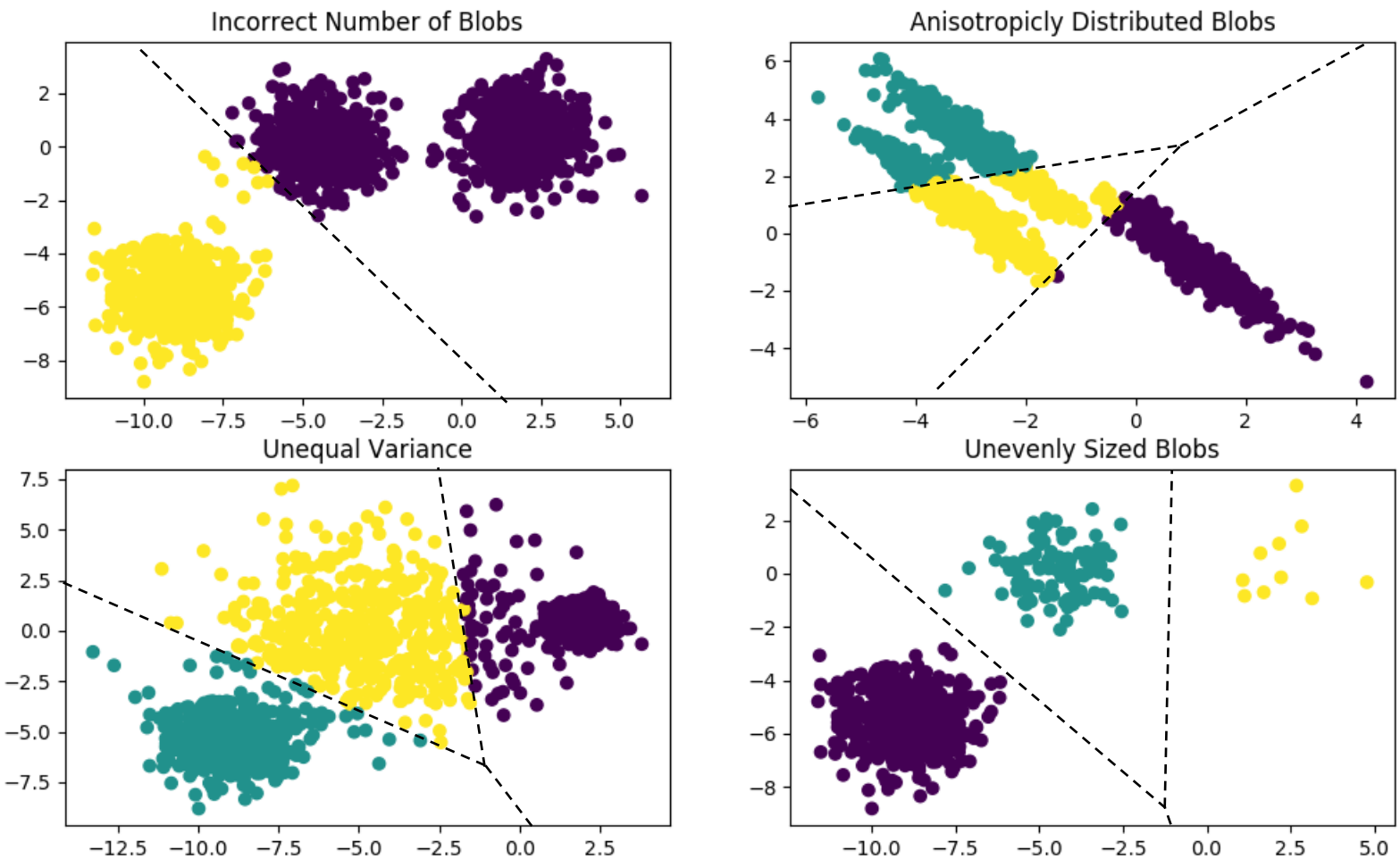
- Partitioning clustering approach
- Each cluster is associated with a centroid (center point)
- Each point is assigned to the cluster with the closest centroid
- Number of clusters, K, must be specified

```
Select K points as the initial centroids.
repeat
    Form K clusters by assigning all points to the closest centroid.
    Recompute the centroid of each cluster.
until The centroids don't change
```

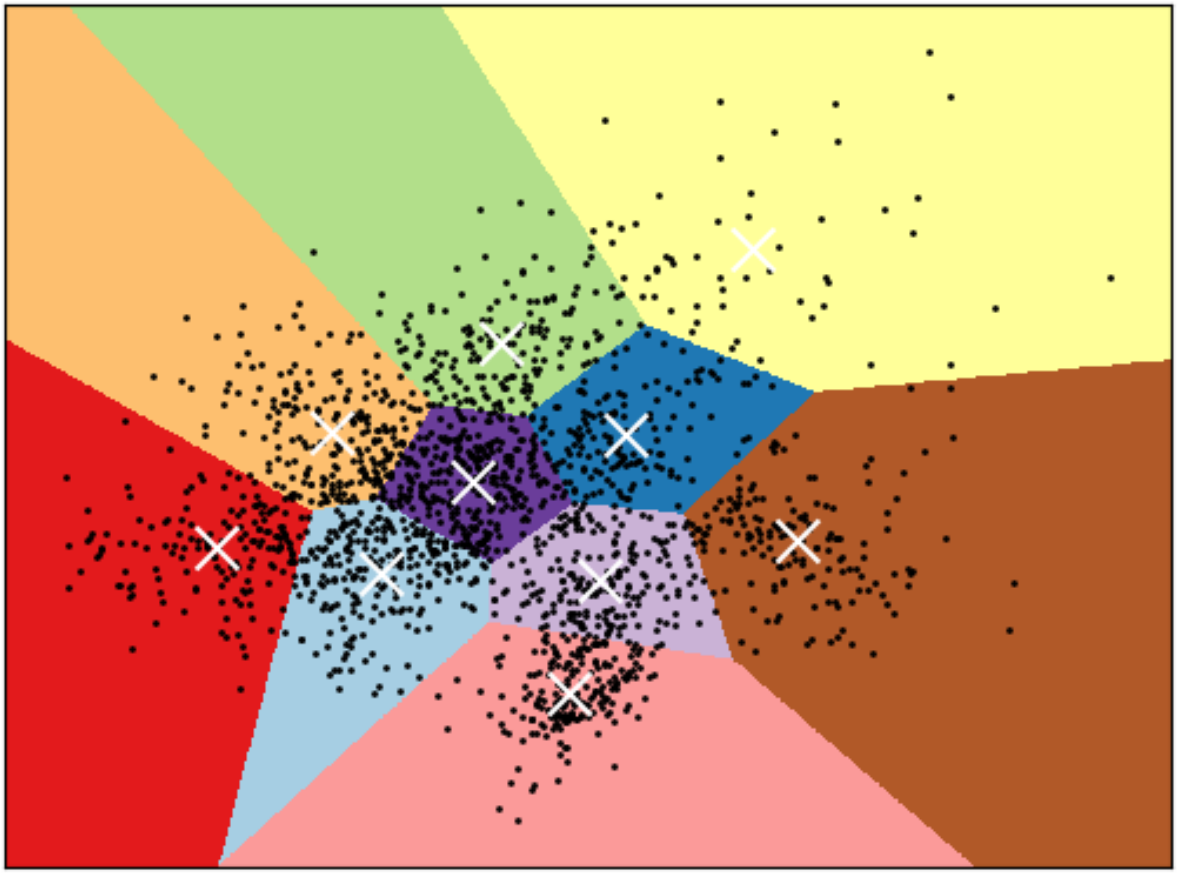
KMeans

- Inertia, or the within-cluster sum of squares criterion, can be recognized as a measure of how internally coherent clusters are.
- Drawbacks:
  - Inertia makes the assumption that clusters are convex and isotropic. It responds poorly to elongated clusters, or manifolds with irregular shapes.
  - Inertia is not a normalized metric
  - Curse of dimensionality

Drawbacks of KMeans



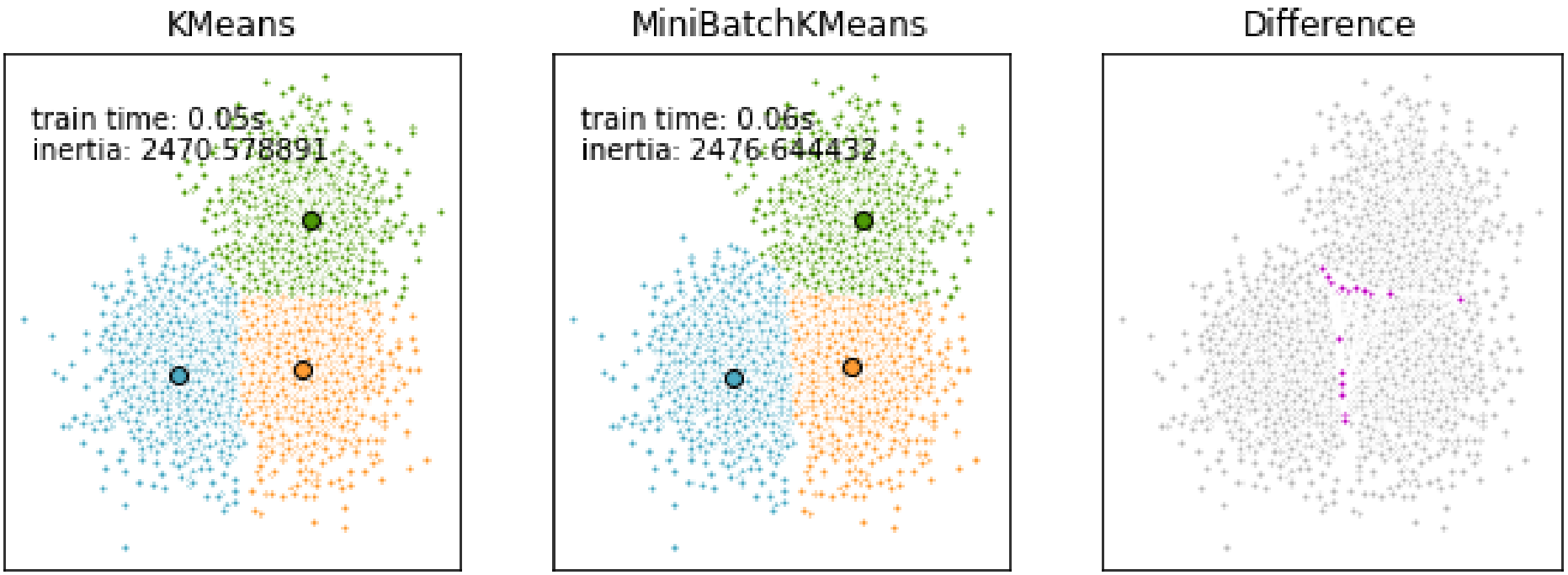
Voronoi Diagram



Mini Batch K-Means

- Mini-batches are subsets of the input data, randomly sampled in each training iteration.
- The algorithm iterates between two steps.
- Step 1: *b* samples are drawn randomly from the dataset, to form a mini-batch. These are then assigned to the nearest centroid.
- Step 2: the centroids are updated. For each sample in the mini-batch, the assigned centroid is updated by taking the streaming average of the sample and all previous samples assigned to that centroid.

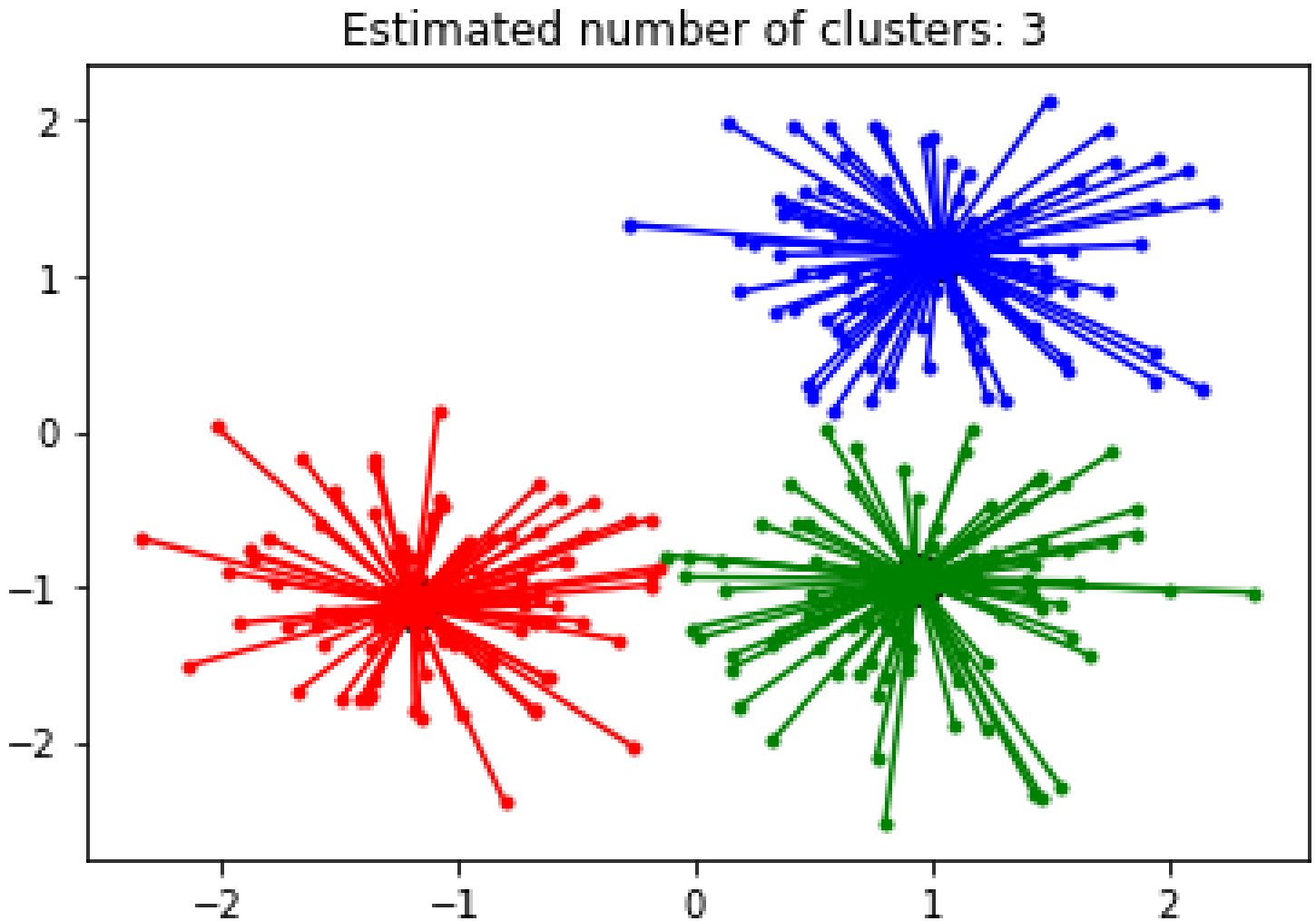
K-Means and


KMeans
train time: 0.05s
inertia: 2470.578891
MiniBatchKMeans
train time: 0.06s
inertia: 2476.644432
Difference

Affinity propagation

- A dataset is described using a small number of exemplars, which are identified as those most representative of other samples.
- The messages sent between pairs represent the suitability for one sample to be the exemplar of the other, which is updated in response to the values from other pairs.
- Updating happens iteratively until convergence to the final clustering.

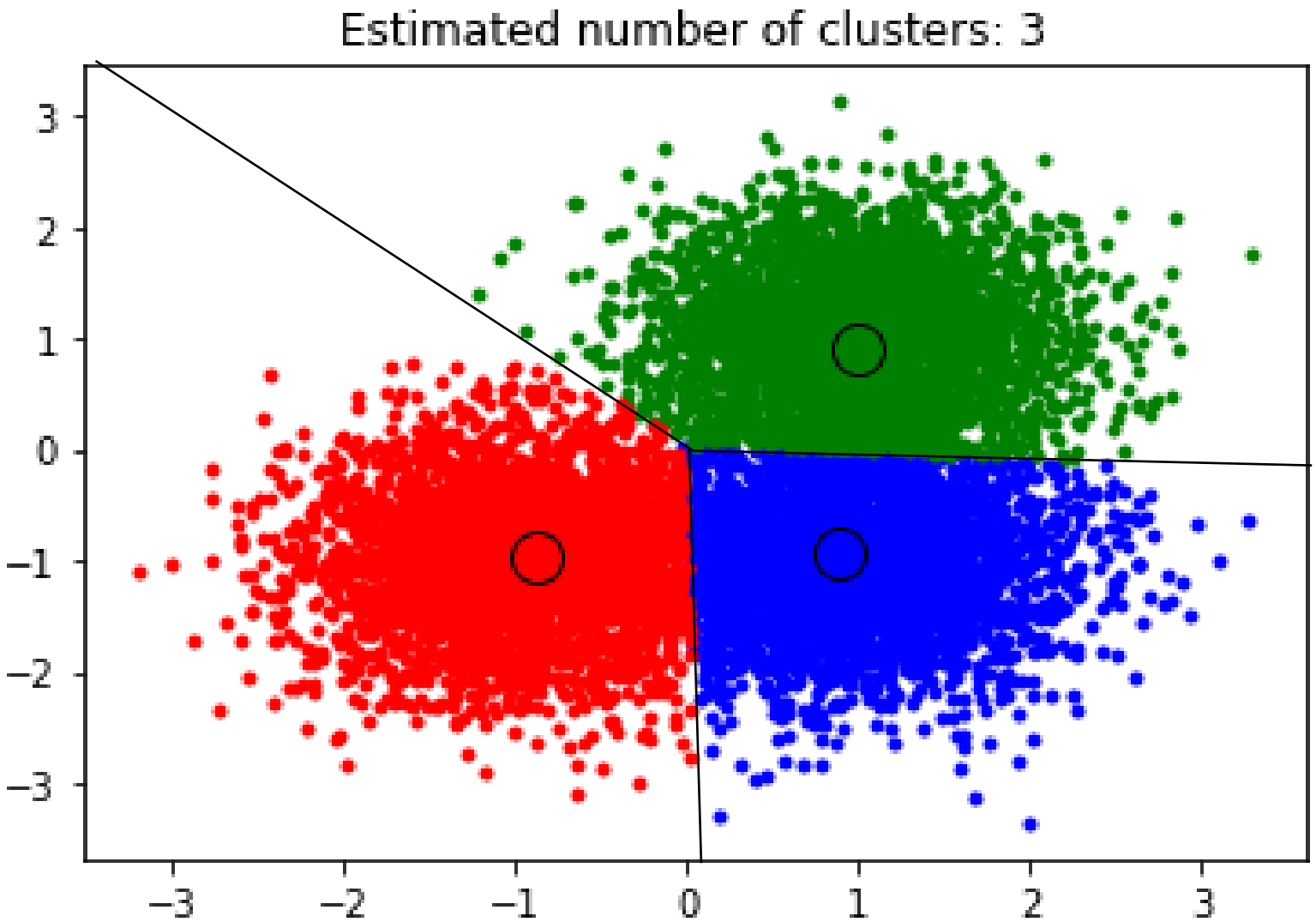
Affinity propagation



Mean-shift

- MeanShift clustering is a centroid based algorithm, updating candidates for centroids to be the mean of the points within a given region.
- Then filtered in a post-processing stage to eliminate near-duplicates to form the final set of centroids.
- $x_i^{\{t+1\}} = x_i^t + m(x_i^t)$
- $m(x_i) = \dfrac{\sum_{x_j \in N(x_i)} K(x_j - x_i) x_j}{\sum_{x_j \in N(x_i)} K(x_j - x_i)}$

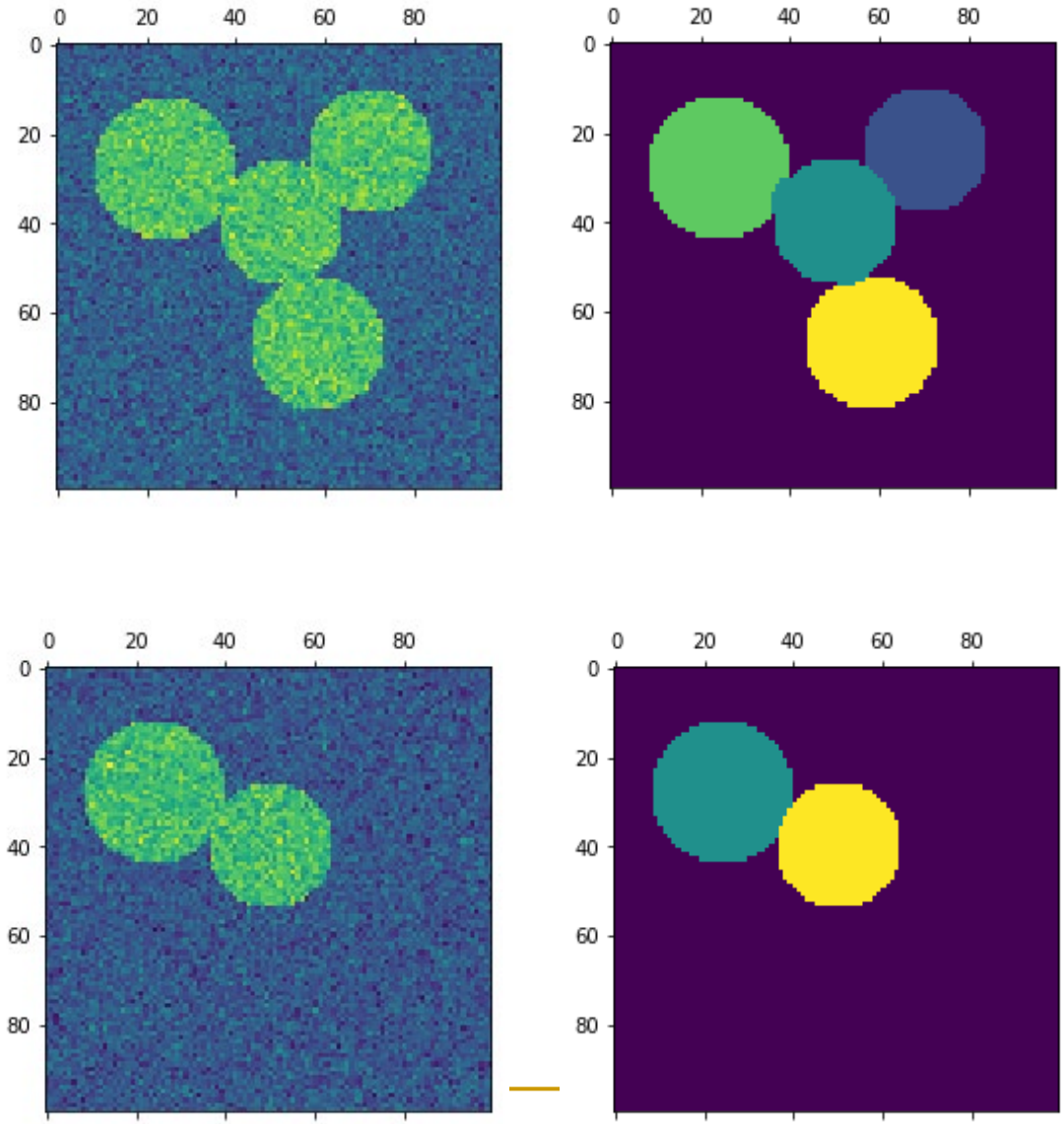
Mean-shift



Spectral clustering

- SpectralClustering does a low-dimension embedding of the affinity matrix between samples, followed by a KMeans in the low dimensional space.
- SpectralClustering requires the number of clusters to be specified.
- It works well for a small number of clusters.
- This criteria is especially interesting when working on images: graph vertices are pixels, and edges of the similarity graph are a function of the gradient of the image.

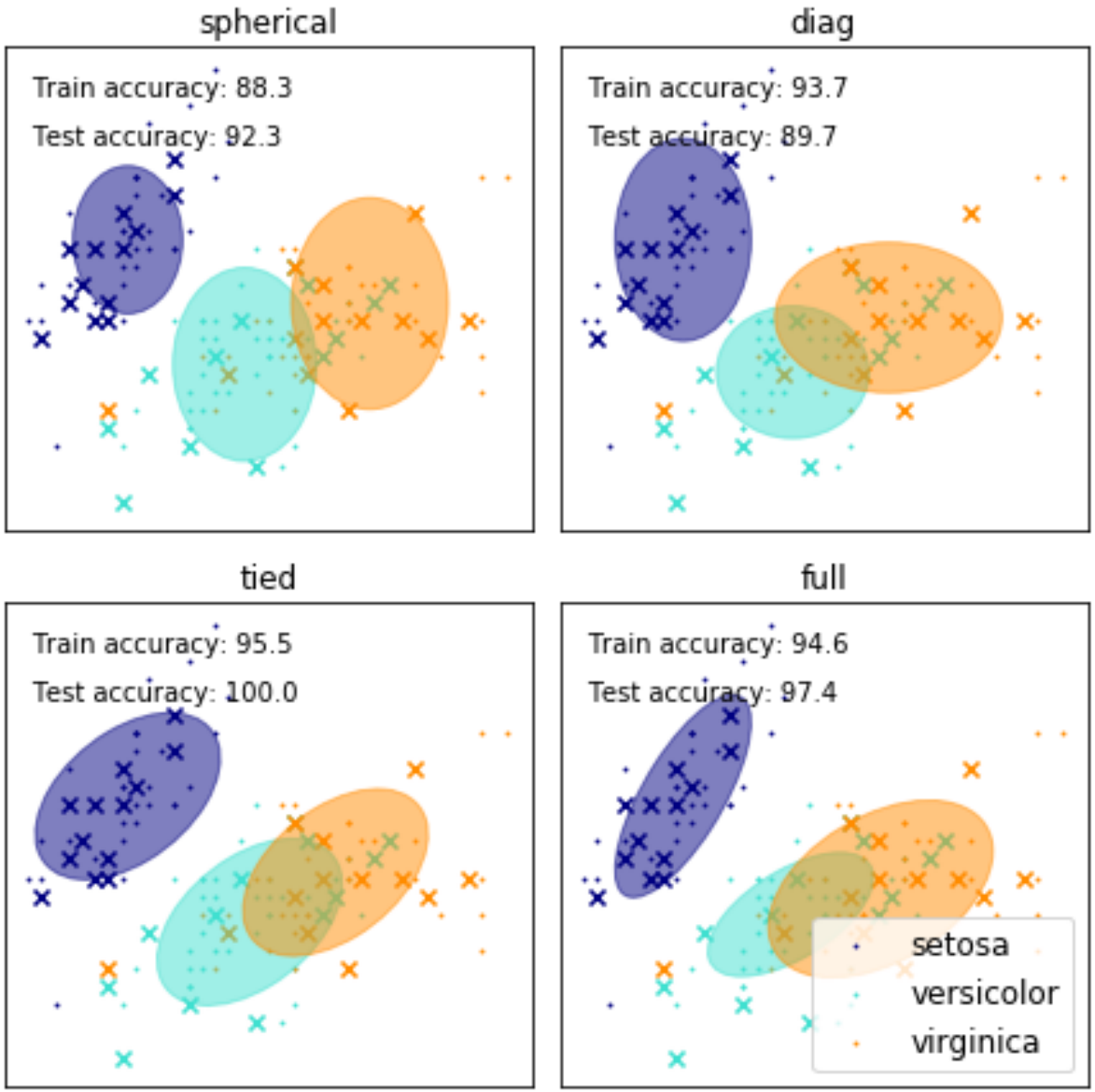
Spectral clustering



Gaussian mixtures

- The GaussianMixture object implements the expectation-maximization (EM) algorithm for fitting mixture-of-Gaussian models.
- It can also draw confidence ellipsoids for multivariate models, and compute the Bayesian Information Criterion to assess the number of clusters in the data.

Gaussian Mixture Model



Hierarchical clustering

- Hierarchical clustering builds nested clusters by merging or splitting them successively.
- This hierarchy of clusters is represented as a tree (or dendrogram).
- The root of the tree is the unique cluster that gathers all the samples, the leaves being the clusters with only one sample.
- The Agglomerative Clustering object performs a hierarchical clustering using a bottom up approach: each observation starts in its own cluster, and clusters are successively merged together.

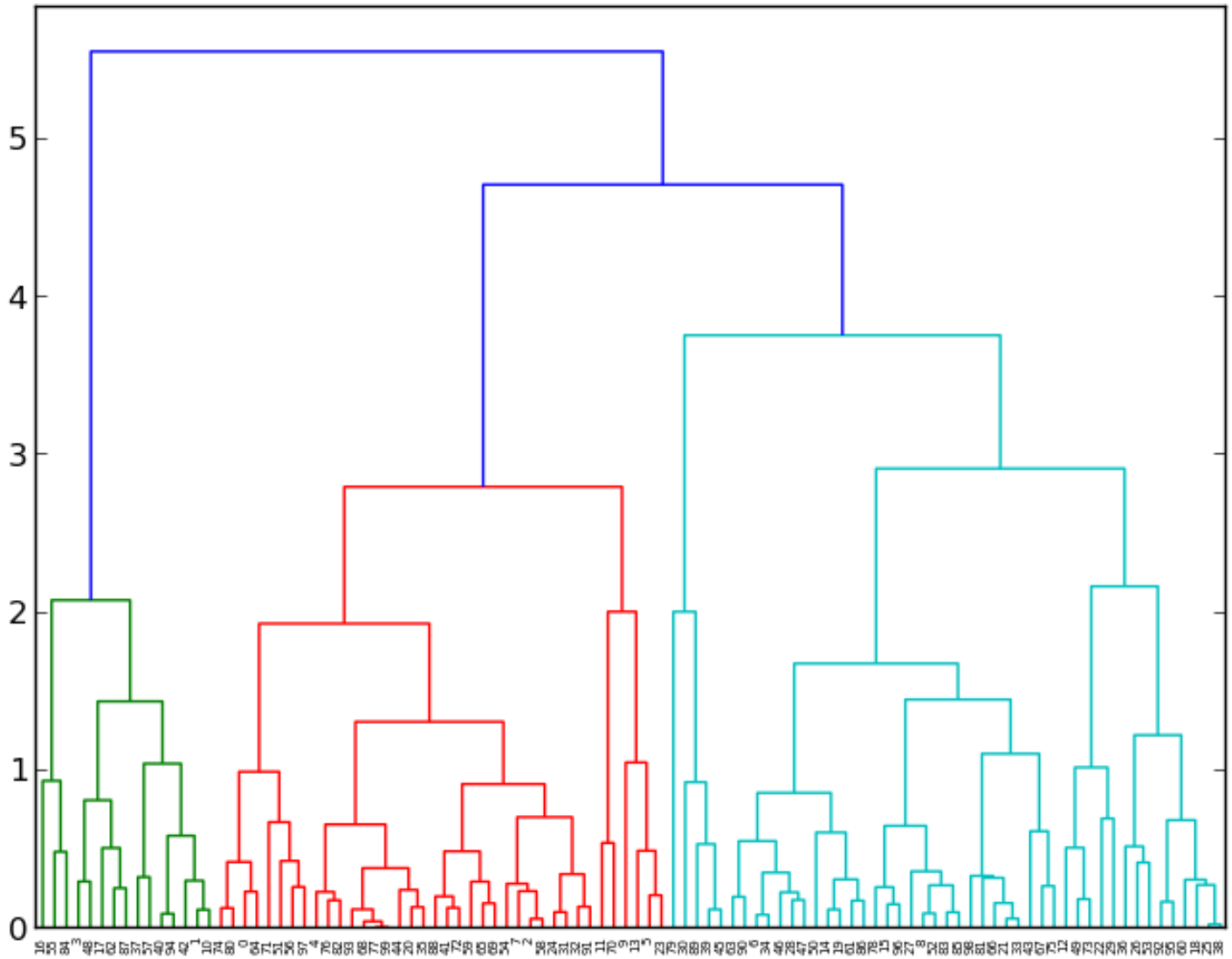
Agglomerative clustering

- Agglomerative Clustering can also scale to large number of samples when it is used jointly with a connectivity matrix, but is computationally expensive when no connectivity constraints are added between samples.
- It considers at each step all the possible merges.

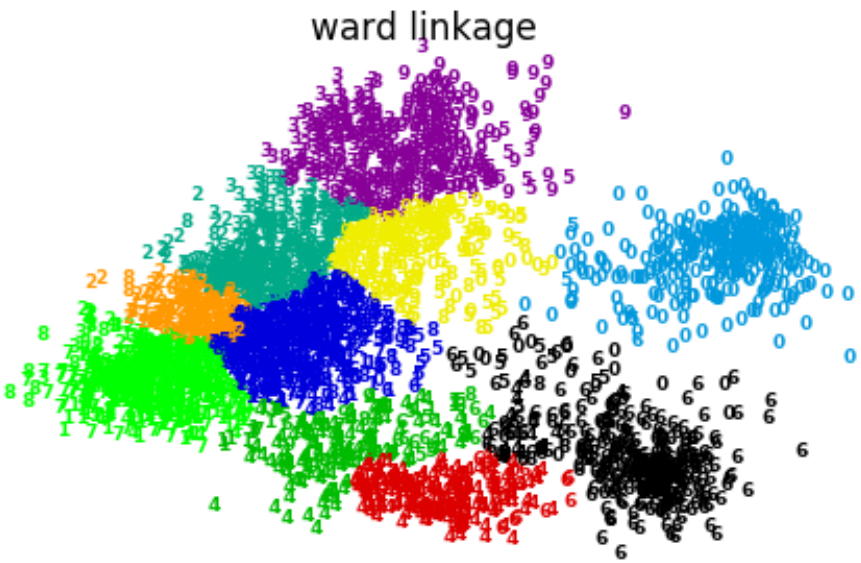
Linkage criteria

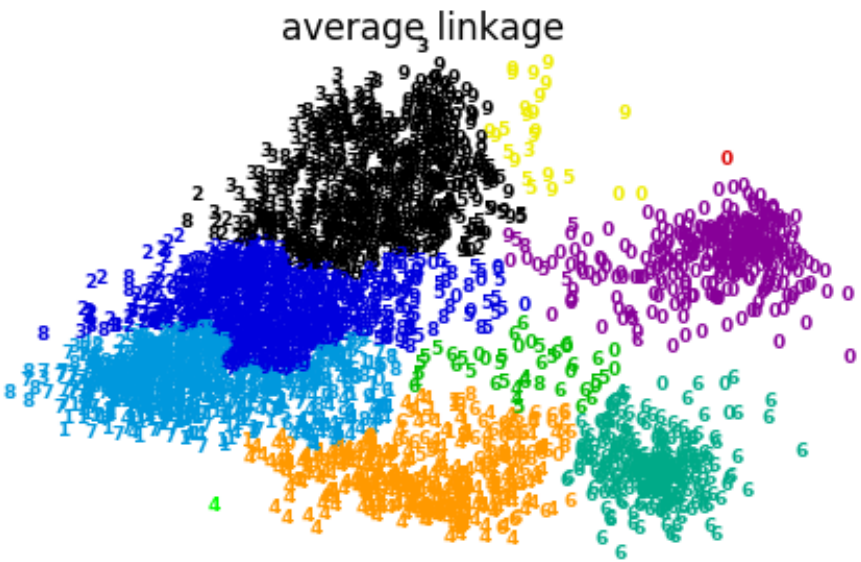
- The linkage criteria determines the metric used for the merge strategy.
- Minimum or single linkage
- Maximum or complete linkage minimizes the maximum distance between observations of pairs of clusters.
- Average linkage minimizes the average of the distances between all observations of pairs of clusters.
- Ward hierarchical clustering minimizes the sum of squared differences within all clusters.
- It is a variance-minimizing approach and in this sense is similar to the k-means objective function but tackled with an agglomerative hierarchical approach.

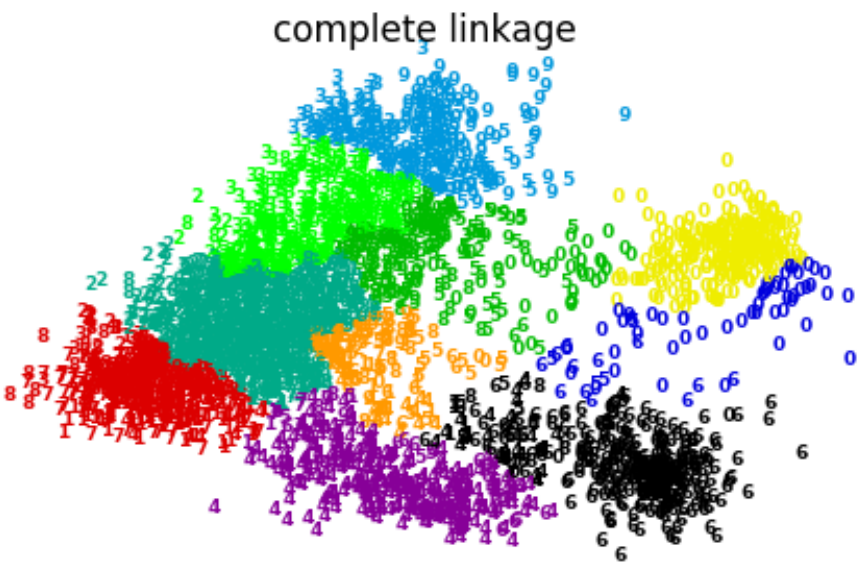
Dendrogram



Hierarchical Clustering Comparison







BIRCH

- The BIRCH builds a tree called the Characteristic Feature Tree (CFT) for the given data.
- The data is essentially lossy compressed to a set of Characteristic Feature nodes (CF Nodes).
- The CF Nodes have a number of subclusters called Characteristic Feature subclusters (CF Subclusters) and these CF Subclusters located in the non-terminal CF Nodes can have CF Nodes as children.
- The CF Subclusters hold the necessary information for clustering which prevents the need to hold the entire input data in memory. This information includes:

BIRCH

- Number of samples in a subcluster.
- Linear Sum - an N-dimensional vector holding the sum of all samples
- Squared Sum - Sum of the squared L2 norm of all samples.
- Centroids - To avoid recalculation linear sum / n_samples.
- Squared norm of the centroids.
- The BIRCH algorithm has two parameters: the threshold and the branching factor.
- The branching factor limits the number of subclusters in a node and the threshold limits the distance between the entering sample and the existing subclusters.

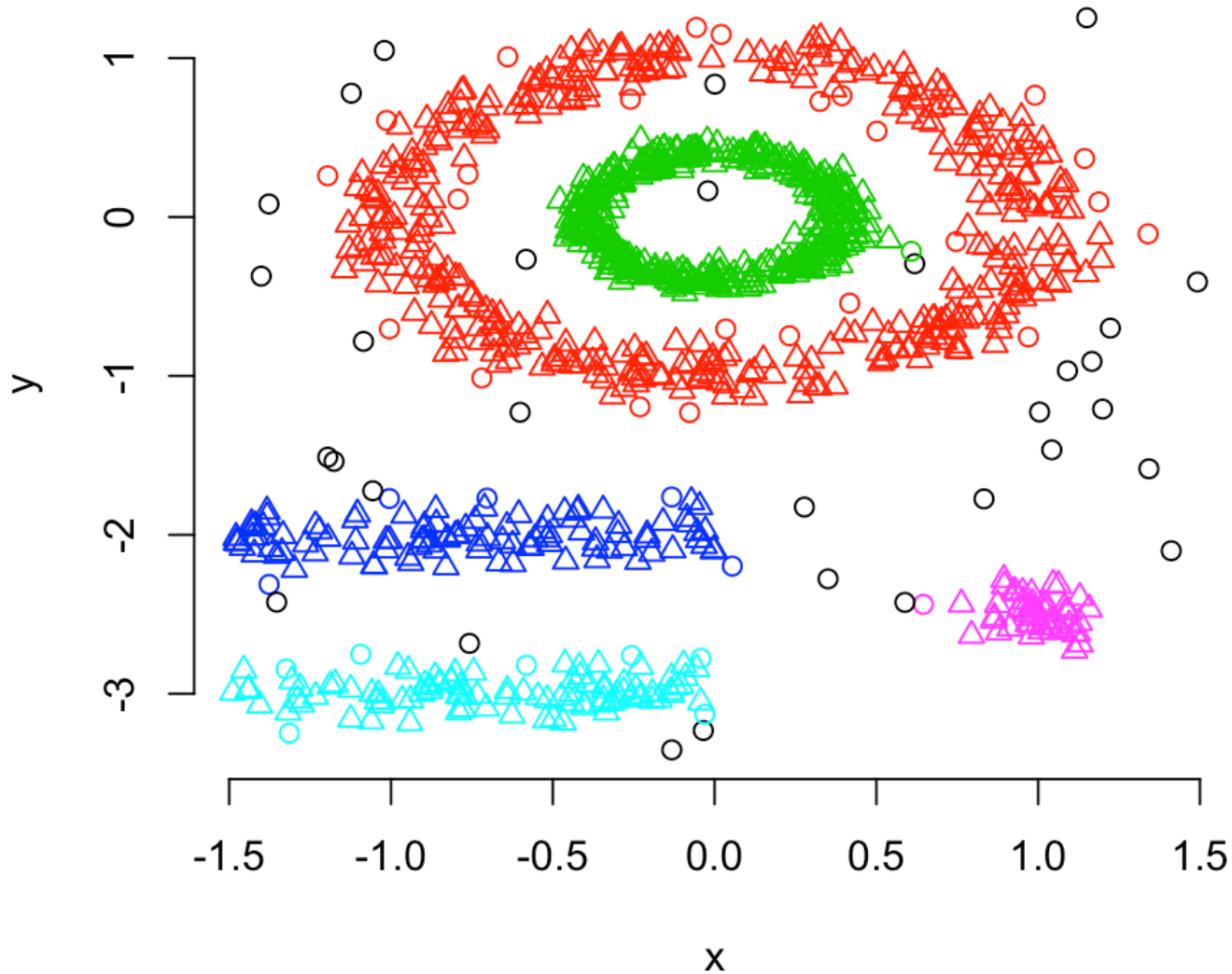
DBSCAN

- The DBSCAN algorithm views clusters as areas of high density separated by areas of low density.
- Clusters found by DBSCAN can be any shape.
- The central component to the DBSCAN is the concept of core samples, which are samples that are in areas of high density.
- A cluster is a set of core samples, each close to each other (measured by some distance measure) and a set of non-core samples that are close to a core sample (but are not themselves core samples).
- There are two parameters to the algorithm, min_samples and eps, which define formally what we mean when we say dense.
- Higher min_samples or lower eps indicate higher density necessary to form a cluster.

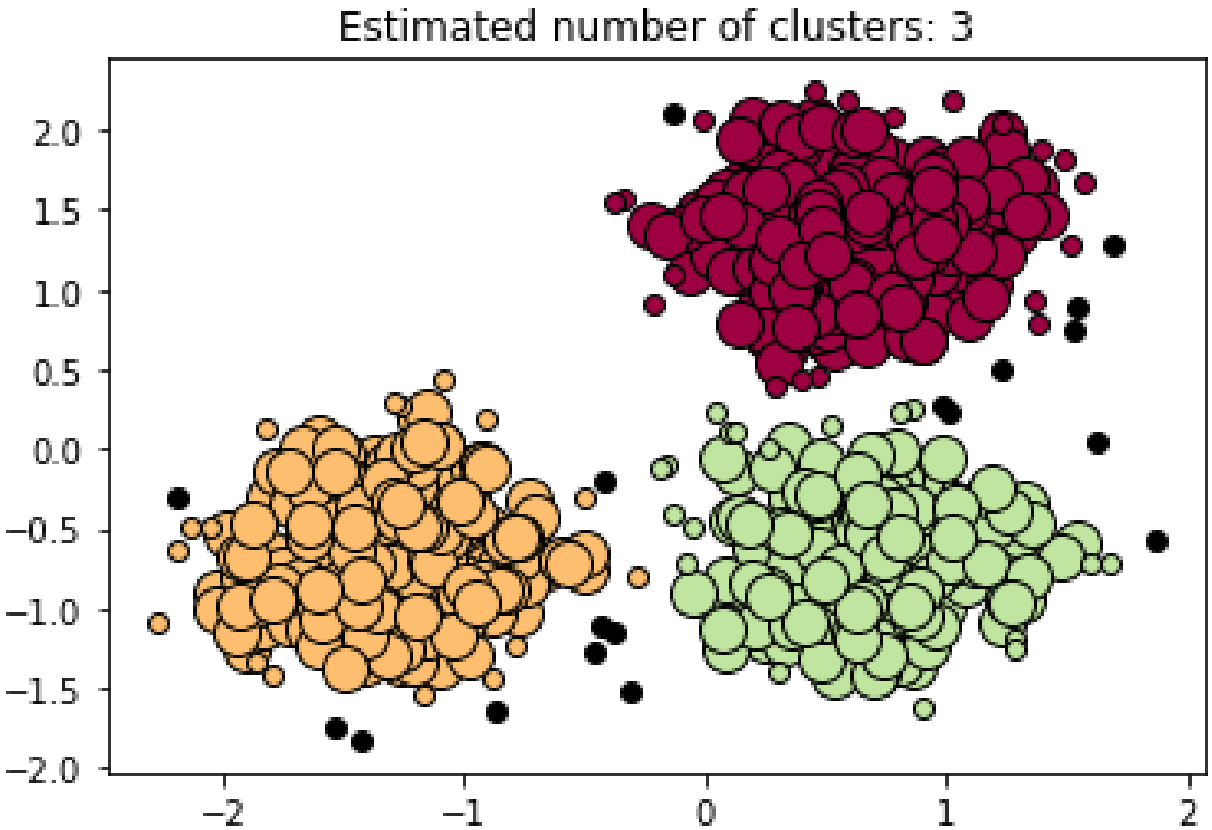
DBSCAN

- A core sample as being a sample in the dataset such that there exist min_samples other samples within a distance of eps, which are defined as neighbors of the core sample.
- The core sample is in a dense area of the vector space.
- A cluster is a set of core samples that can be built by recursively taking a core sample, finding all of their neighbors that are core samples, and so on.
- A cluster also has a set of non-core samples, which are samples that are neighbors of a core sample in the cluster but are not themselves core samples called border.
- Any sample that is not a core sample or border is an outlier.

DBSCAN



DBSCAN



Clustering performance evaluation

- Adjusted Mutual Information
- Rand index
- Calinski and Harabaz score
- Davies-Bouldin score
- Completeness metric of a cluster labeling given a ground truth.
- Contingency matrix describing the relationship between labels.
- Fowlkes Mallows Score

Clustering performance evaluation

- Homogeneity
- Completeness
- V-Measure scores
- Mutual Information between two clusterings
- Normalized Mutual Information between two clusterings
- Silhouette Score
- Silhouette Samples

Cohesion and Separation

- Cohesion: Measures how closely related are objects in a cluster
- Separation: Measure how distinct or well-separated a cluster is from other clusters
- Sum Squared Error (SSE)
  - Cohesion is measured by the within cluster sum of squares (SSE)

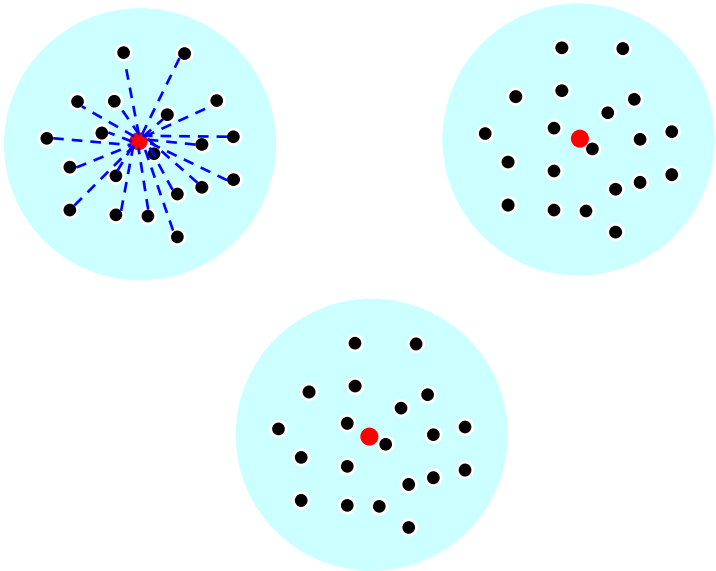
$$\boldsymbol{WSS} = \sum_{\boldsymbol{i}} \sum_{\boldsymbol{x} \in \boldsymbol{C}_i} (\boldsymbol{x} - \boldsymbol{m}_{\boldsymbol{i}})^2$$

  - Separation is measured by the between cluster sum of squares

$$BSS = \sum_i |C_i| (m - m_i)^2$$

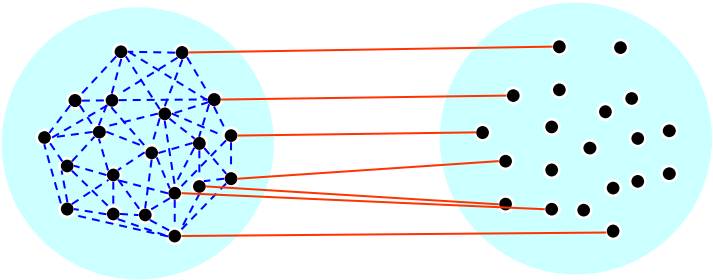
Sum Squared Error (SSE)

$$SSE = \sum_{j=1}^{C} \sum_{i=1}^{N_j} \left(dist(\mathbf{x}_i, \mathbf{m}_j) \right)^2$$



Cohesion & Separation

- A proximity graph based approach
  - Cluster cohesion is the sum of the weight of all links within a cluster.
  - Cluster separation is the sum of the weights between nodes in the cluster and nodes outside the cluster.

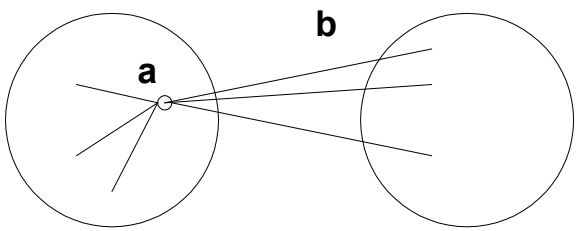


Silhouette Coefficient

- Silhouette combine ideas of both cohesion and separation, but for individual points, as well as clusters and clusterings
- For an individual point, i
  - Calculate a = average distance of i to the points in its cluster
  - Calculate b = min (average distance of i to points in another cluster)
  - The silhouette coefficient for a point is then given by

    $s = 1 - a/b$ if $a < b$, (or $s = b/a - 1$ if $a \geq b$, not the usual case)

  - Typically between 0 and 1.
  - The closer to 1 the better.



- Can calculate the Average Silhouette width for a cluster or a clustering

Summary

- Large amount of data are without labels
- Clustering or Unsupervised learning, machine learning data without targets or labels
- Popular clustering algorithms are based on partitional techniques, connectivity or hierarchical techniques, and density based techniques.
- Each cluster technique has different drawbacks and advantages.
- To select which technique for a particular jobs, we must compare based on the clustering performances, such as SSE, Silhouette, etc.

https://scikit-learn.org/stable/unsupervised_learning.html

2. Unsupervised learning — scikit

scikit-learn.org/stable/unsupervised_learning.html

Install User Guide API Examples Community More Go

Prev Up Next

scikit-learn 1.3.0
Other versions

Please cite us if you use the software.

Toggle Menu

2. Unsupervised learning