BITS Pilani

BITS Pilani
Hyderabad Campus

Dr. Manik Gupta
Assistant Professor
Department of CSIS

Data Mining (CS F415)
Lecture 16 – Clustering

Tuesday, 25th February 2020

Research Project Mid-Semester Demos

Please fill in your preferences for the Demo next week:

https://doodle.com/poll/i6ebv5nfxasg3emf

Few important things (PLEASE FOLLOW THEM!):

- Please be on time else your assessment will not be carried out. All the team members need to be present.
- Please do not contact us regarding slot swaps. Coordinate amongst yourself and inform us via email.
- Please carry your own laptops and make sure the demo runs as well as report/slides are present and working.
- Create 5 slides (time limit 5 min!!)
  - Team work division
  - Project and data overview
  - Reasons for choosing the techniques used
  - Shortcomings
  - Next steps

Today's Agenda

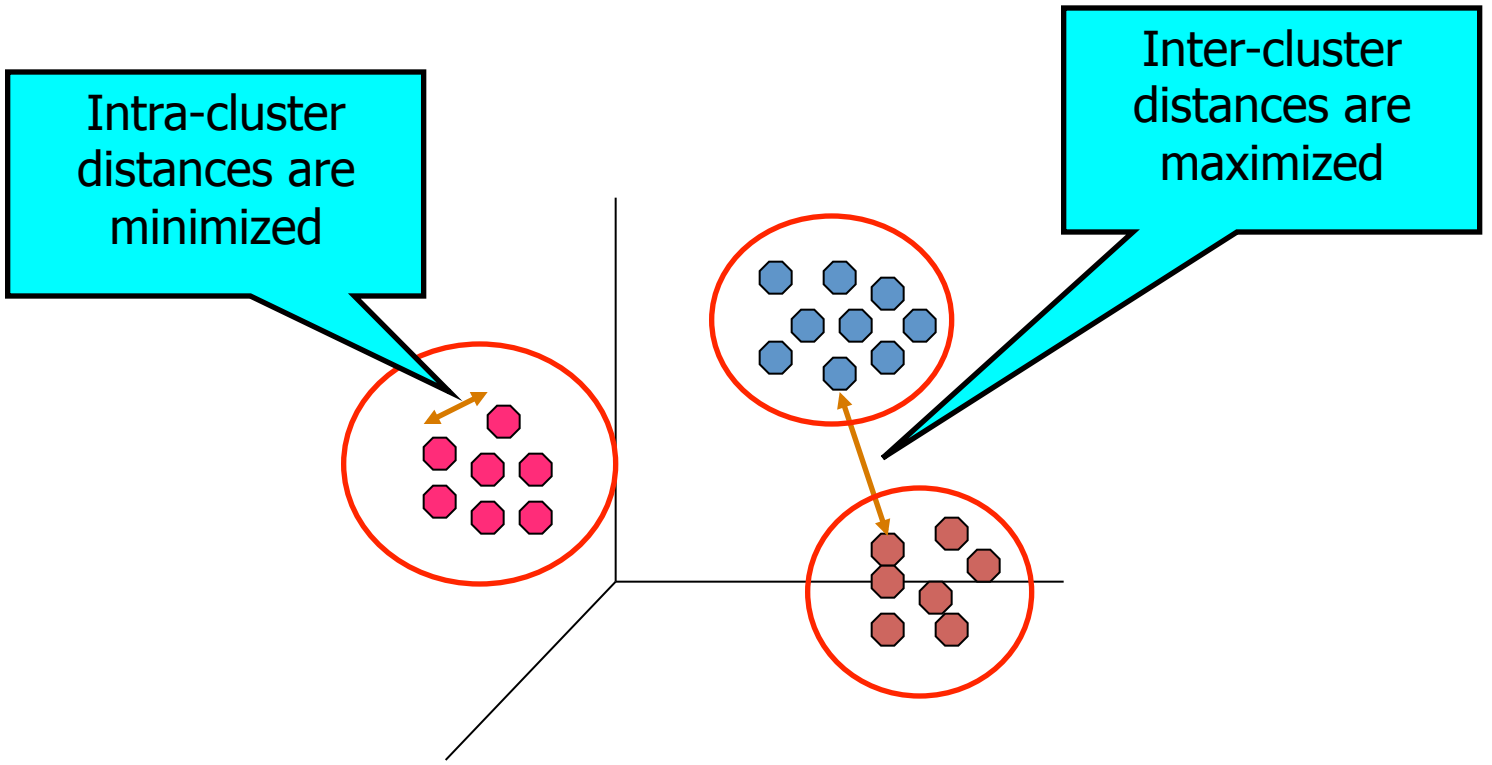
- Introduction to Clustering

What is a cluster?

- Cluster: A collection of data objects
  - Similar (or related) to one another within the same group
  - Dissimilar (or unrelated) to the objects in other groups

What is Cluster Analysis?

- Finding groups of objects such that the objects in a group will be similar (or related) to one another and different from (or unrelated to) the objects in other groups



What are clustering applications?

- Typical applications
  - As a stand-alone tool to get insight into data distribution
  - As a preprocessing step for other algorithms

Clustering for Data Understanding and Applications

- Biology: taxonomy of living things: kingdom, phylum, class, order, family, genus and species
- Information retrieval: document clustering
- Land use: Identification of areas of similar land use in an earth observation database
- Marketing: Help marketers discover distinct groups in their customer bases, and then use this knowledge to develop targeted marketing programs
- City-planning: Identifying groups of houses according to their house type, value, and geographical location
- Earth-quake studies: Observed earth quake epicenters should be clustered along continent faults
- Climate: understanding earth climate, find patterns of atmospheric and ocean
- Economic Science: market research

Clustering as a Preprocessing Tool (Utility)

- Summarization
  - Preprocessing for regression, PCA, classification, and association analysis
  - Apply to reduced set consisting of cluster prototypes
- Compression
  - Image processing: vector quantization
  - Each object represented by index of prototype associated with a cluster
- Finding Nearest Neighbors
  - Localizing search to one or a small number of clusters
  - Use prototypes to reduce number of distance computations that are necessary to find NN of an object.
- Outlier detection
  - Outliers are often viewed as those “far away” from any cluster

What is not Cluster Analysis?

- Simple segmentation
  - Dividing students into different registration groups alphabetically, by last name
- Results of a query
  - Groupings are a result of an external specification
  - Clustering is a grouping of objects based on the data
- Supervised classification
  - Have class label information

Few Announcements

- Team Demonstrations 27th Feb (12 pm - 2pm)
  TeamID_33
  TeamID_27
  TeamID_12
  TeamID_09

- Team Demonstrations 28th Feb (1pm onwards)

  TeamID_25
  TeamID_37
  TeamID_23
  Team 11
  TeamID_02
  TeamID_29
  Team_07
  TeamID_06

- Rest of the demonstrations after mid semester on 11th and 13th March
- No Lecture on Feb 29th
- Mid semester exam on March 6th, 9am to 10:30 am

Notion of a Cluster can be Ambiguous

How many clusters?

Six Clusters

Two Clusters

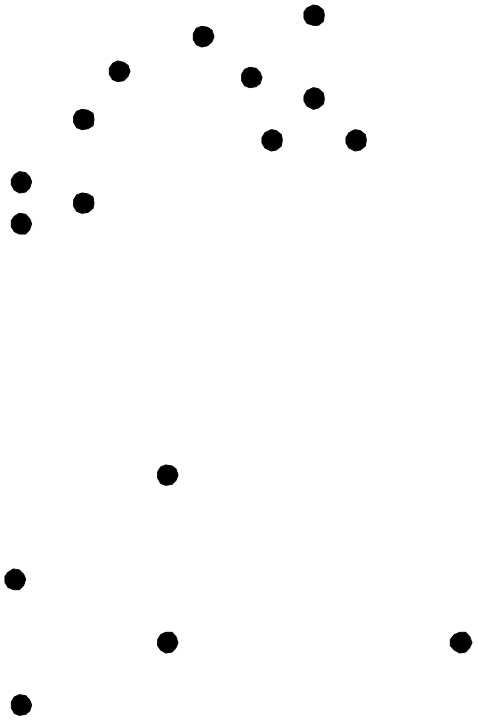
Four Clusters

Definition of a cluster is imprecise and best definition depends on the nature of data and desired results.

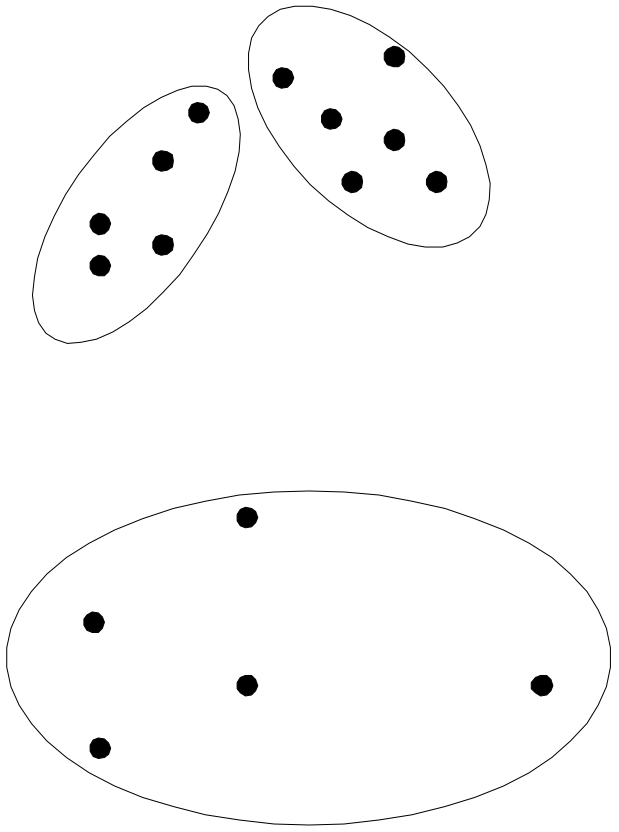
Types of Clustering

- Important distinction between hierarchical and partitional sets of clusters
  - Partitional Clustering
    - A division data objects into non-overlapping 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

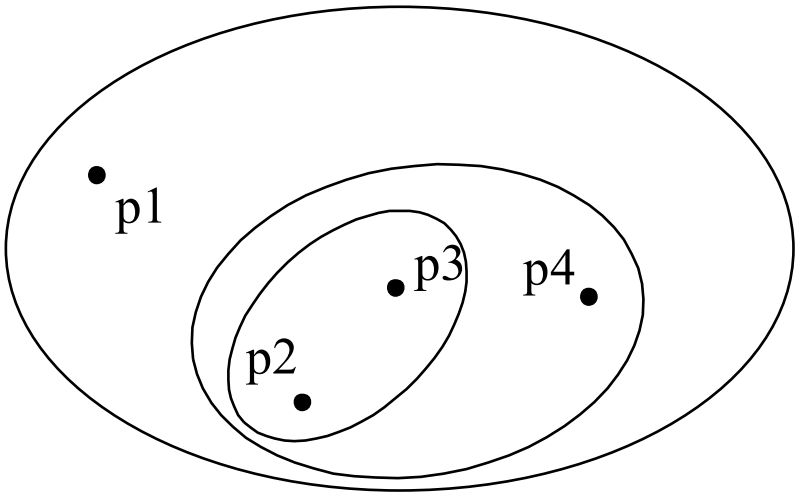


Original Points

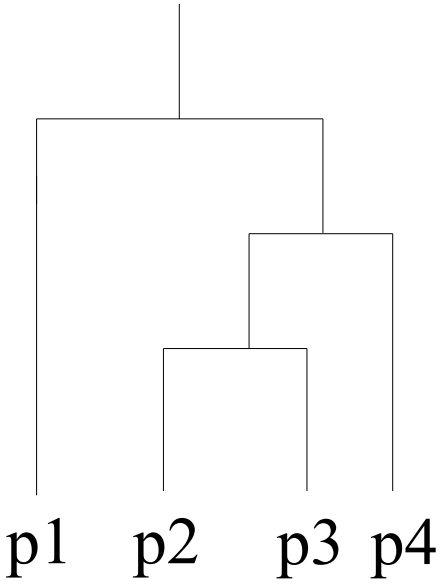


Partitional Clustering

Hierarchical Clustering



Hierarchical Clustering



Dendrogram

Other Distinctions Between Sets of Clusters

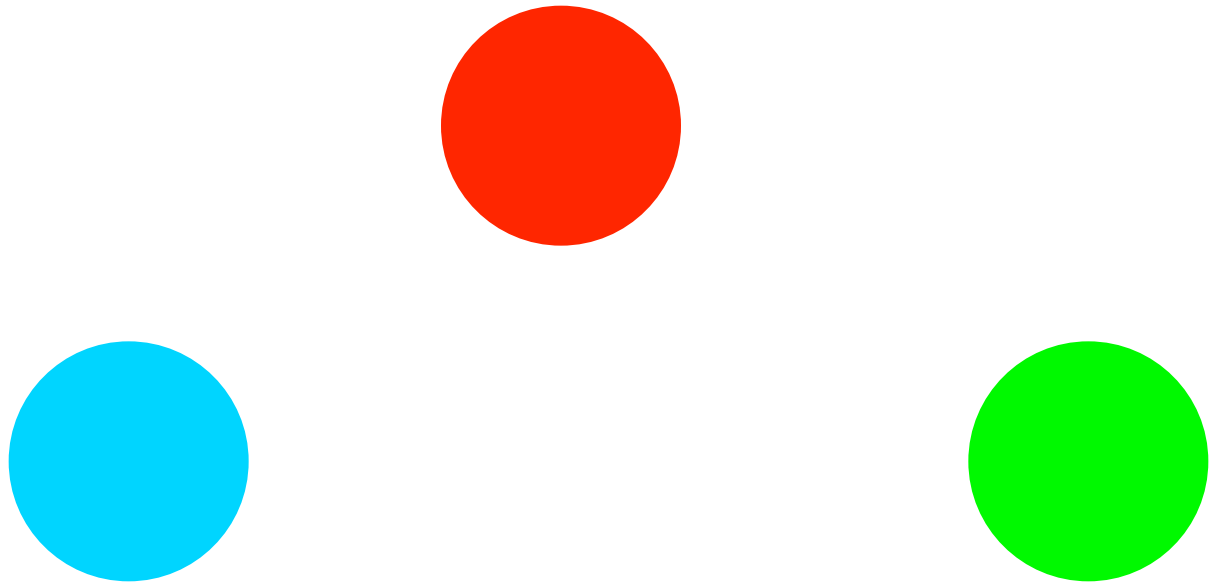
- **Exclusive versus non-exclusive**
  - In non-exclusive clustering, points may belong to multiple clusters.
  - Can represent multiple classes or 'border' points
- **Fuzzy versus non-fuzzy**
  - In fuzzy clustering, a point belongs to every cluster with some weight between 0 and 1
  - Weights for each object must sum to 1
  - Probabilistic clustering has similar characteristics
- **Partial versus complete**
  - In some cases, we only want to cluster some of the data
- **Heterogeneous versus homogeneous**
  - Cluster of widely different sizes, shapes, and densities

Types of Clusters

- Well-separated clusters
- Center-based clusters
- Contiguous clusters
- Density-based clusters
- Property or Conceptual

Types of Clusters: Well-Separated

- Well-Separated Clusters
  - A cluster is a set of points such that any point in a cluster is closer (or more similar) to every other point in the cluster than to any point not in the cluster.



3 well-separated clusters

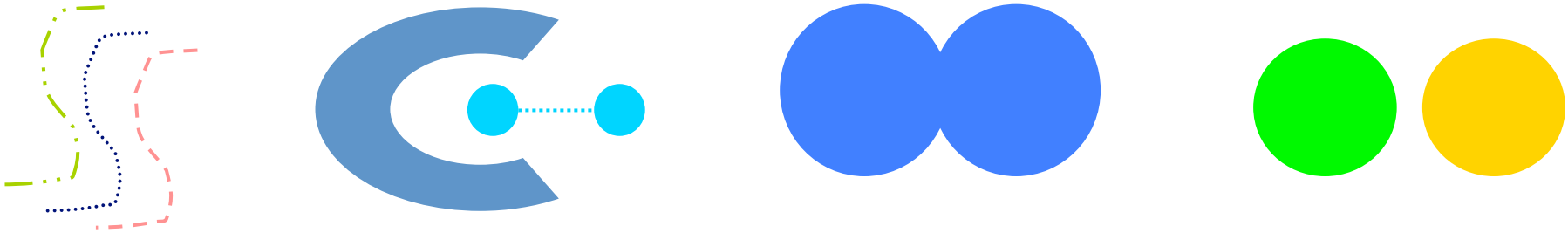
Types of Clusters: Center-Based

- Center-based or Prototype based
  - A cluster is a set of objects such that an object in a cluster is closer (more similar) to the "center" of a cluster, than to the center of any other cluster
  - The center of a cluster is often a centroid, the average of all the points in the cluster, or a medoid, the most "representative" point of a cluster

4 center-based clusters

Types of Clusters: Contiguity-Based

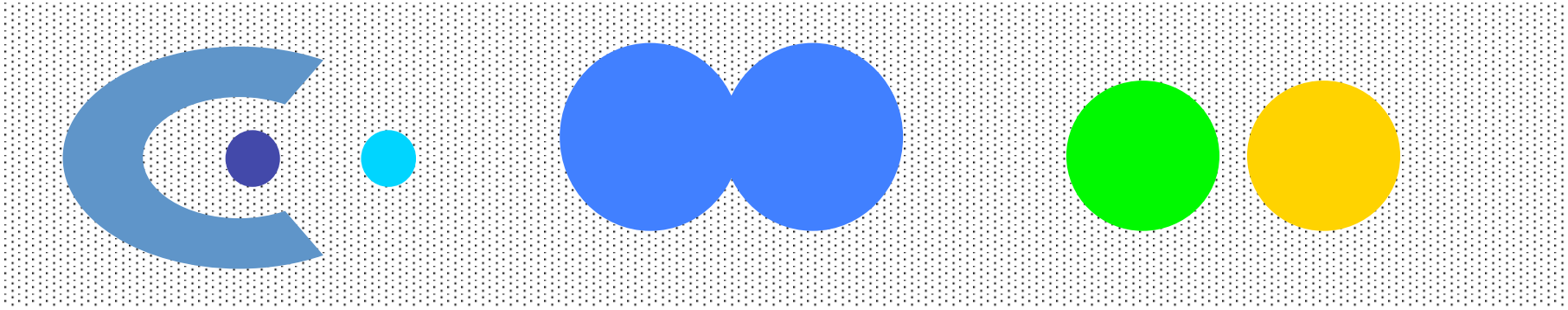
- Contiguous Cluster
  - A cluster is a set of points such that each point is closer (or more similar) to some other point in the cluster than to any point in another cluster.
  - Group of objects that are connected to one another, but have no connection to objects outside the group



8 contiguous clusters

Types of Clusters: Density-Based

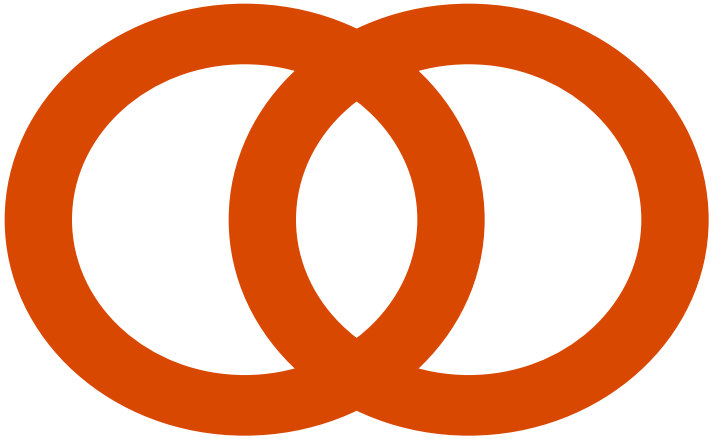
- Density-based
  - A cluster is a dense region of points, which is separated by low-density regions, from other regions of high density.
  - Used when the clusters are irregular or intertwined, and when noise and outliers are present.



6 density-based clusters

Types of Clusters: Conceptual Clusters

- Shared Property or Conceptual Clusters
  - Finds clusters that share some common property or represent a particular concept.
  - Specific concept of a cluster is needed to detect these clusters



2 Overlapping Circles

Quality: What Is Good Clustering?

- A good clustering method will produce high quality clusters
  - high intra-class similarity: cohesive within clusters
  - low inter-class similarity: distinctive between clusters
- The quality of a clustering method depends on
  - the similarity measure used by the method
  - its implementation, and
  - Its ability to discover some or all of the hidden patterns

K-Means Clustering

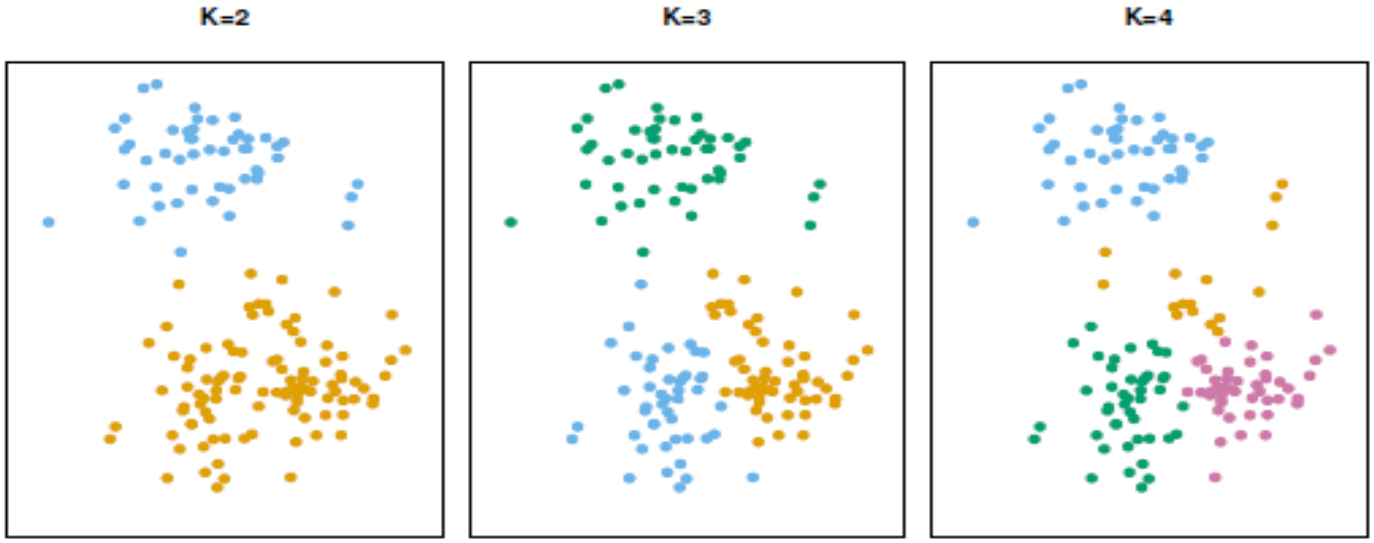
- K-Means algorithm belongs to the category of prototype-based clustering.
- **Prototype-based** clustering means that each cluster is represented by a prototype
  - **Centroid** (*average*) of similar points with continuous features
  - **Medoid** (the most *representative* or most frequently occurring point) in the case of categorical features.

K-means Clustering

- Partitional 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
- 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

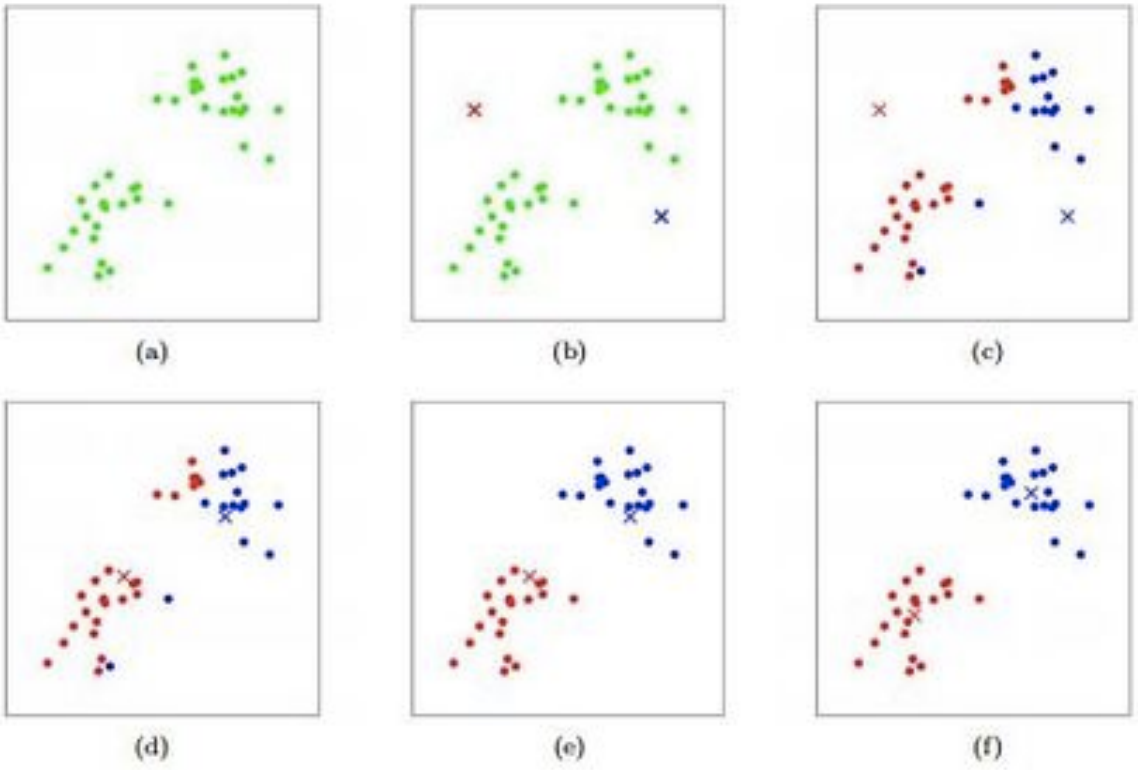


In **K-means** clustering, the observations are partitioned into a pre-specified number of clusters.

Steps for K-Means Algorithm

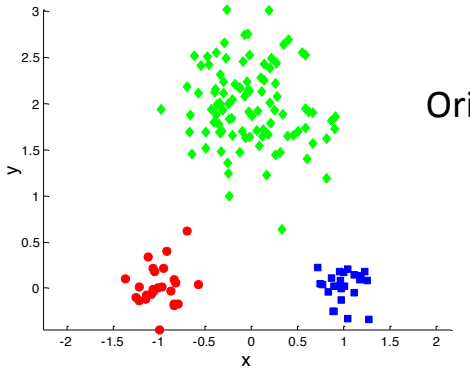
- Randomly pick *k* centroids from the sample points as initial cluster centers.
- Assign each sample to the nearest centroid $\mu(j)$, $j \in \{1,...,k\}$.
- Move the centroids to the center of the samples that were assigned to it.
- Repeat the steps 2 and 3 until the cluster assignment do not change or a user defined tolerance or a maximum number of iterations is reached.

K-means demo with K=2

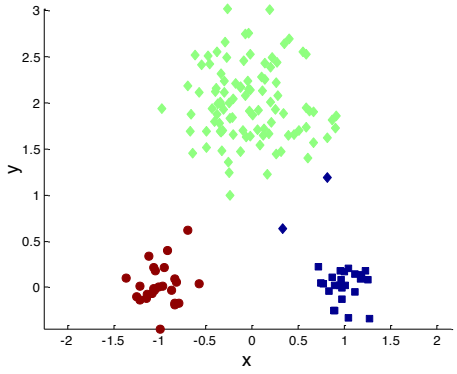


Choosing Initial Centroids

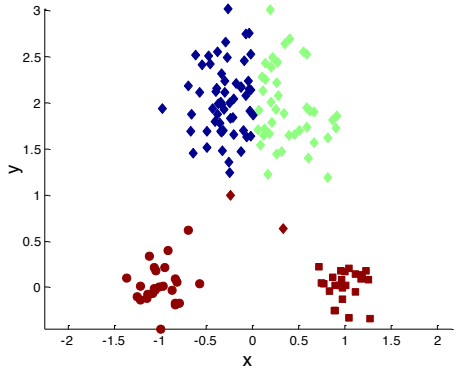
- Initial centroids are often chosen randomly.
  - Clusters produced vary from one run to another.



Original Points

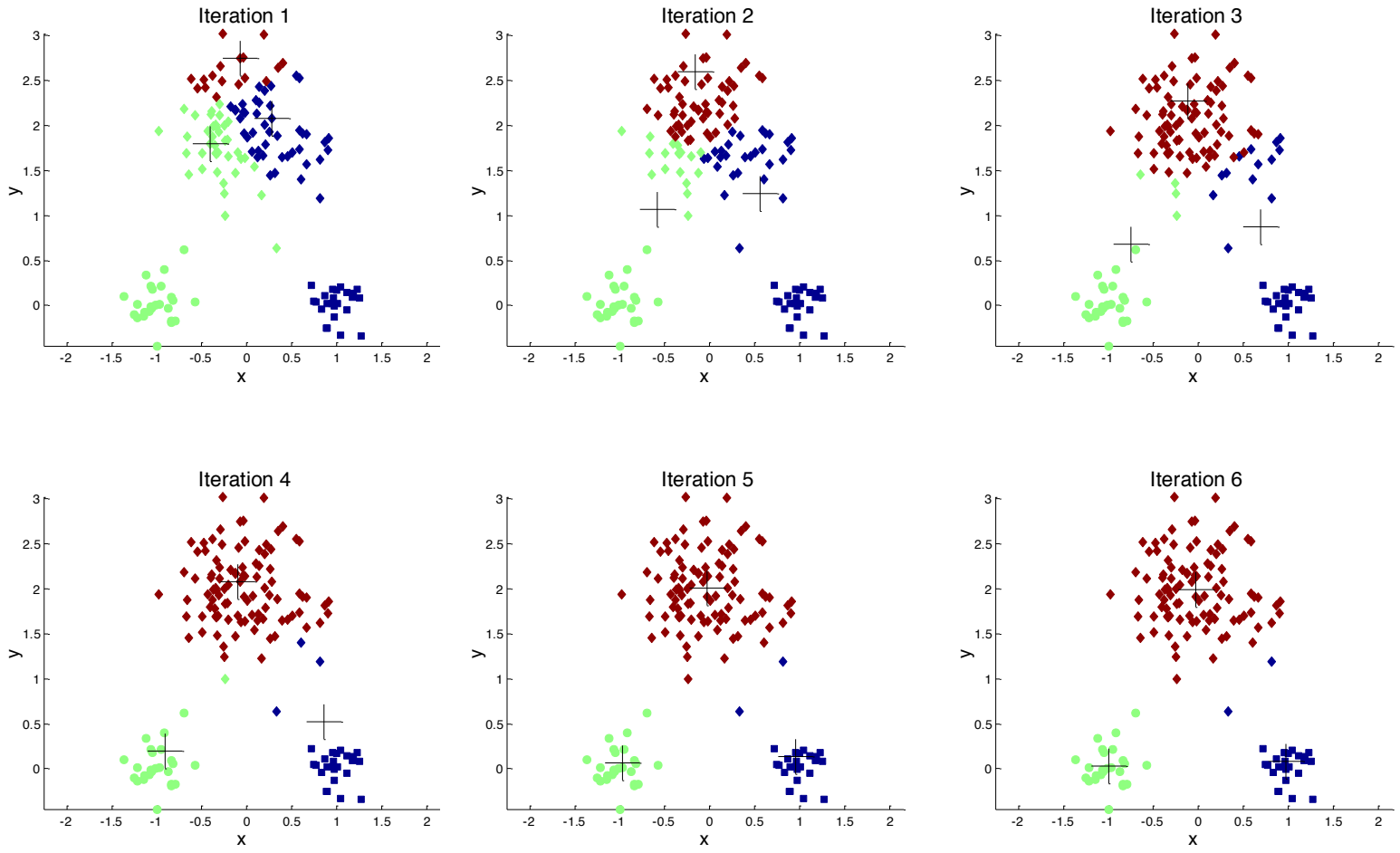


Clustering 1

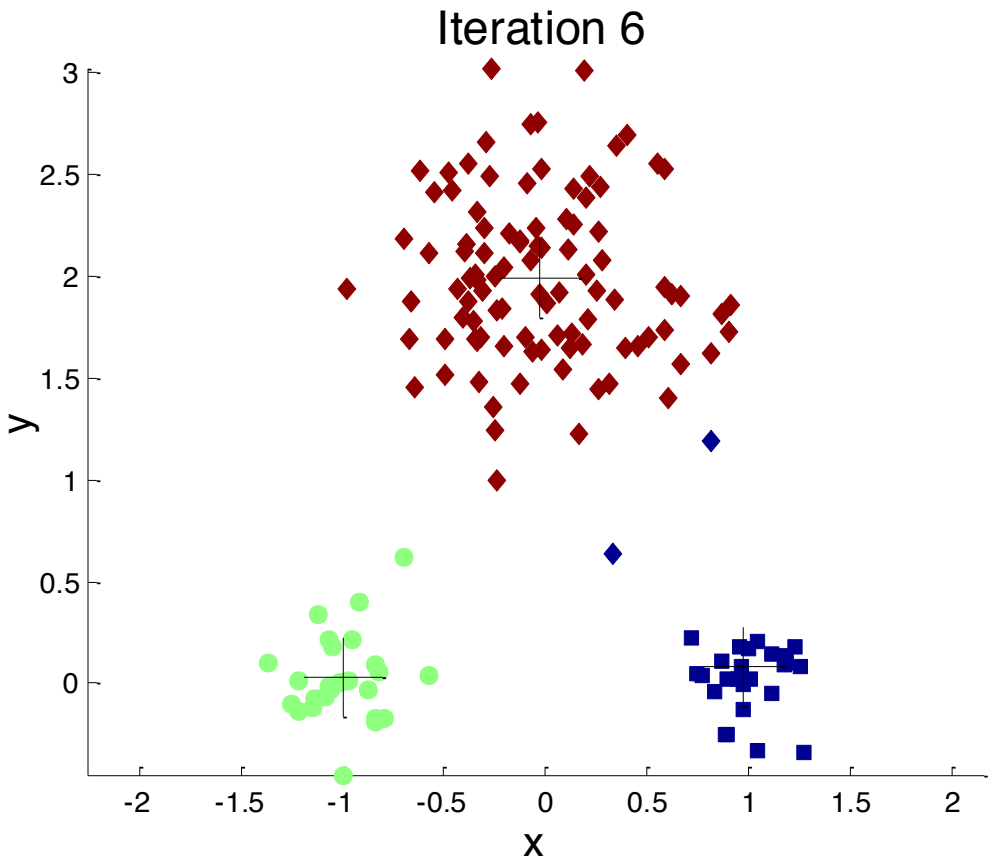


Clustering 2

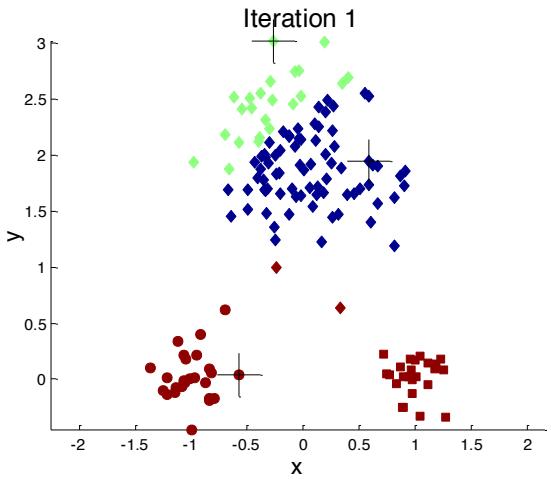
Good Initialization

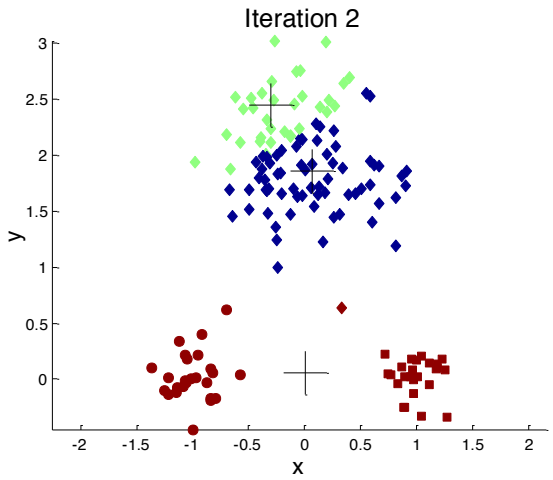


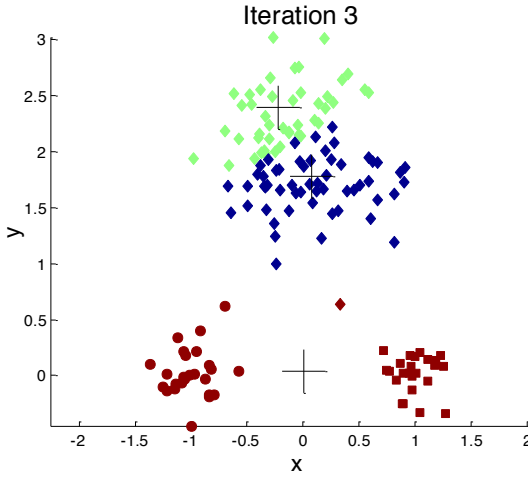
Good Result

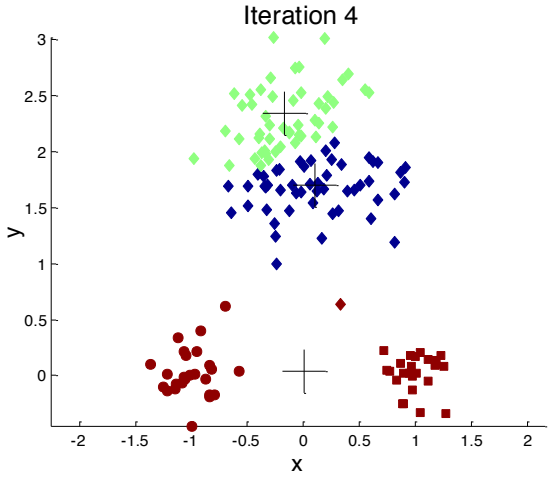


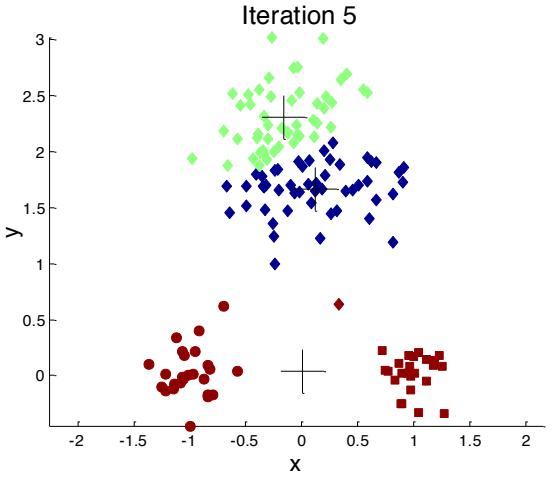
Bad Initialization



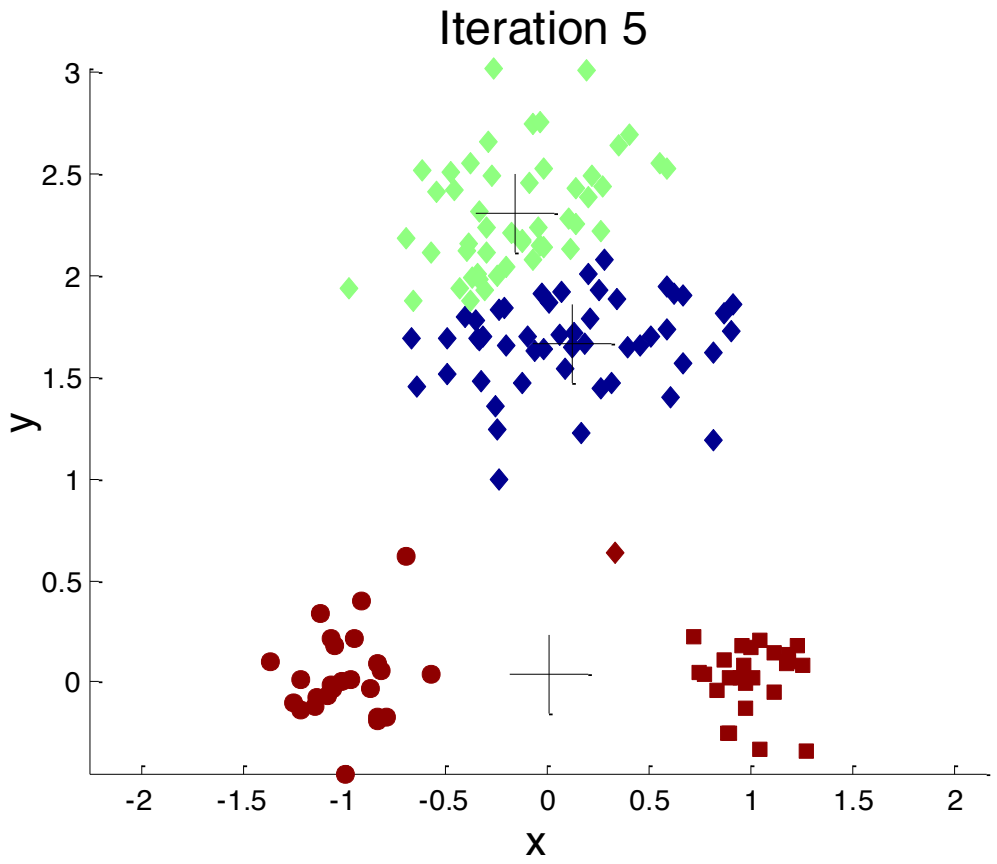








Bad Result



How to measure clustering result?

- 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.

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

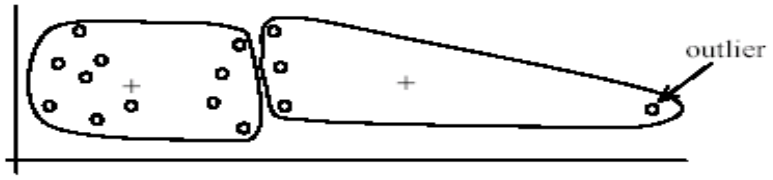
  - x is a data point in cluster C_i and m_i is the representative point for cluster C_i

Advantages of K-Means

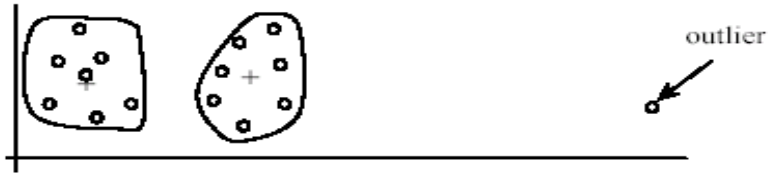
- Easy to understand and implement
- Most popular clustering algorithm

Drawbacks of K-Means

- Need to specify the number of clusters *K* a priori
- K-means is sensitive to outliers
  - Outliers are data points that are very far away from other data points.
  - Outliers could be errors in the data recording or some special data points with very different values.



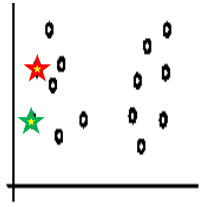
(A): Undesirable clusters



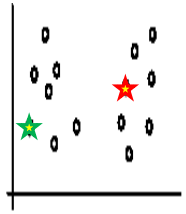
(B): Ideal clusters

Drawbacks of K-Means

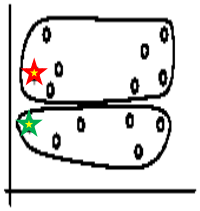
Sensitivity to initial seeds



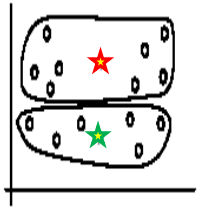
Random selection of seeds (centroids)



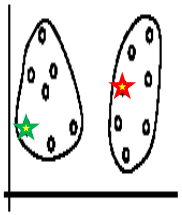
Random selection of seeds (centroids)



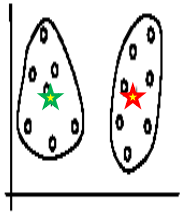
Iteration 1



Iteration 2

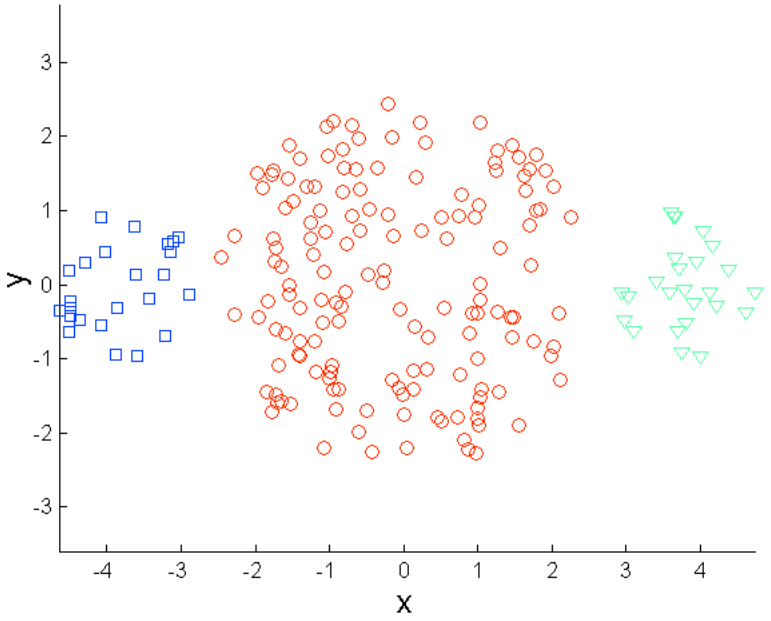


Iteration 1

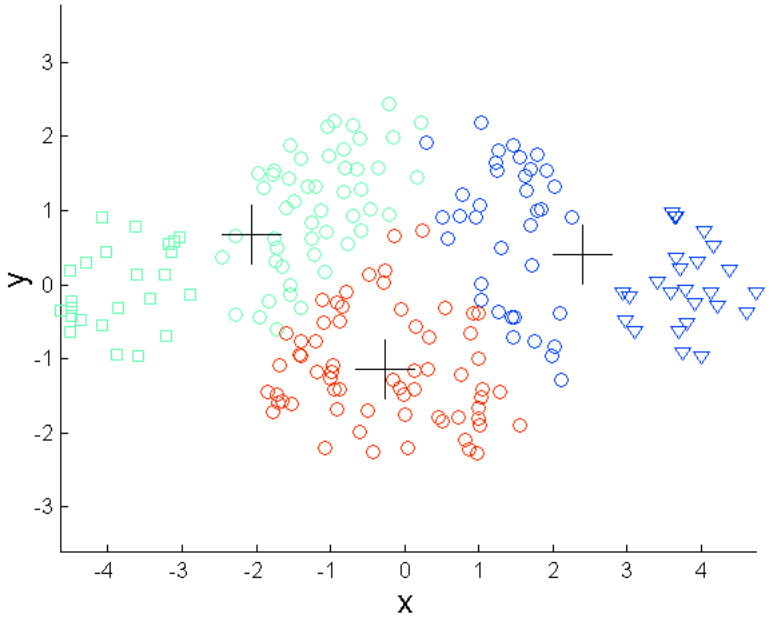


Iteration 2

Drawbacks of K-Means Differing Sizes

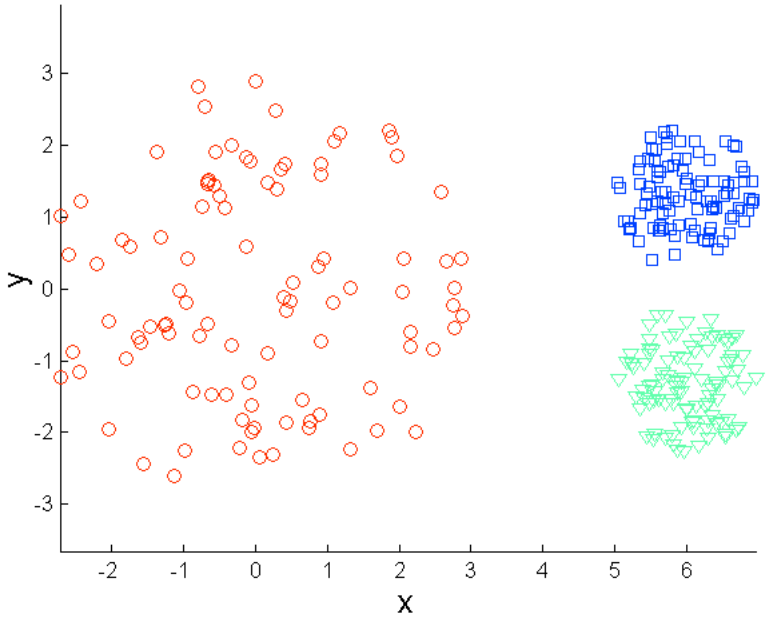


Original Points

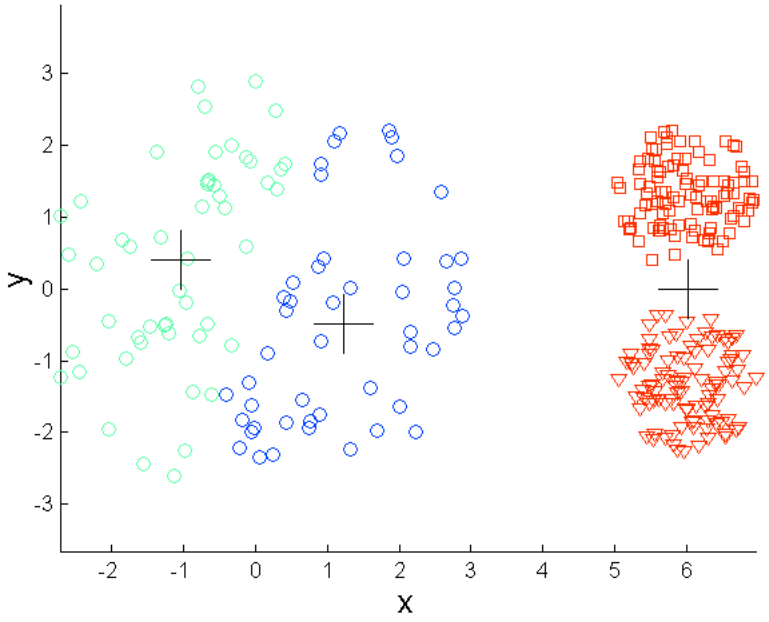


K-means (3 Clusters)

Drawbacks of K-means: Differing Density

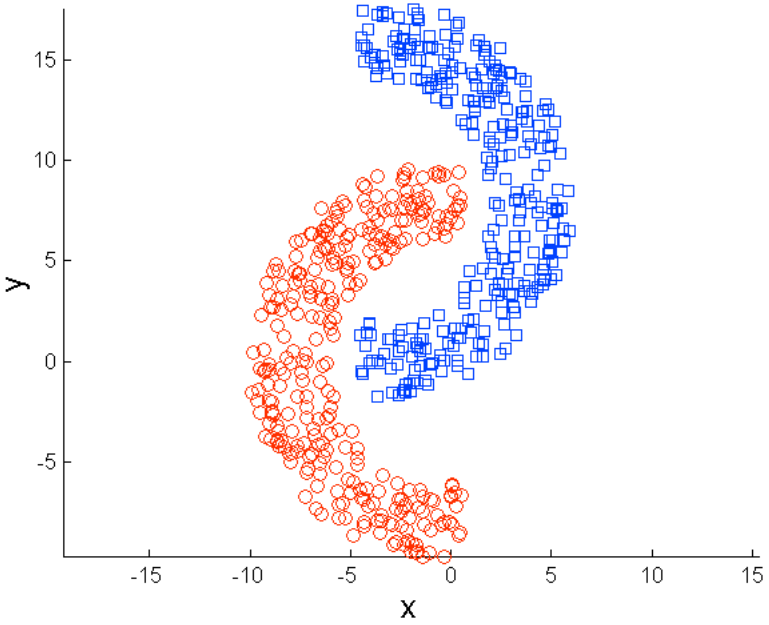


Original Points

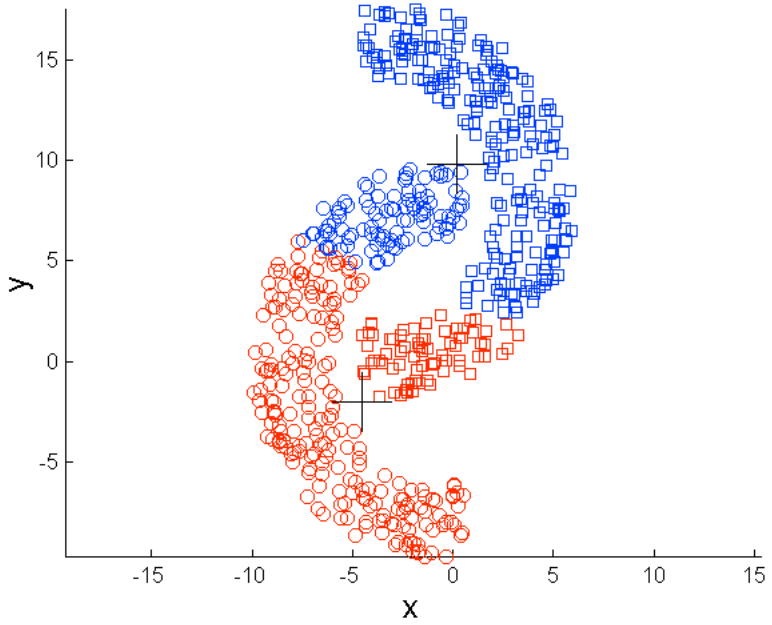


K-means (3 Clusters)

Drawbacks of K-means: Non-globular Shapes



Original Points



K-means (2 Clusters)

Thanks!

Next Lecture:

- Variants of K-Means Clustering
- Hierarchical Clustering
- DBSCAN

Readings:

– Chapter 8 - Tan & Kumar