DATA MINING 1
Centroid-based Clustering

Dino Pedreschi, Riccardo Guidotti

Revisited slides from Lecture Notes for Chapter 7 “Introduction to Data Mining”, 2nd Edition by Tan, Steinbach, Karpatne, Kumar

UNIVERSITÀ DI PISA

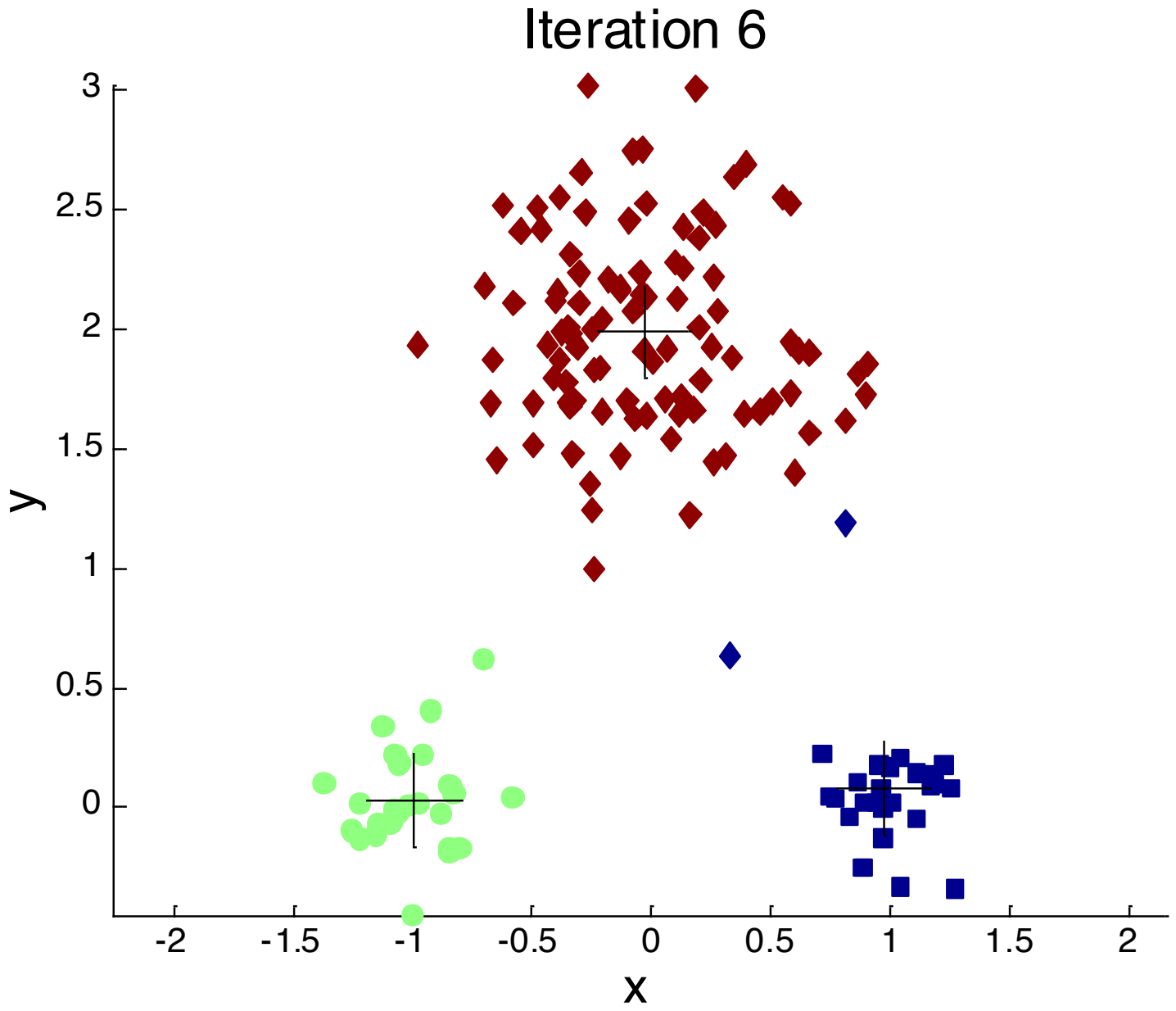
K-Means

K-Means Clustering

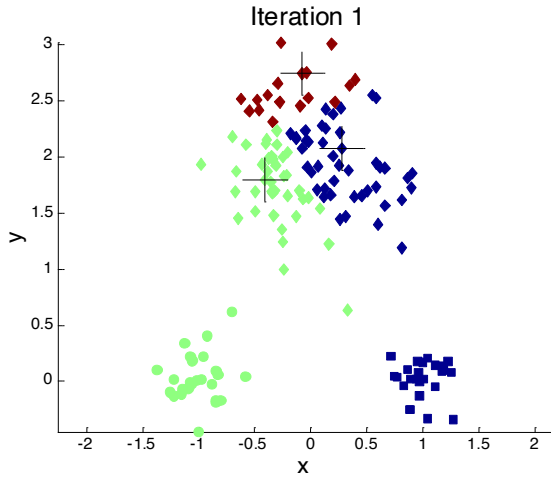
- Partitional clustering approach
- Number of clusters, K, must be specified
- Each cluster is associated with a **centroid** (center point)
- Each point is assigned to the cluster with the **closest centroid**
- The basic algorithm is very simple

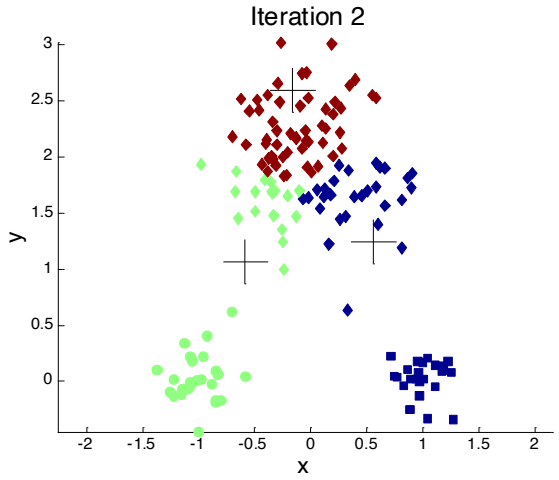
1: Select K points as the initial centroids.
2: **repeat**
3: Form K clusters by assigning all points to the closest centroid.
4: Recompute the centroid of each cluster.
5: **until** The centroids don't change

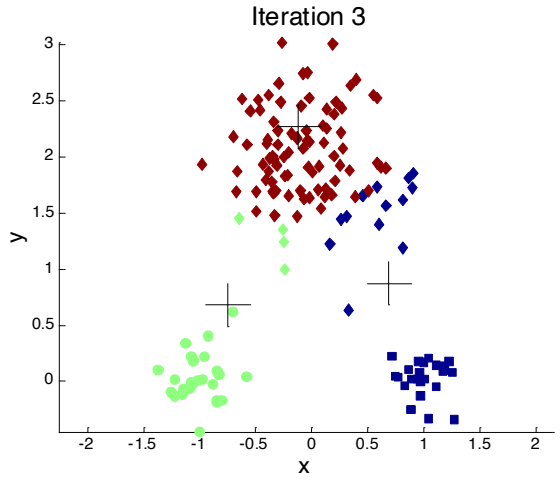
Example of K-Means Clustering

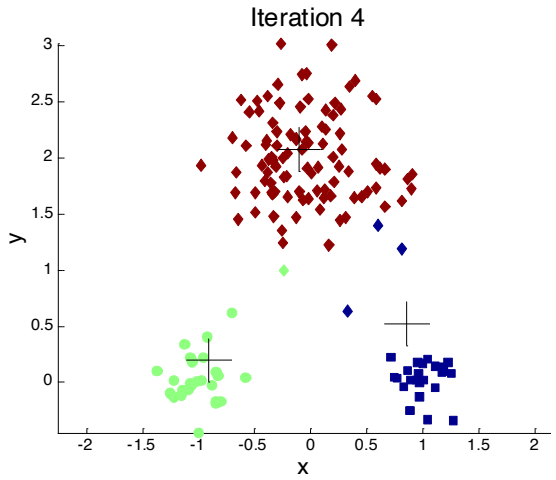


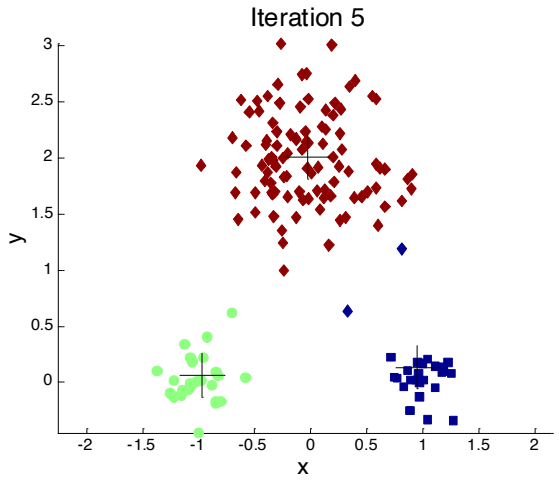
Example of K-Means Clustering

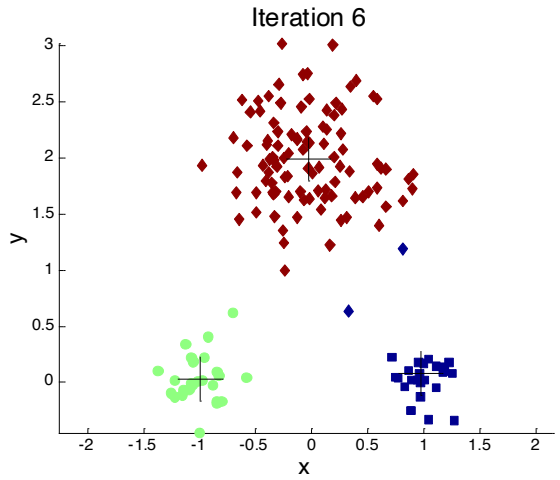








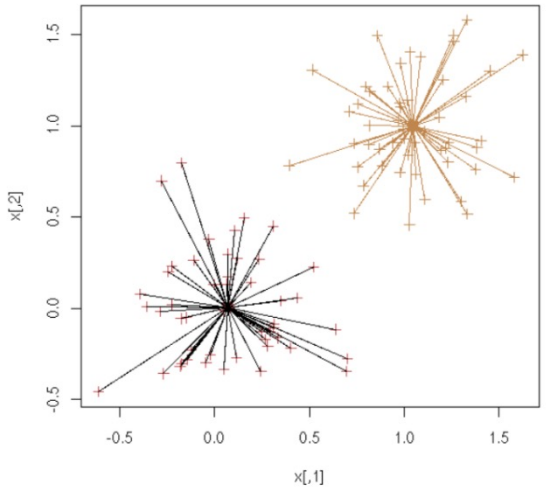




K-Means Clustering – Details

- Initial centroids are often chosen randomly.
  - Clusters produced vary from one run to another.
- The centroid is (typically) the mean of the points in the cluster.
- ‘Closeness’ is measured by Euclidean distance, cosine similarity, correlation, etc.
- K-means will converge for common similarity measures mentioned above.
- Most of the convergence happens in the first few iterations.
  - Often the stopping condition is changed to ‘Until relatively few points change clusters’
- Complexity is O(n * K * I * d)
  - n = number of points, K = number of clusters,
    I = number of iterations, d = number of attributes

Evaluating K-Means Clusters

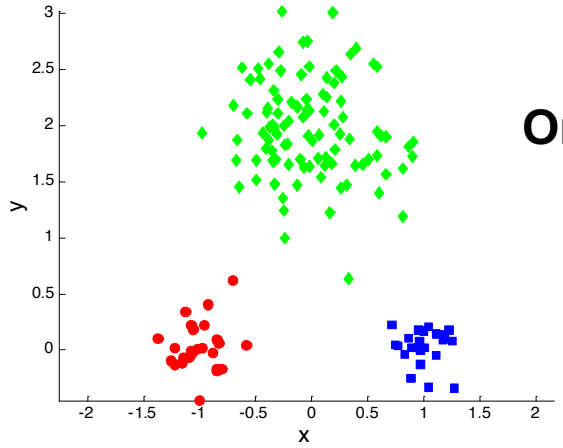


- Most common measure is Sum of Squared Error (SSE)
  - For each point, the error is the distance to the nearest cluster
  - To get SSE, we square these errors and sum them.

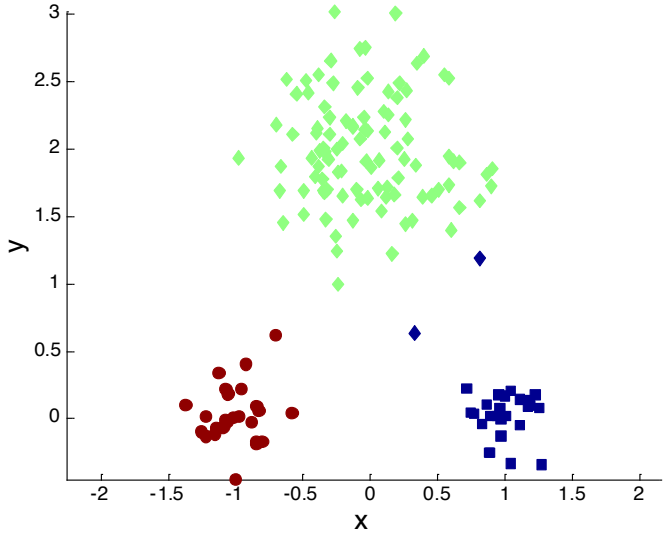
$$SSE = \sum_{i=1}^{K} \sum_{x \in C_i} dist^2(m_i, x)$$

- x is a data point in cluster C_i and m_i is the representative point for cluster Ci
  - can show that mi corresponds to the center (mean) of the cluster
- Given two sets of clusters, we prefer the one with the smallest error
- One easy way to reduce SSE is to increase K, the number of clusters
- A good clustering with smaller K can have a lower SSE than a poor clustering with higher K

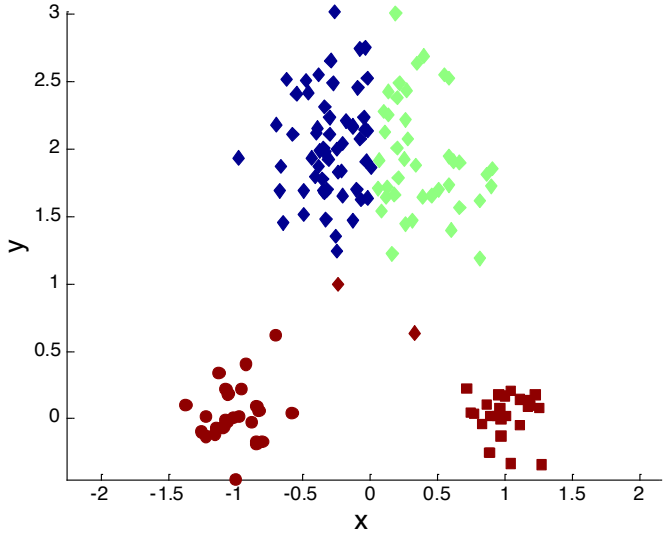
Two different K-Means Clusterings



Original Points



Optimal Clustering



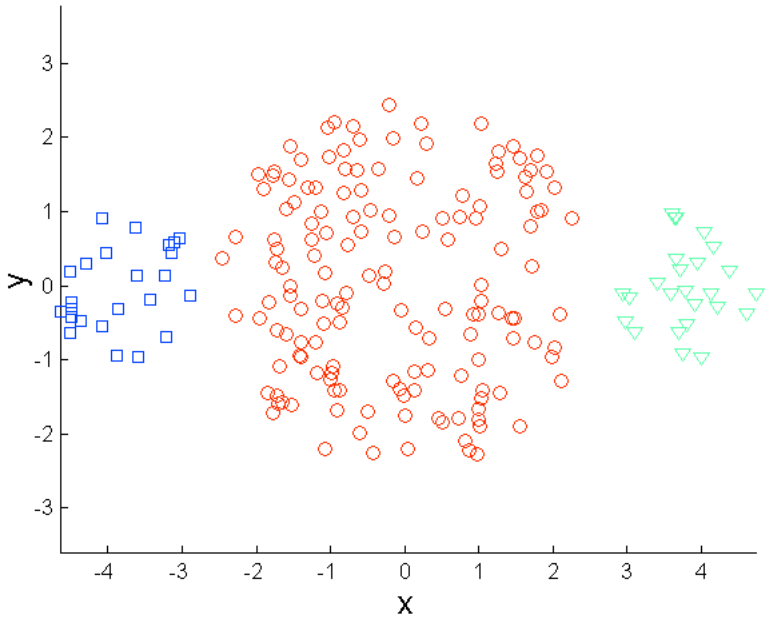
Sub-optimal Clustering

Limitations of K-Means

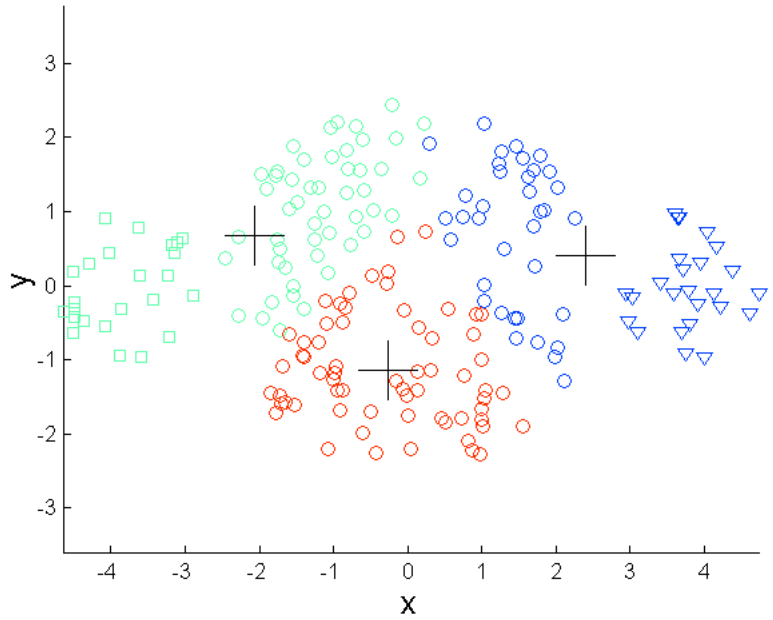
- K-Means has problems when clusters are of differing
  - Sizes
  - Densities
  - Non-globular shapes

- K-Means has problems when the data contains outliers.

Limitations of K-Means: Differing Sizes

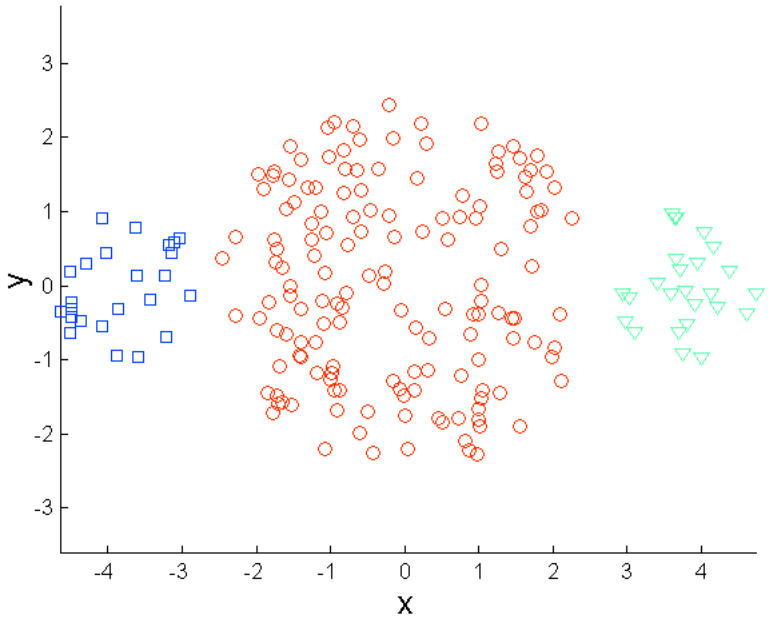


Original Points

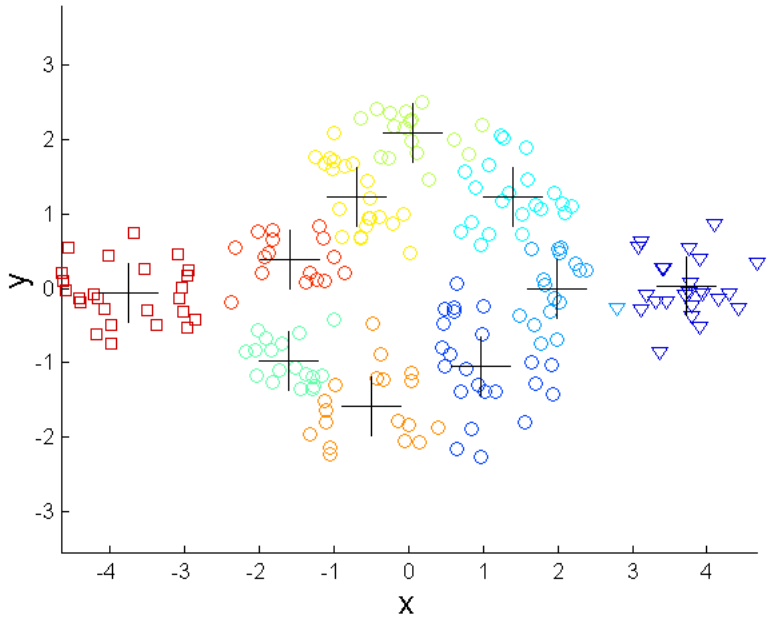


K-means (3 Clusters)

Overcoming K-Means Limitations



Original Points

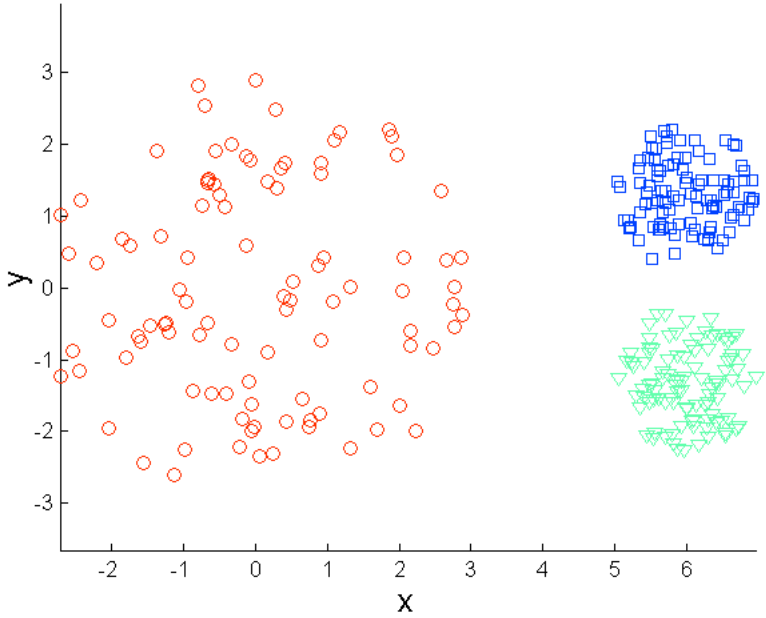


K-means Clusters

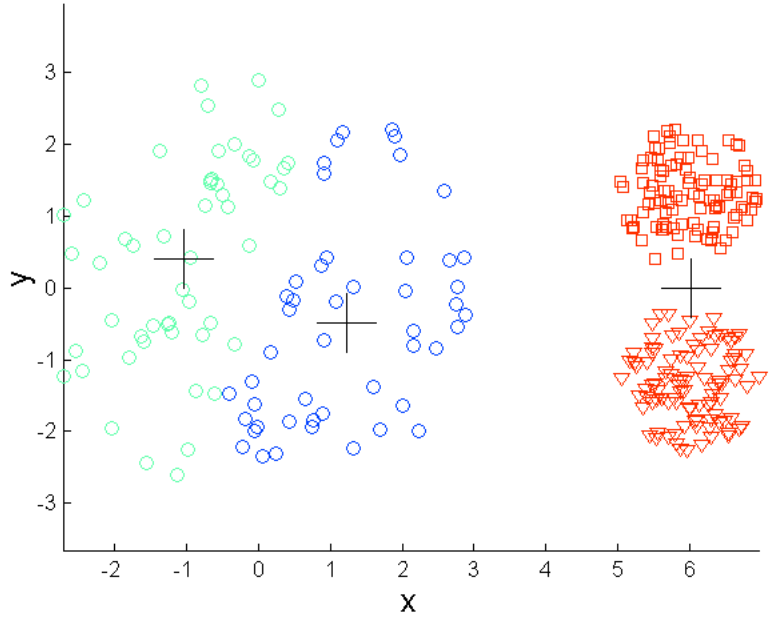
One solution is to use many clusters.

Find parts of clusters, but need to put together.

Limitations of K-Means: Differing Density

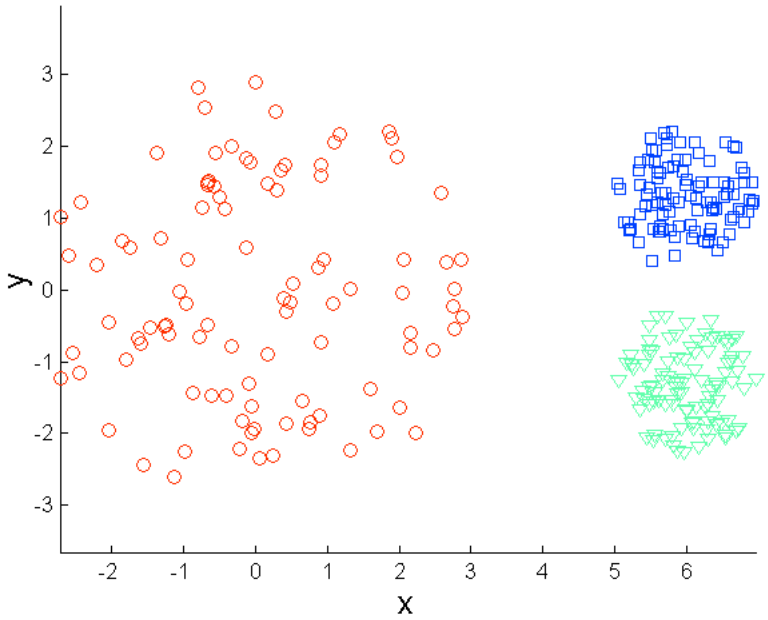


Original Points

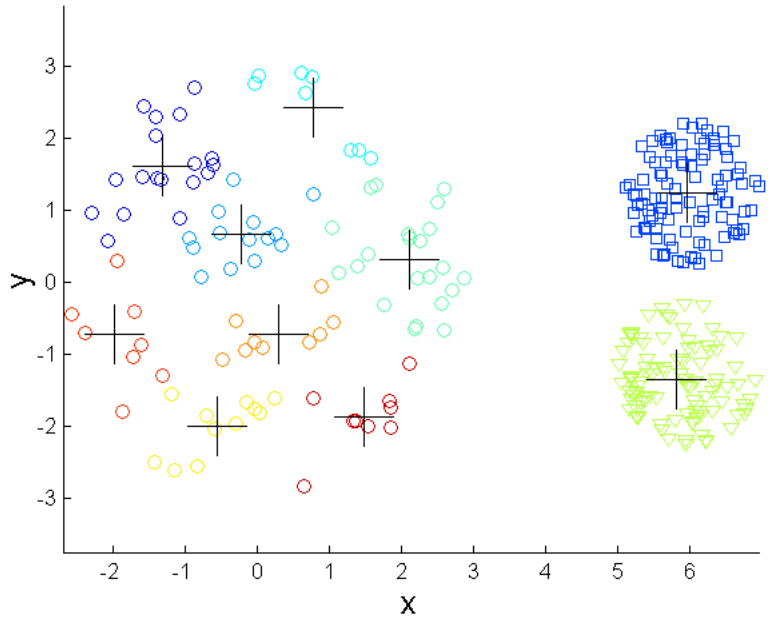


K-means (3 Clusters)

Overcoming K-Means Limitations

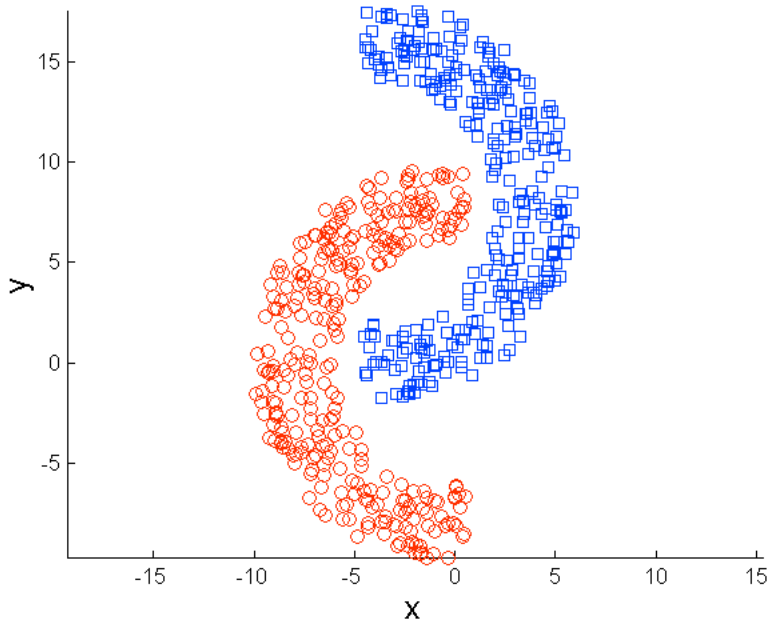


Original Points

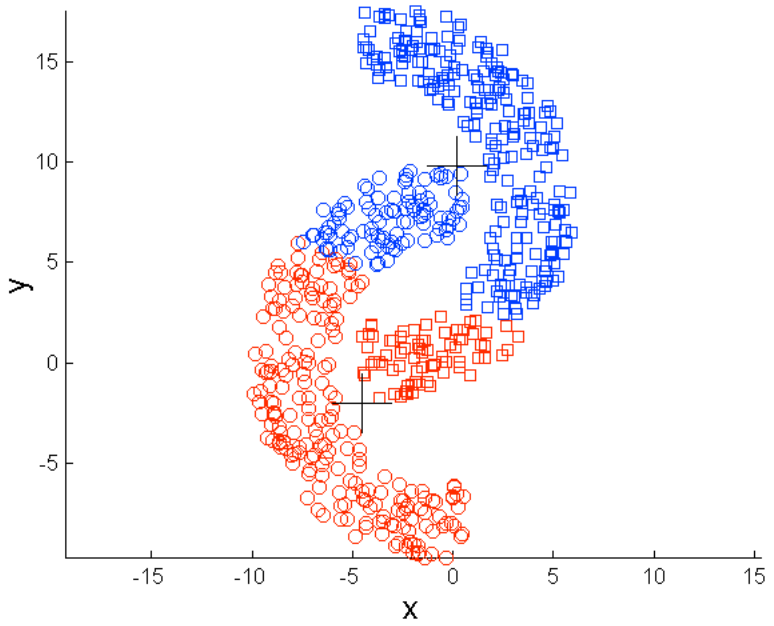


K-means Clusters

Limitations of K-Means: Non-globular Shapes

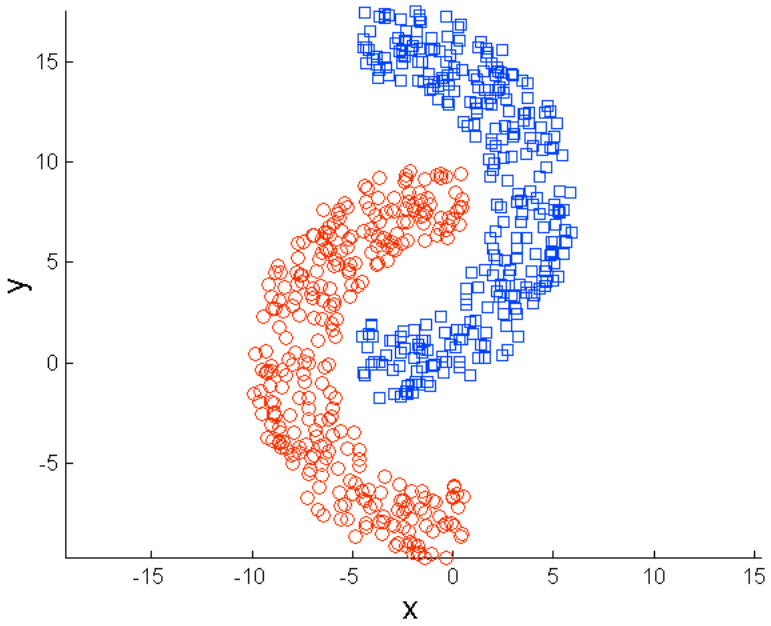


Original Points

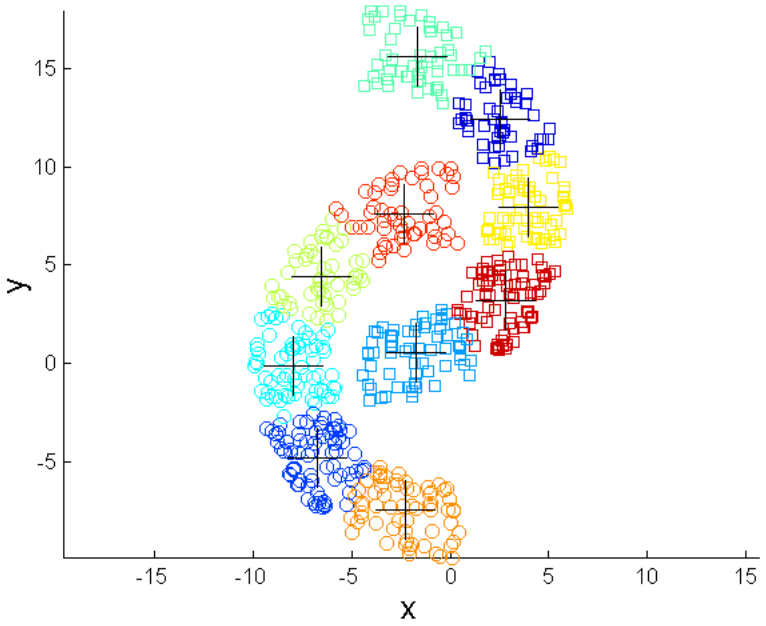


K-means (2 Clusters)

Overcoming K-Means Limitations



Original Points

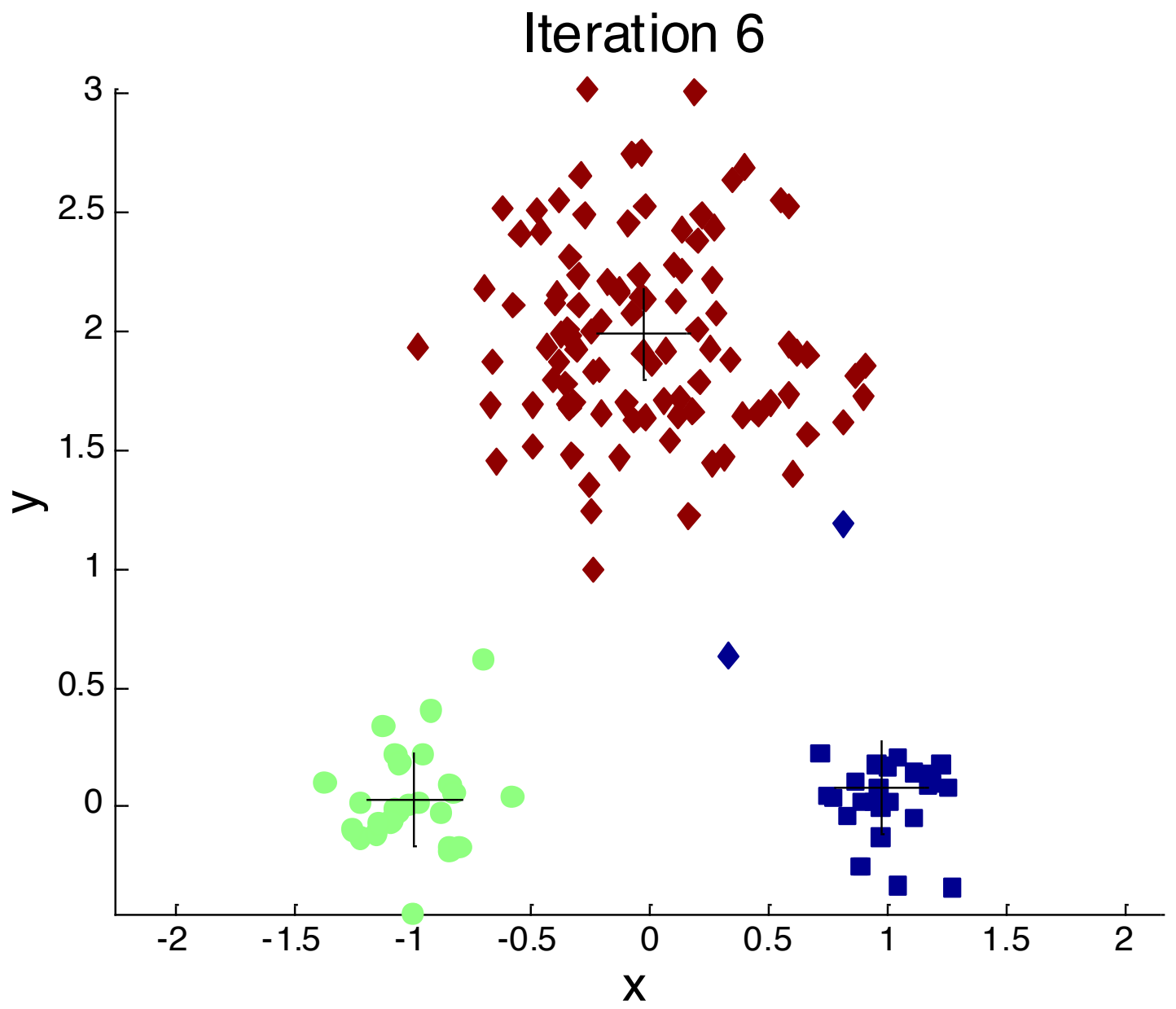


K-means Clusters

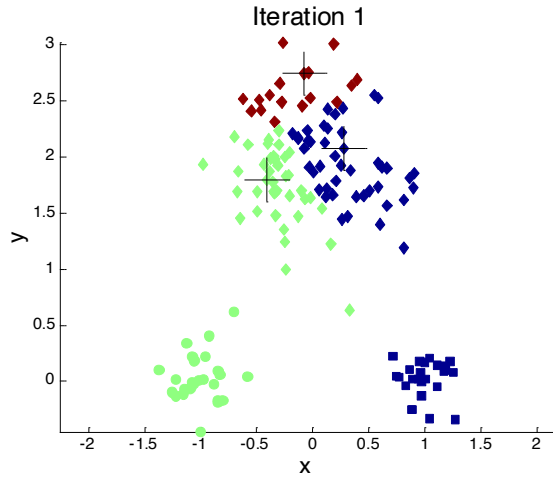
Pre-processing and Post-processing

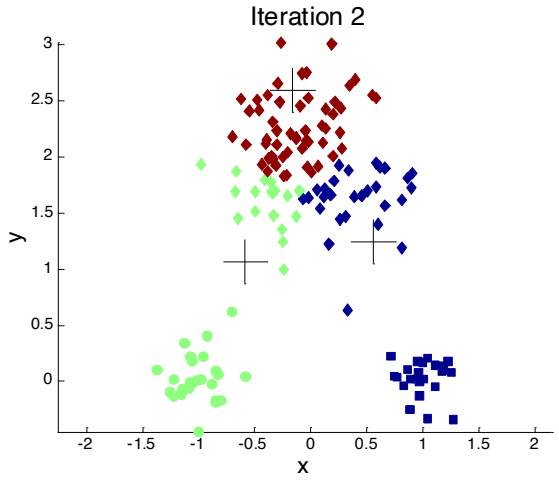
- Pre-processing
  - Normalize the data
  - Eliminate outliers
- Post-processing
  - Eliminate small clusters that may represent outliers
  - Split ‘loose’ clusters, i.e., clusters with relatively high SSE
  - Merge clusters that are ‘close’ and that have relatively low SSE
  - Can use these steps during the clustering process
    - ISODATA

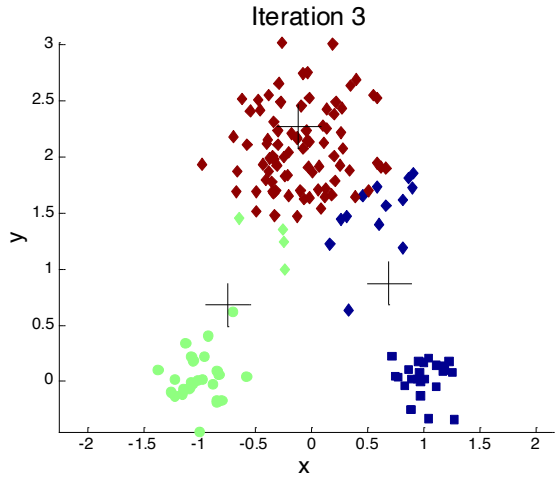
Importance of Choosing Initial Centroids

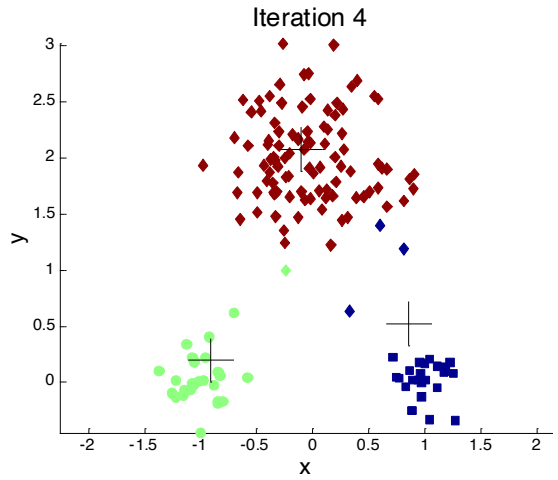


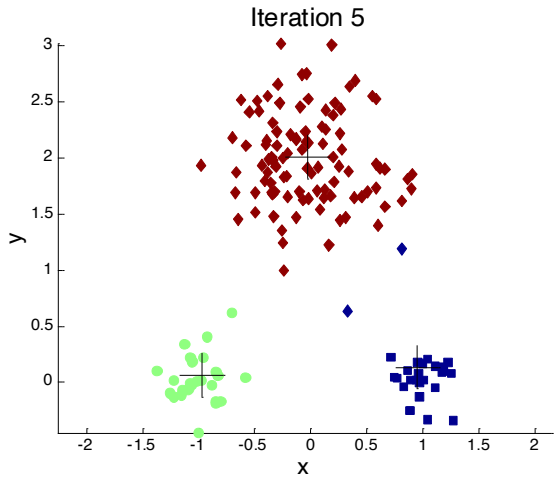
Importance of Choosing Initial Centroids

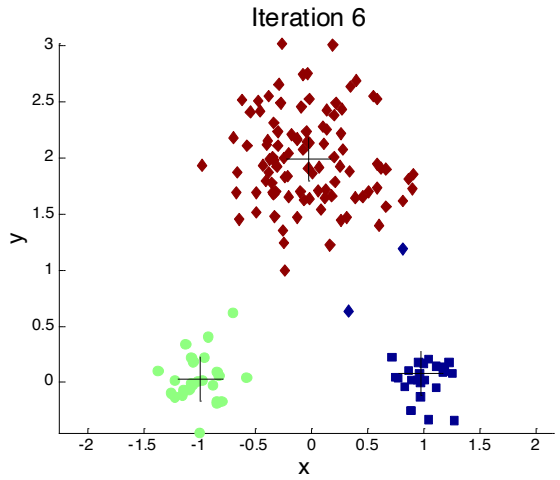




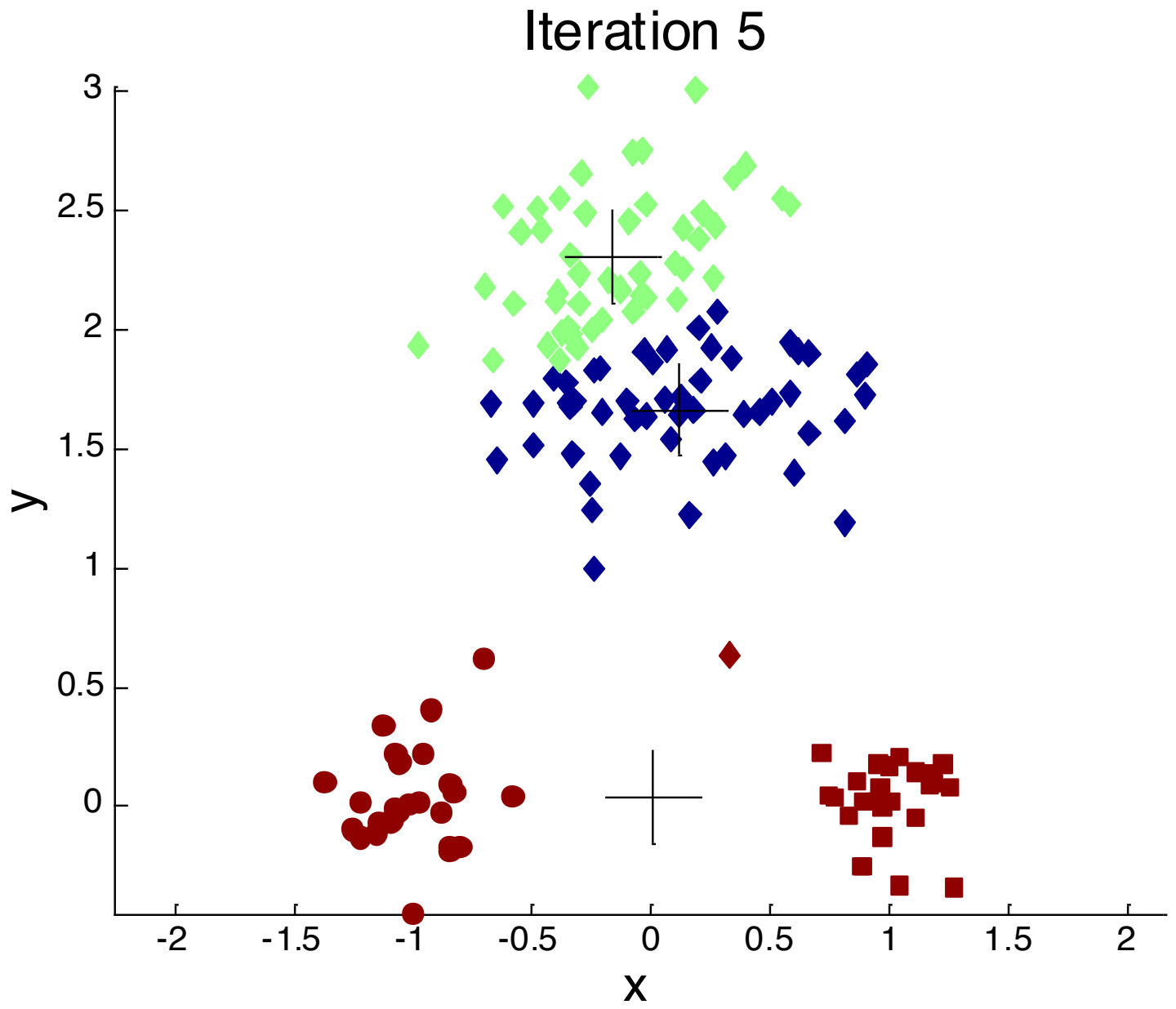




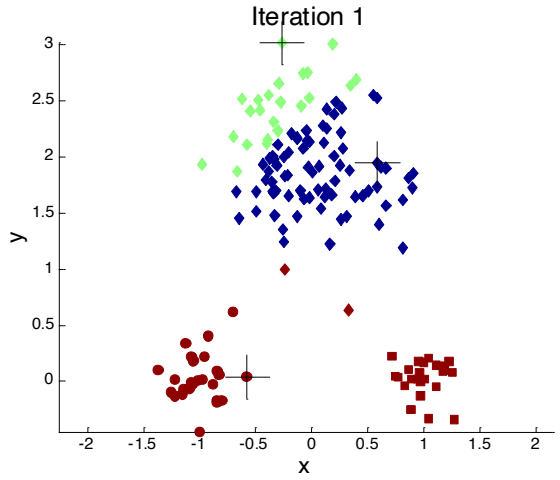


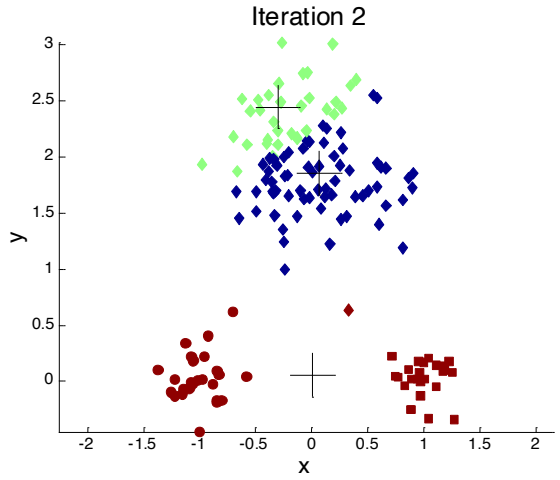


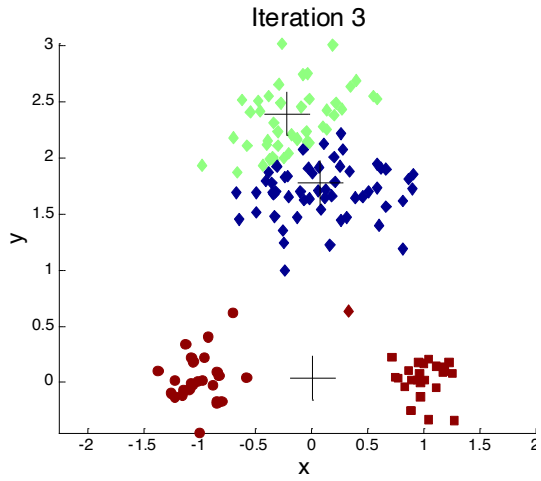
Importance of Choosing Initial Centroids …

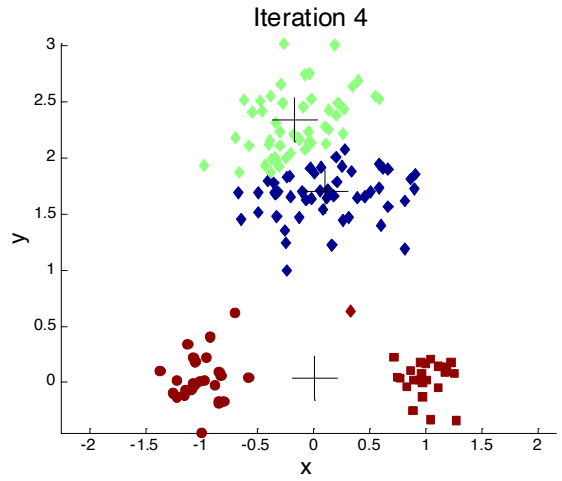


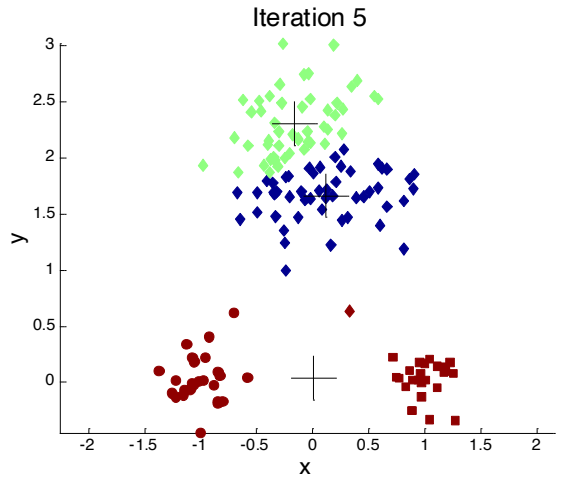
Importance of Choosing Initial Centroids …



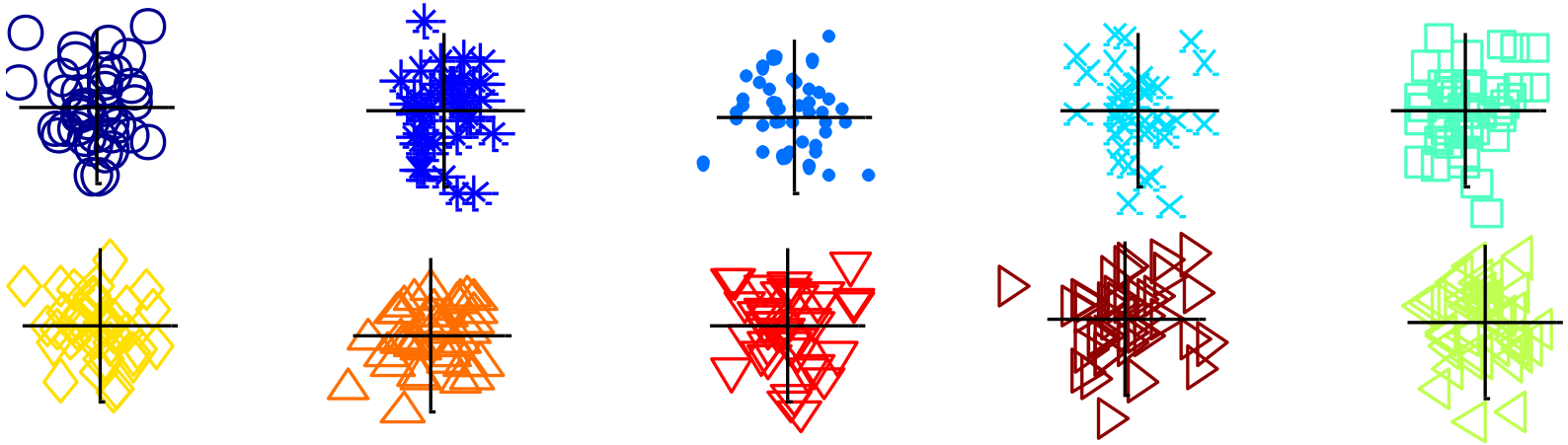






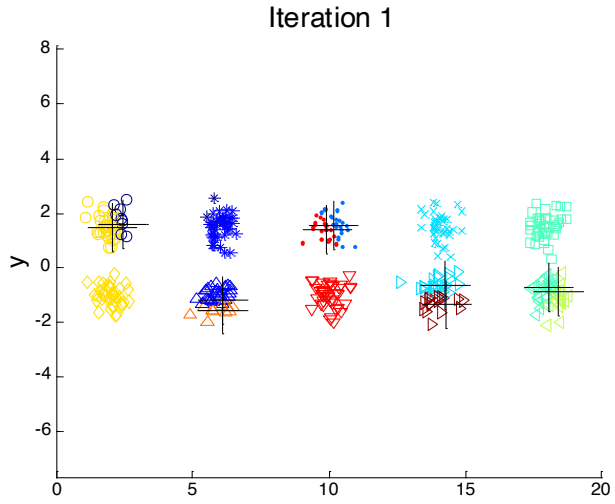


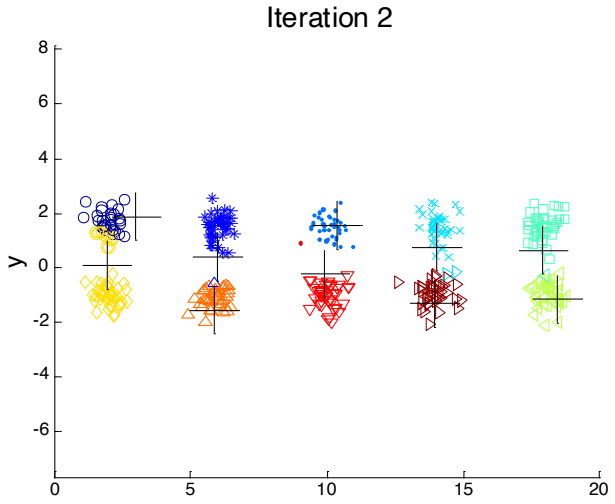
10 Clusters Example

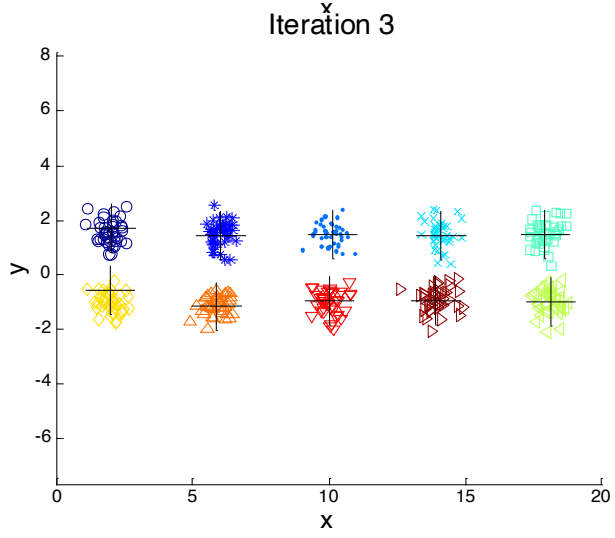


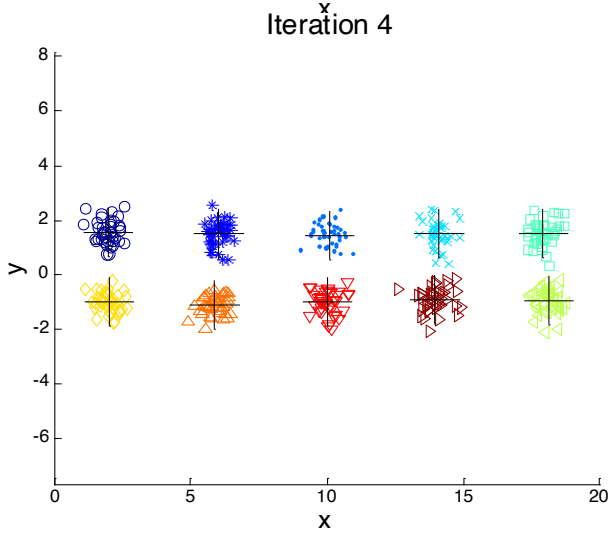
Starting with two initial centroids in one cluster of each pair of clusters

10 Clusters Example



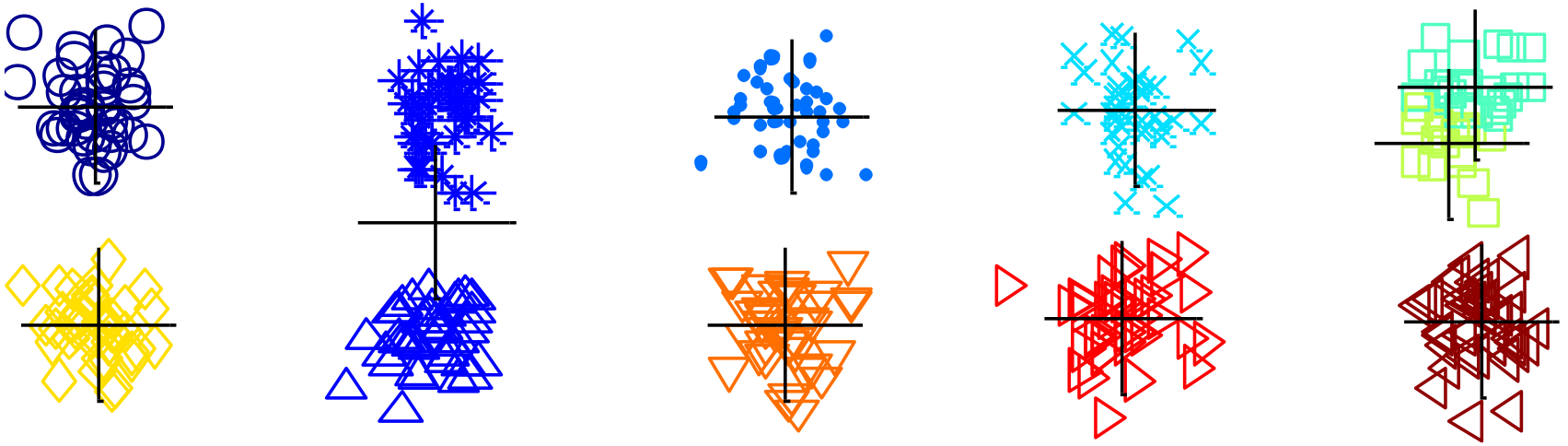






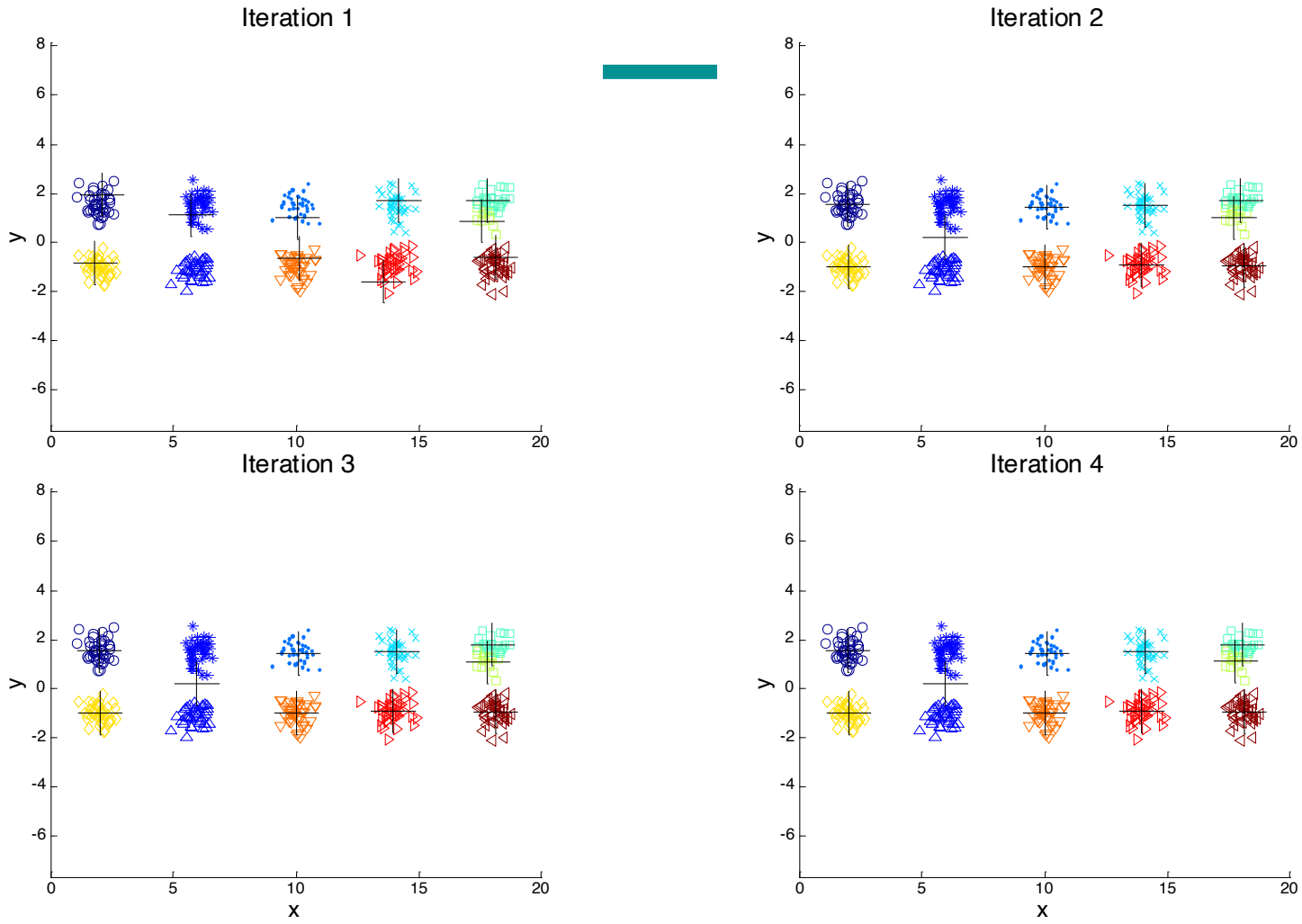
Starting with two initial centroids in one cluster of each pair of clusters

10 Clusters Example



Starting with some pairs of clusters having three initial centroids, while other have only one.

10 Clusters Example



Starting with some pairs of clusters having three initial centroids, while other have only one.

Solutions to Initial Centroids Problem

- Multiple runs
  - Helps, but probability is not on your side
- **Sample and use hierarchical clustering to determine initial centroids**
- Select more than k initial centroids and then select among these initial centroids
  - Select most widely separated
- Postprocessing
- Generate a larger number of clusters and then perform a hierarchical clustering
- Bisecting K-means
  - Not as susceptible to initialization issues

K-Means Extensions

Bisecting K-Means

Bisecting K-means

- Variant of K-Means that can produce a hierarchical clustering
- The number of clusters K must be specified.
- Start with a unique cluster containing all the points.

1: Initialize the list of clusters to contain the cluster containing all points.
2: **repeat**
3: Select the cluster with the highest SSE to the list of clusters
4: **for** $i = 1$ to *number_of_iterations* **do**
5: Bisect the selected cluster using basic 2-Means
6: **end for**
7: Add the two clusters from the bisection to the list of clusters.
8: **until** Until the list of clusters contains K clusters

Bisecting K-means Limitations

- The algorithm can be also exhaustive and terminating at a singleton clusters if K is not specified.
- Terminating at singleton clusters
  - Is time consuming
  - Singleton clusters are meaningless (i.e., over-splitting)
  - Intermediate clusters are more likely to correspond to real classes

- Bisecting K-Means do not use any criterion for stopping bisections before singleton clusters are reached.