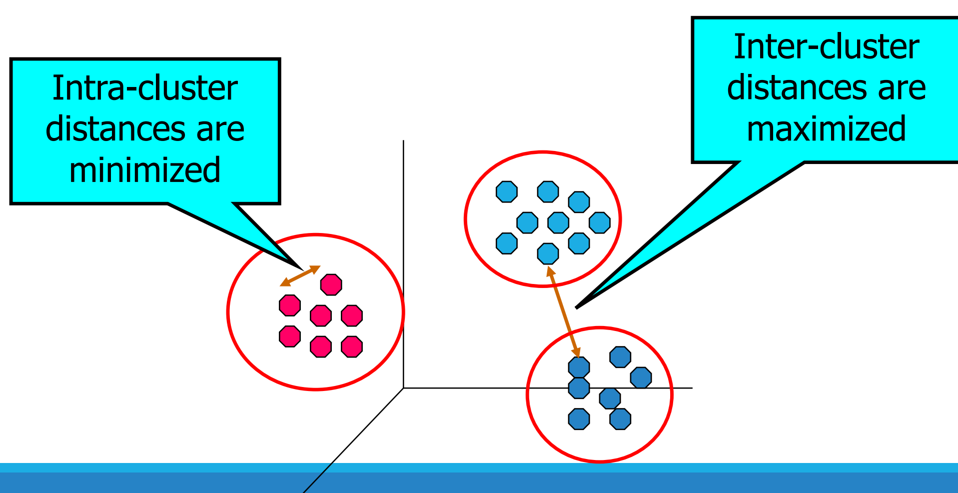
Beginner-friendly ML school:

Clustering and K-Means

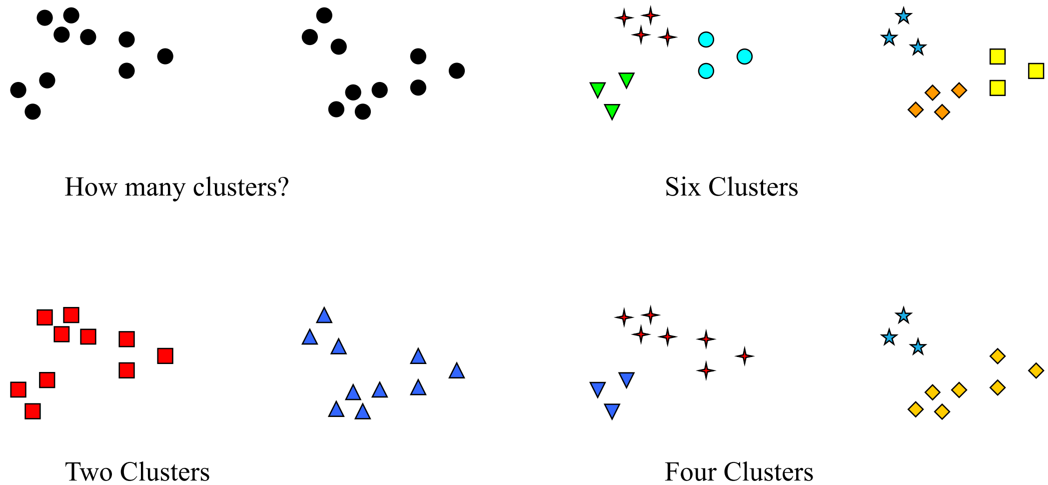
Viviana Acquaviva (CUNY)
w many many thanks to
Ashwin Satyanarayana who has taught clustering in my class many times and put together most of these slides!

What is Clustering?

In general a grouping of objects such that the objects in a group (cluster) are similar (or related) to one another and different from (or unrelated to) the objects in other groups



Notion of a Cluster can be Ambiguous



Types of Clustering

A clustering is a set of clusters

Important distinction between hierarchical and partitional sets of clusters

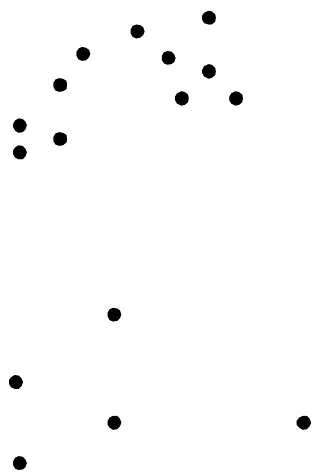
Partitional Clustering

- A division data objects into subsets (clusters) such that each data object is in exactly one subset

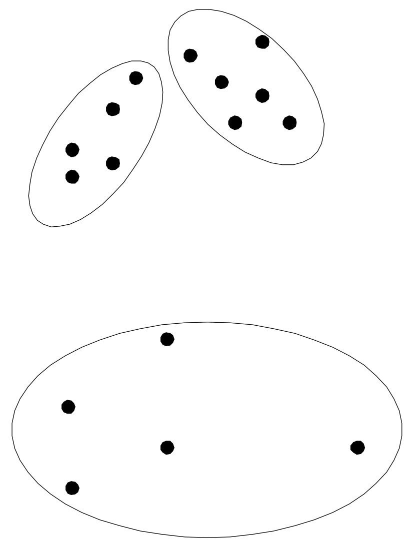
Hierarchical clustering

- A set of nested clusters organized as a hierarchical tree

Partitional Clustering

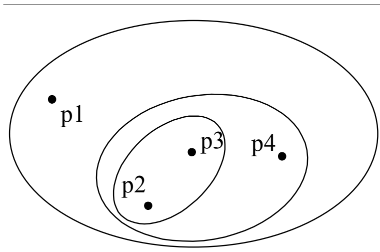


Original Points

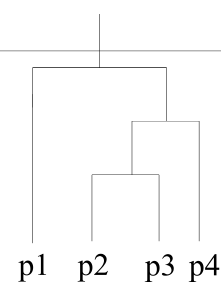


A Partitional Clustering

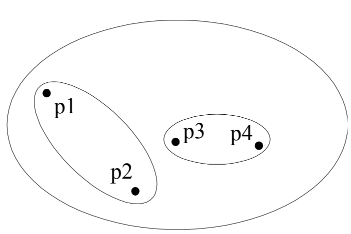
Hierarchical Clustering



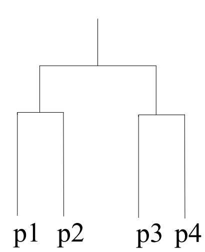
Traditional Hierarchical Clustering



Traditional Dendrogram



Non-traditional Hierarchical Clustering



Non-traditional Dendrogram

Objective Function

Clustering as an optimization problem

- Finds clusters that minimize or maximize an objective function.
- Enumerate all possible ways of dividing the points into clusters and evaluate the `goodness' of each potential set of clusters by using the given objective function. (NP Hard)
- Can have global or local objectives.
  - Hierarchical clustering algorithms typically have local objectives
  - Partitional algorithms typically have global objectives
- A variation of the global objective function approach is to fit the data to a parameterized model.
  - The parameters for the model are determined from the data, and they determine the clustering
  - E.g., Mixture models assume that the data is a 'mixture' of a number of statistical distributions.

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 objective is to minimize the sum of distances of the points to their respective centroid

K-means Clustering

Problem: Given a set X of n points in a d-dimensional space and an integer K, group the points into K clusters $C = \{C_1, C_2, \ldots, C_k\}$ such that

$$Cost(C) = \sum_{i=1}^{k} \sum_{x \in C_i} dist(x, c)$$

is minimized, where c_i is the centroid of the points in cluster C_i

K-means Clustering

- Most common definition is with euclidean distance, minimizing the Sum of Squares Error (SSE) function
  - Sometimes K-means is defined like that

Problem: Given a set X of n points in a d-dimensional space and an integer K group the points into K clusters $C = \{C_1, C_2, \ldots, C_k\}$ such that

$$Cost(C) = \sum_{i=1}^{k} \sum_{x \in C_i} (x - c_i)^2$$

is minimized, where c_i is the mean of the points in cluster C_i

K-means Algorithm

Pseudocode:

```
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
```

Problem - Initialization

Initial centroids are often chosen randomly, and clusters produced vary from one run to another.

Do multiple runs and select the clustering with the smallest error

Select original set of points by methods other than random . E.g., pick the most distant (from each other) points as cluster centers (K-means++ algorithm)

K-means Algorithm – Convergence

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 = dimensionality

In general a fast and efficient algorithm

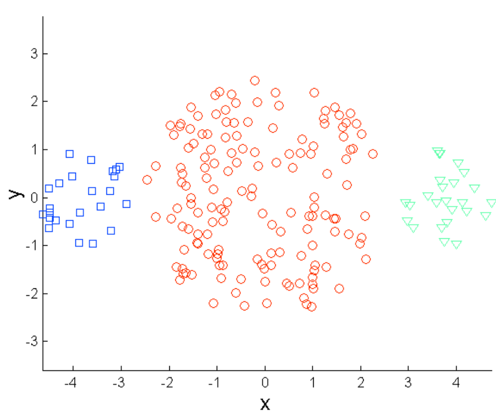
Limitations of K-means

You need to specify the number of clusters in advance (although there are ways to find the optimal one.

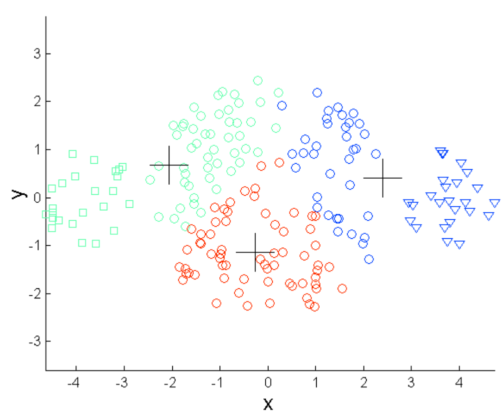
K-means has problems when clusters are of different
- Sizes
- Densities
- Non-globular shapes

K-means has problems when the data contains outliers.

Limitations of K-means: Differing Sizes

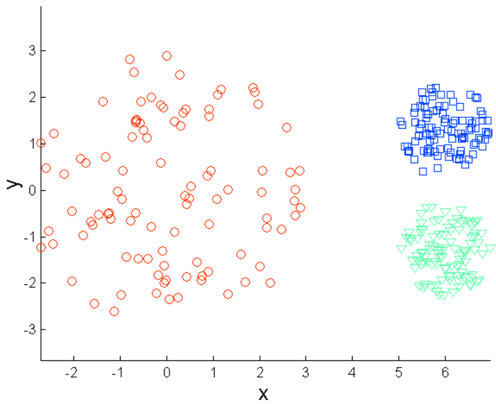


Original Points

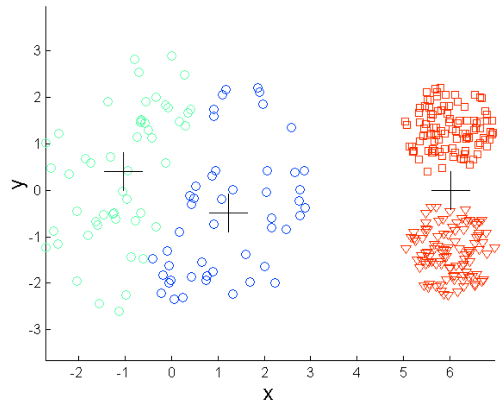


K-means (3 Clusters)

Limitations of K-means: Differing Density

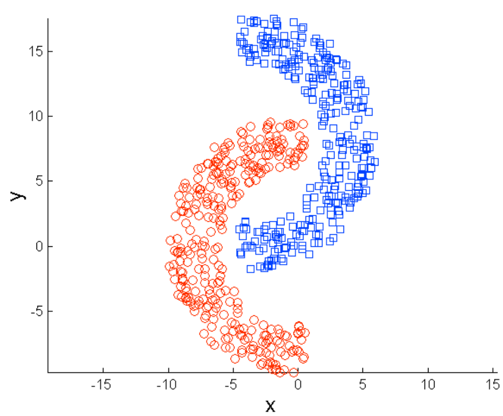


Original Points

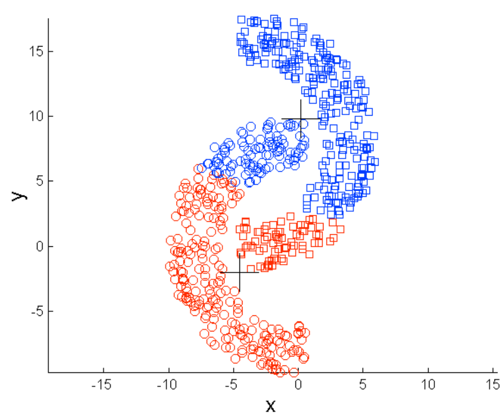


K-means (3 Clusters)

Limitations of K-means: Non-globular Shapes

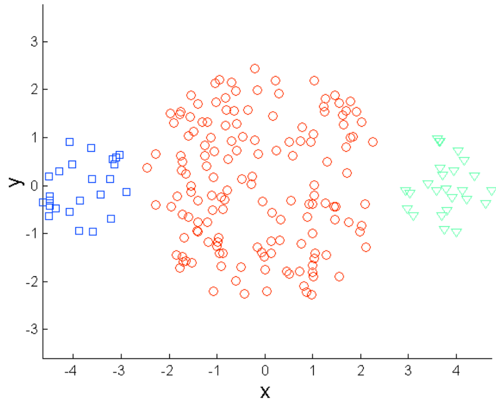


Original Points

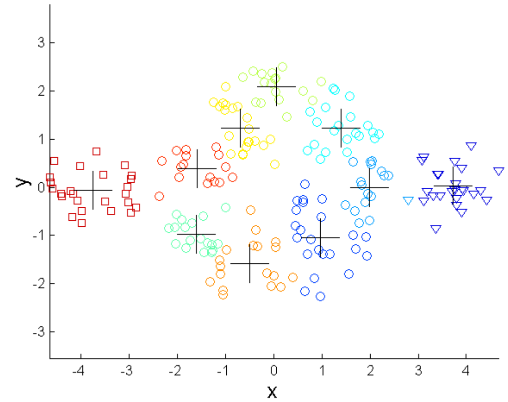


K-means (2 Clusters)

Overcoming K-means Limitations



Original Points

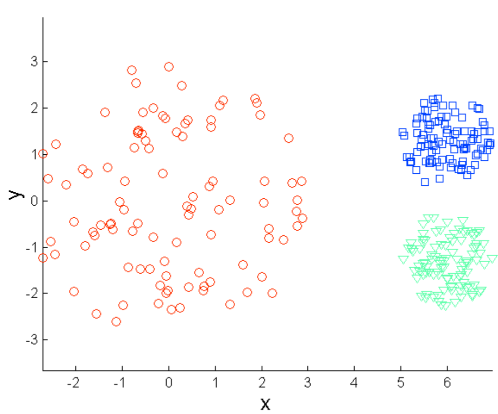


K-means Clusters

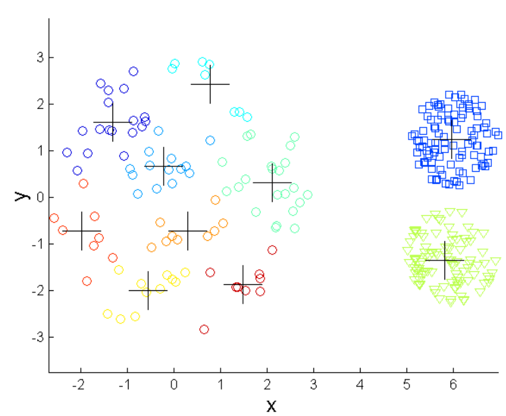
One solution is to use many clusters.

Find parts of clusters, but need to put together.

Overcoming K-means Limitations

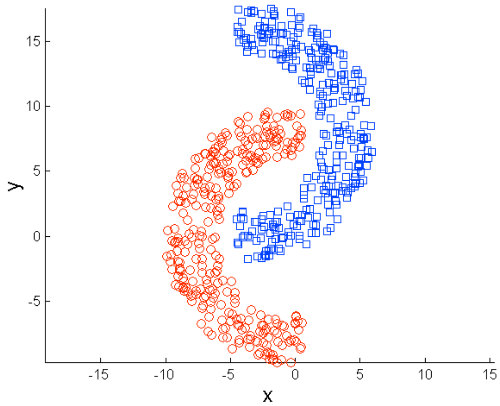


Original Points

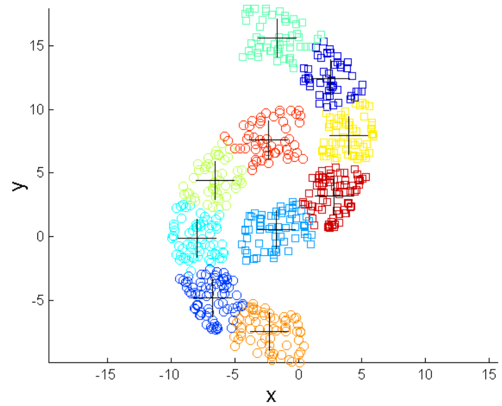


K-means Clusters

Overcoming K-means Limitations



Original Points



K-means Clusters

Let’s play:

https://www.naftaliharris.com/blog/visualizing-k-means-clustering/