UNSUPERVISED LEARNING

CLUSTER ANALYSIS

CS576 MACHINE LEARNING

Dr. Jin S. Yoo, Professor
Department of Computer Science
Purdue University Fort Wayne

Reference

- P. Tan et al., Introduction to Data Mining, 2nd ed., Ch 2.4 Similarity Measures, Ch7. Cluster Analysis (Ch 8 in 1st ed.)
- Kelleher et al., Fundamentals of Machine Learning (2nd ed.), Ch 10. Beyond Prediction: Unsupervised Learning
- G. James et al., An Introduction to Statistical Learning, Ch 10.3

Outline

- Supervised vs. Unsupervised Learning
- Cluster Analysis
  - Introduction
  - Similarity/Dissimilarity Metrics
  - Clustering Methods
    - *K*-mean Clustering
    - Hierarchical Clustering
  - Evaluation of Clustering

Supervised Learning: Regression and Classification

- **Supervised learning** is often described as "***learning with a teacher***."
- In this framework, each instance is characterized by a feature set $(X_1, X_2, \ldots, X_p)$ and is associated with a response variable Y.
- The **objective** is to construct a model that can predict the response variable Y using the features $(X_1, X_2, \ldots, X_p)$.
- Model training is performed using a dataset of labeled examples $\{(x_{1i}, x_{2i}, \ldots, x_{pi}, y_i)\}_{i=1}^{N}$, where each y_i is the known outcome for the corresponding feature set.
- **Examples**: Regression and Classification

Unsupervised Learning

- **Unsupervised learning** can be thought of as "***learning without a teacher***."
- Here, we only have access to the features $(X_1, X_2, \ldots, X_p)$ without any associated response variables.
- The focus is on identifying patterns, groupings, or inherent structures within the data rather than making predictions.
- **Examples** of unsupervised learning tasks include clustering, dimensionality reduction, density estimation, association rule mining, outlier detection, and so on.

Objectives of Unsupervised Learning

- The essence of unsupervised learning lies in uncovering structures and insights from data without labeled outcomes.
- Key questions include:
  - Can we discover subgroups or patterns within the variables or observations?
  - What are the most effective ways to represent or visualize inherent structures of the data?

- Consider a dataset comprising N observations $(x_1, \ldots, x_N)$ of a random d-vector X having a joint probability distribution $\Pr(X)$
- The aim of unsupervised learning is to deduce the characteristics of $\Pr(X)$ autonomously, without guidance or corrective feedback (i.e., without a "teacher" to verify the findings)

Unsupervised Learning Methods

- Unsupervised learning methods are numerous and diverse, each suited for different kinds of data and analysis goals.
  - Clustering
  - Dimensionality Reduction
  - Association Rue Mining
  - Anomaly Detection
  - Feature Extraction and Filtering
  - Neural Networks and Deep Leering
    - Self-Organizing Maps (SOMs)
  - Visualization and Mapping
    - Multidimensional Scaling (MDS)
    - Self-Organizing Maps (SOMs)
  - Manifold Learning
  - Matrix Factorization

Clustering

- **Clustering** is a suite of techniques aimed at partitioning a dataset into clusters.
- Each cluster groups data points that are more similar to each other than to those in different clusters.
- The objective of cluster analysis is to identify distinct regions within the feature space where the probability density $\Pr(X)$ is locally maximized.

Dimensionality Reduction

- **Dimensionality reduction** includes techniques such as Principal Component Analysis (PCA), Multidimensional Scaling (MDS), Self-Organizing Maps (SOMs), and Principal Curves, all of which aim to simplify the complexity of data by reducing the number of variables under consideration.
- Principal Component Analysis (PCA) vs. Clustering
  - PCA seeks a lower-dimensional representation that captures a significant portion of the variance in the data, aiding in the visualization and understanding of its structure.
  - In construct to PCA, clustering algorithms strive to discover homogeneous groups within the data, not necessarily reducing dimensionality but clarifying the grouping structure.

Association Rule Mining

- **Association Rule Mining** uncovers interesting relationships between variables within large sets of data, typically found in transactional databases.
- It aims to identify patterns or rules that explain the presence or absence of certain items within transactions – like what items often appear together in a shopping cart.
- These rules are typically presented in the form of "If-Then" statements, revealing the likelihood of item co-occurrence in binary-valued datasets, emphasizing frequent combinations.

Outlier Detection

- **Outlier (Anomaly) detection** focuses on discovering data points that deviate noticeably from the majority of the data.
- These anomalies or outliers could signify errors, but they may also indicate a novel or significant discovery, warranting further investigation.
- Effective detection helps in areas such as fraud detection, system health monitoring, and removing noise from data.

Usefulness of Unsupervised Learning

- Unsupervised learning methods **use data without pre-given labels.**
  - Getting unlabeled data, like readings from scientific equipment or user activity logs, is usually less work than labeling data by hand.
- Unsupervised learning is becoming increasingly **useful in various areas**. For example:
  - Sorting customers based on what they look at and buy online,
  - Organizing films based on how viewers rate them,
  - Categorizing breast cancer patients by patterns in their genetic data.

Challenges of Unsupervised Learning

- The number of features (or ‘dimensions’) in unsupervised learning can often exceed that found in supervised learning tasks.
- The patterns or properties being sought are usually more complex than those in basic classification tasks.
- Unsupervised learning tends to be subjective because it lacks a definitive goal like predicting an outcome.
- Verifying the accuracy of results obtained from unsupervised learning algorithms is challenging due to the absence of a clear benchmark.
- Consequently, the unsupervised learning field includes a wide variety of methods, as effectiveness can vary based on perspective and specific application needs.

Outline

- Supervised vs. Unsupervised Learning
- ☞ **Cluster Analysis**
  - Introduction
  - Similarity/Dissimilarity Metrics
  - Clustering Methods
  - Evaluation of Clustering

General Idea for Clustering

- Clusters should be naturally occurring in data.
- Clustering analysis should discover hidden patterns in the data.
- Data objects within a cluster should be similar.
- Data objects in two different clusters should not be similar.

Cluster Analysis

- **Cluster analysis**, or **data segmentation**, has many purposes.
- It involves a wide range of methods used to identify natural groupings, or clusters, within a dataset.
- In these clusters, items are more alike to each other than they are to items in different clusters.

Typical Usage of Clustering

- **Clustering for understanding**
  - Clustering is essential in various applications to categorize data points into naturally similar groups.
  - Through clustering, we can discover underlying patterns within datasets that may not be immediately apparent.
- **Clustering for utility**
  - Clustering servers as a fundamental preprocessing step for data mining and machine learning workflows.
  - Efficiently categorizing a voluminous set of data points into fewer, manageable clusters is crucial for data summarization and facilitates a deeper comprehension of the dataset's structure.

Example: Color Quantization

- **Basics of Image Data**
  - An image is composed of pixel data arranged in a grid, each pixel containing a triplet of values representing Red, Green and Blue (RGB) components.
- **Purpose of Color Quantization**: **Color Quantization** is the technique used to reduce the number of distinct colors in an image, which is particularly useful when many pixel values are similar
- **Processes and Benefits:**
  - During quantization, pixels are grouped through clustering, with each group begin replaced by a representative color (such as the cluster's centroid value in RGB space)
  - The transformed image then displays with a simplified color scheme, limited to the number of clusters formed (e.g., an image with only 2 or 3 distinct colors).
  - Reducing the color palette is beneficial for decreasing memory usage and file size.

Color Quantization Results

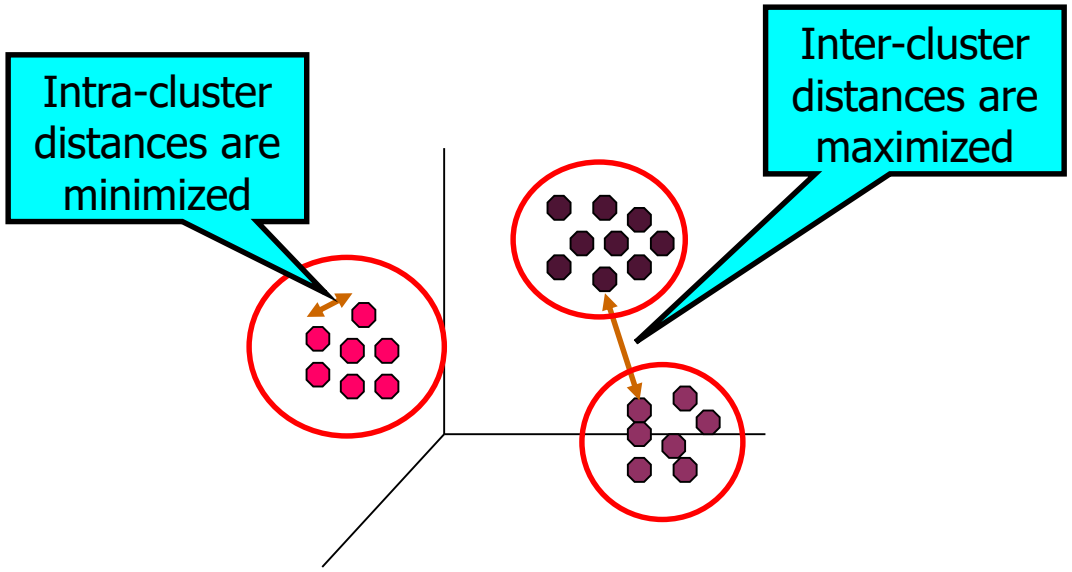
http://opencvpython.blogspot.com/2012/12/k-means-clustering-2-working-with-scipy.html

Outline

- Supervised vs. Unsupervised Learning
- Cluster Analysis
  - Introduction
  - ☞ **Similarity/Dissimilarity Metrics**
  - Clustering Methods
  - Evaluation of Clustering

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 result depends on both the similarity measure used by the method and its implementation
- The quality of a clustering method is also measured by its ability to discover some or all of the hidden patterns



Defining Similarity

- To effectively cluster, we need to establish what '***similar***' or '***different***' means for the items we're examining.
- A clustering technique then organizes items into groups based on a set rule or criteria for similarity or difference that we provide.

Similarity, Dissimilarity and Proximity

- **Similarity** measure or **similarity function**
  - A real-valued function that quantifies the similarity between two objects
  - Measure how two data objects are alike. The higher value, the more alike
  - Often falls in the range [0,1] - 0: no similarity; 1: completely similar
- **Dissimilarity** (or **distance**) measure
  - Numerical measure of how different two data objects are
  - In some sense, the inverse of similarity: The lower, the more alike
  - Minimum dissimilarity is often 0 (i.e., completely similar)
  - Range [0, 1] or [0, ∞) , depending on the definition
- **Proximity** : usually refers to either similarity or dissimilarity

Dissimilarity Matrix

- A **dissimilarity matrix** is a tool used by most clustering algorithms to measure how different, or dissimilar, each pair of observations is within a dataset.
- A dissimilarity matrix for observations x_{ij} for i=1, 2, …, N , on attributes j=1,2, ..., p is

$$\begin{array}{c} \\ x_1 \\ x_2 \\ \vdots \\ x_N \end{array} \begin{array}{c} \begin{array}{cccc} X_1 & X_2 & \cdots & X_p \end{array} \\ \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & \cdots & & \vdots \\ \vdots & \vdots & & \\ x_{N1} & \cdots & & x_{NP} \end{bmatrix} \end{array}$$

Data Matrix

$$\begin{array}{c} \\ x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_N \end{array} \begin{array}{c} \begin{array}{cccc} x_1 & x_2 & \cdots & x_N \end{array} \\ \begin{bmatrix} 0 & & & \\ D(x_2, x_1) & 0 & & \\ D(x_3, x_1) & D(x_3, x_2) & \ddots & \vdots \\ \vdots & \vdots & & \\ D(x_N, x_1) & D(x_N, x_2) & \cdots & 0 \end{bmatrix} \end{array}$$

Dissimilarity Matrix

Here, $D(x_i, x_{i'})$ is dissimilarity between two objects (or data points) i and i'

Dissimilarity with Minkowski Distance

- $r = 1$: (L_1 norm) Manhattan (or city block) distance.

  $D(x_i, x_{i'}) = |x_{i1} - x_{i'1}| + |x_{i2} - x_{i'2}| + \ldots + |x_{ip} - x_{i'p}|$

  - A common example of this is the Hamming distance, which is just the number of bits that are different between two binary vectors

- $r = 2$: (L_2 norm) Euclidean distance

  $$D(x_i, x_{i'}) = \sqrt{|x_{i1} - x_{i'1}|^2 + |x_{i2} - x_{i'2}|^2 + \ldots + |x_{ip} - x_{i'p}|^2}$$

  Squared Euclidean distance

  $$D(x_i, x_{i'}) = |x_{i1} - x_{i'1}|^2 + |x_{i2} - x_{i'2}|^2 + \ldots + |x_{ip} - x_{i'p}|^2$$

- $r \to \infty$: (L_{max} norm, L_∞ norm) "supremum" distance

  - The maximum difference between any component of the vectors

  $$D(x_i, x_{i'}) = \lim_{p \to \infty} \sqrt[m]{|x_{i1} - x_{i'1}|^m + |x_{i2} - x_{i'2}|^m + \ldots + |x_{ip} - x_{i'p}|^m}$$

  $$= \max_{1 \le f \le p} |x_{if} - x_{i'f}|$$

Similarity and Dissimilarity of Other Attribute Types

- Similarity and dissimilarity of nominal, ordinal and interval attribute values

Attribute Type	Dissimilarity	Similarity
Nominal	$d = \begin{cases} 0 & \text{if } p = q \\ 1 & \text{if } p \neq q \end{cases}$	$s = \begin{cases} 1 & \text{if } p = q \\ 0 & \text{if } p \neq q \end{cases}$
Ordinal	$d = \frac{\lvert p-q \rvert}{n-1}$ (values mapped to integers 0 to $n-1$, where n is the number of values)	$s = 1 - \frac{\lvert p-q \rvert}{n-1}$
Interval or Ratio	$d = \lvert p - q \rvert$	$s = -d$, $s = \frac{1}{1+d}$ or $s = 1 - \frac{d - min\_d}{max\_d - min\_d}$

* p and q are the two objects with one attribute of the indicate type.

Outline

- Supervised vs. Unsupervised Learning
- Cluster Analysis
  - Introduction
  - Similarity/Dissimilarity Metrics
  - ☞ **Clustering Methods**
    - *K*-mean Clustering
    - Hierarchical Clustering
  - Evaluation of Clustering

Major Clustering Methods

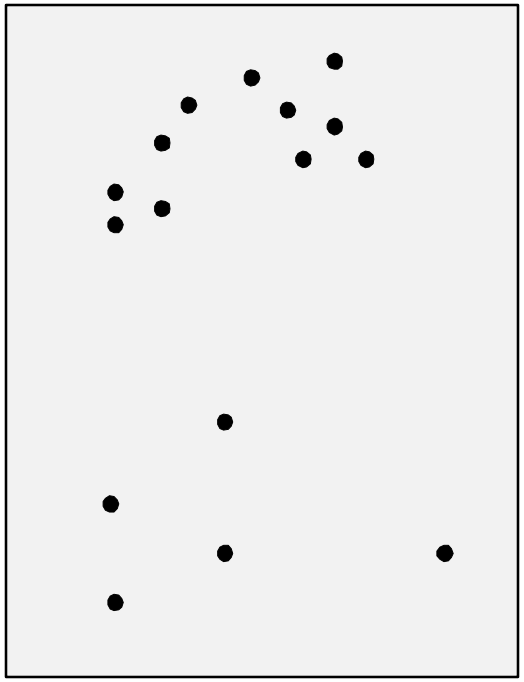
- **Partitioning-based (Representative-based) Clustering**:
  - Construct various partitions and then evaluate them by some criterion, e.g., minimizing the sum of square errors
  - k-means, k-medoids, CLARANS
- **Hierarchical Clustering:**
  - Create a hierarchical decomposition of the set of data objects using some criterion
  - Diana, Agnes, BIRCH, ROCK, CAMELEON
- **Density-based Clustering:**
  - Based on connectivity or density functions
  - DBSACN, OPTICS, DenClue
- **Grid-based Clustering:**
  - Based on a finite number of cells that form a grid structure
  - STING, WaveCluster, CLIQUE

Major Clustering Methods (Cont.)

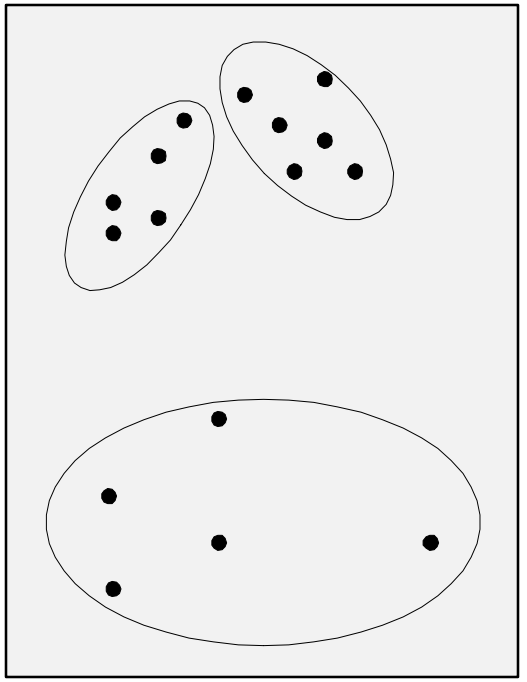
- **Probabilistic Model-based Clustering:**
  - A model is hypothesized for each of the clusters and tries to find the best fit of that model to each other
  - EM, SOM, COBWEB
- **Graph-based Clustering:**
  - Data of virtually any type can be converted to similarity graphs, then clustering
  - **Link-based clustering:**
    - Objects are often linked together in various ways
    - Massive links can be used to cluster objects: SimRank, LinkClus
- **Frequent pattern-based Clustering:**
  - Based on the analysis of frequent patterns
  - p-Cluster
- **User-guided or constraint-based Clustering:**
  - Clustering by considering user-specified or application-specific constraints: COD (obstacles), constrained clustering

Partitioning-based Clustering

- **Partitioning-based clustering** constructs a partition of a dataset *D* of *n* objects into a set of ***k*** clusters



Original Points



A Partitional Clustering (k=3)

Partitioning(/Representative)-based Approach

- Partitioning-based clustering groups data using specific representative points.
- The **partitioning representatives** can be calculated (like the average point of a cluster) or picked from actual data points within each cluster.
- **The aim is to identify representative points that best summarize the clusters.** Good representatives makes for good clusters.
- After choosing the representatives, a distance function is used to assign data points to their closest representatives.

K-means Clustering

- ***K*-means** is a method that divides a dataset of 'n' data points into '$\boldsymbol{k}$' distinct groups or clusters
- Let $C_1, \ldots, C_k$ denote sets containing n data points, $\{1, \ldots, n\}$. These sets satisfy two properties.
  1. Every item is part of a cluster, making sure all items are grouped, i.e., $C_1 \cup C_2 \cup \ \ldots \cup C_k = \{1, \ldots, n\}$. An item i is placed in cluster k, $i \in C_k$, if it's most similar to that cluster compared to the others.
  2. Clusters are exclusive. Each item can belong to one cluster only, i.e., $C_k \cap C_{k'} = \emptyset$ for all $k \neq k'$.
- Clusters are formed based on similarity, and each cluster is distinct from the others

Example: *K*-means Clustering

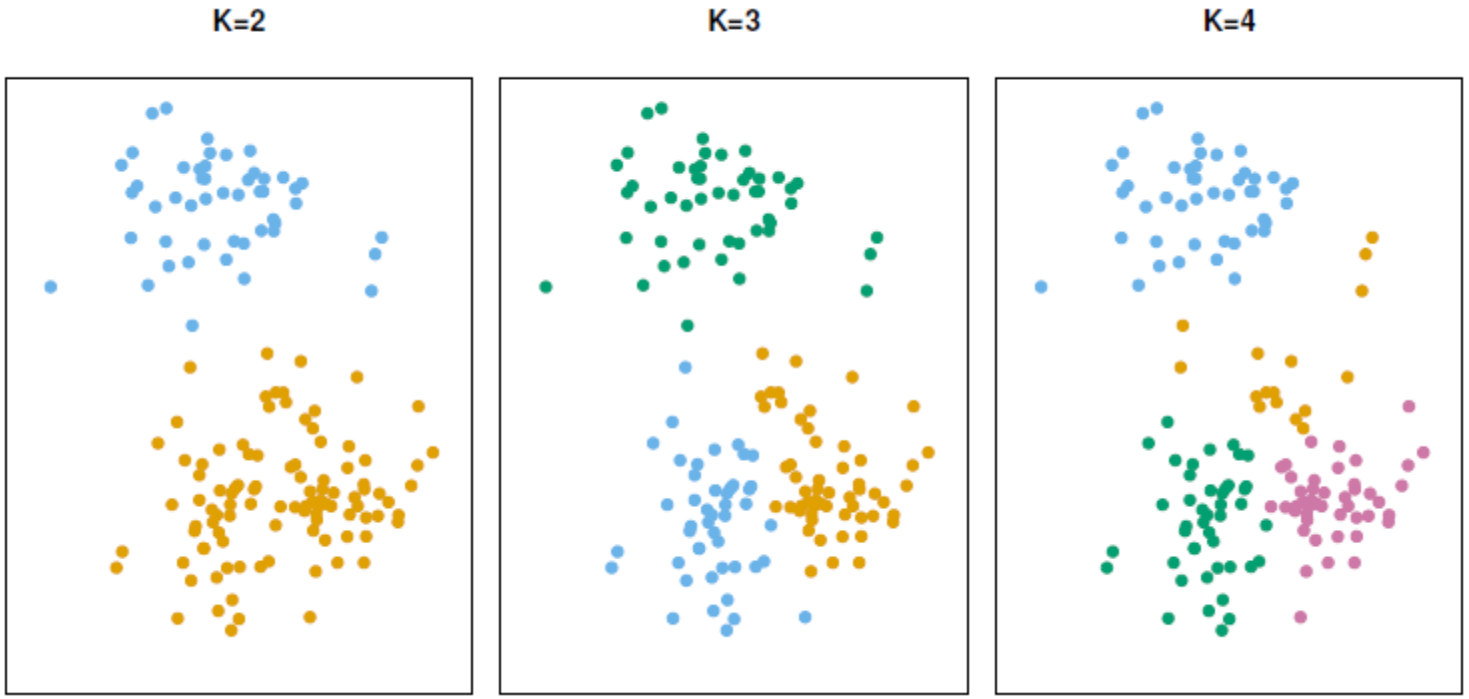


Figure. A simulated data set with 150 observations in 2-dimensional space. Panels show the results of applying K-means clustering with different values of K, the number of clusters. The color of each observation indicates the cluster to which it was assigned using the K-means clustering algorithm. Note that there is no ordering of the clusters, so the cluster coloring is arbitrary. These cluster labels were not used in clustering; instead, they are the outputs of the clustering procedure.

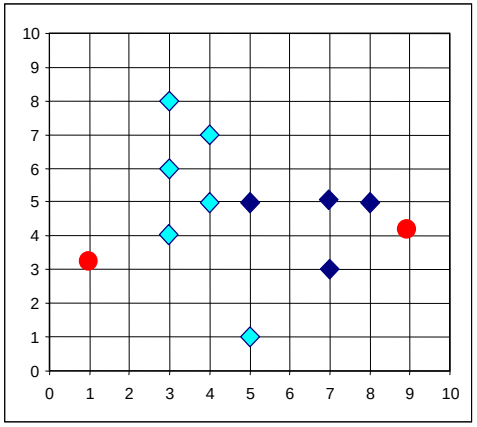
K-means Clustering Steps

- **Starting the Process**:
  1. Begin by giving every data point a cluster number from 1 to *k* at random. This is their starting cluster.
  1. Alternatively, pick *k points* in the data space as the initial cluster centers
- **Refining the Clusters**:
  - Keep reassigning data points to clusters based on the following steps until no more changes happen:
  2. **Find the Center**: For each cluster, calculate its '*centroid*' by finding the average position of all the points in it.
  3. **Reassign Points**: Move each data point to the cluster of the nearest centroid. 'Nearest' is usually measured by straight-line (Euclidean) distance.

K- Means with the Generic Algorithm

Algorithm $K - Means$ (Database: $\mathcal{D}$, Number of Representatives: k)
begin
  Initialize representative set S;
  repeat
    Create clusters (C_1 ... C_j) by assigning each
      point in $\mathcal{D}$ to closest representative in S
      using the distance function $\|x_i - c_j\|_2^2$
    Recreate set S by determining one representative c_j for
      each c_j that minimizes $\sum_{x_i \in C_j} \|x_i - c_j\|_2^2$;
  until convergence;
  return ($\mathcal{C}_1 \ldots \mathcal{C}_k$);
end

Example of *K*-means Clustering Algorithm

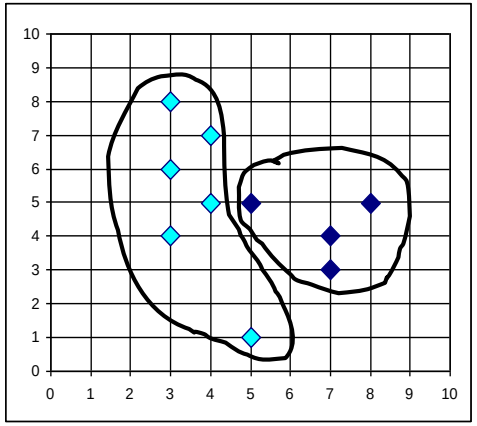


k=2

Arbitrarily choose *k* object as initial cluster center

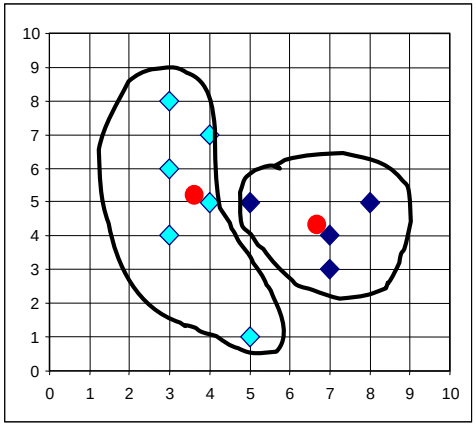
Partition objects into *k* nonempty subsets

Assign each objects to most similar center

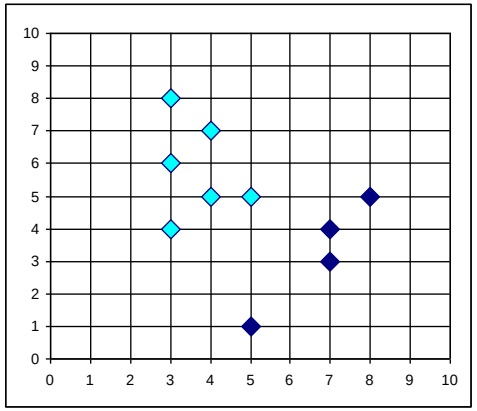


reassign

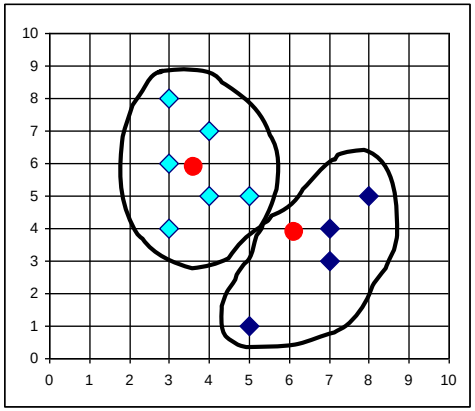
Update the cluster means



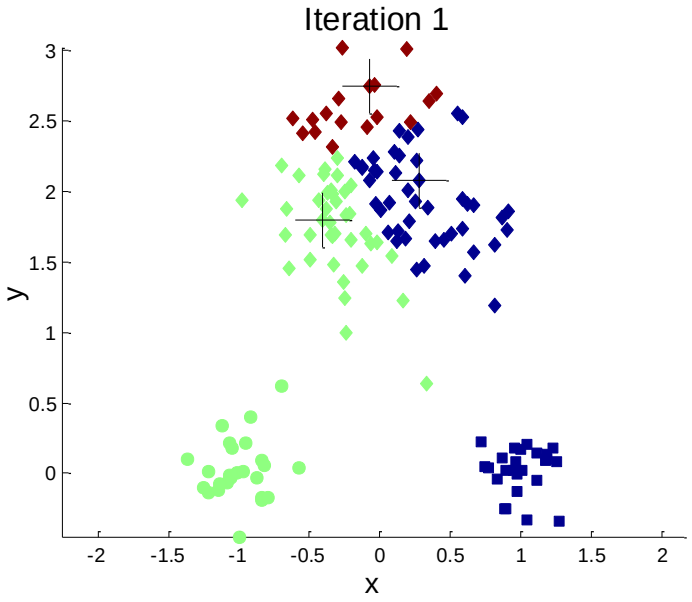
reassign

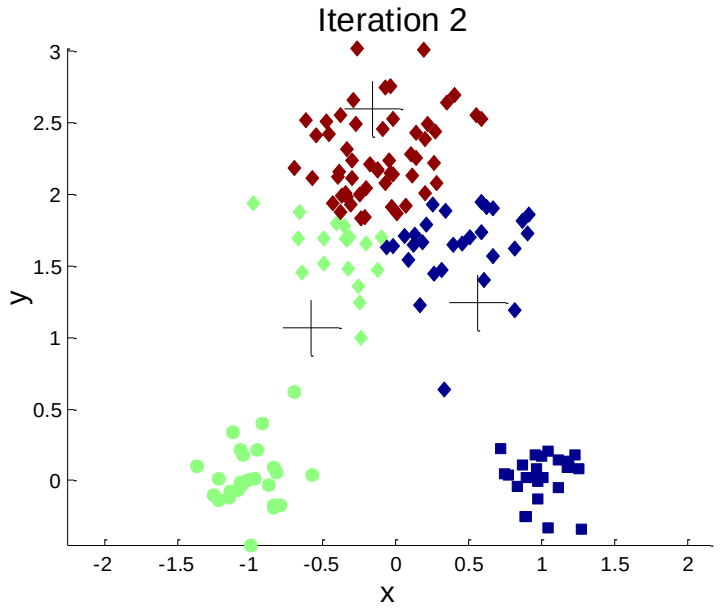


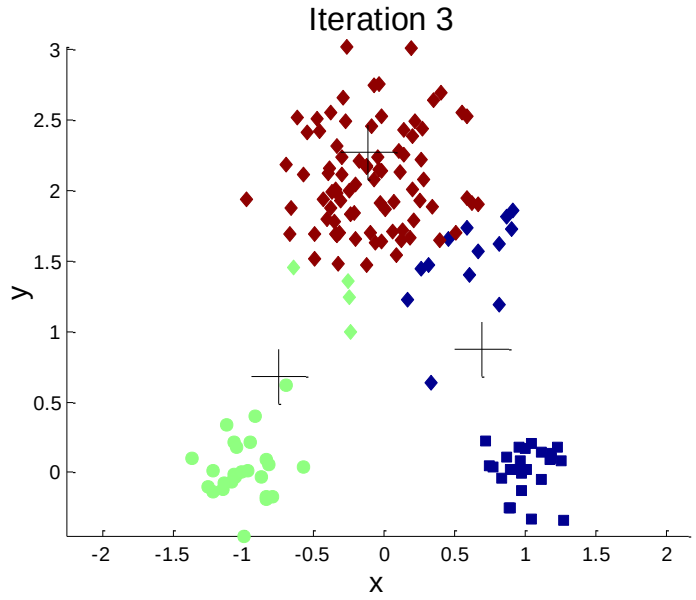
Update the cluster means

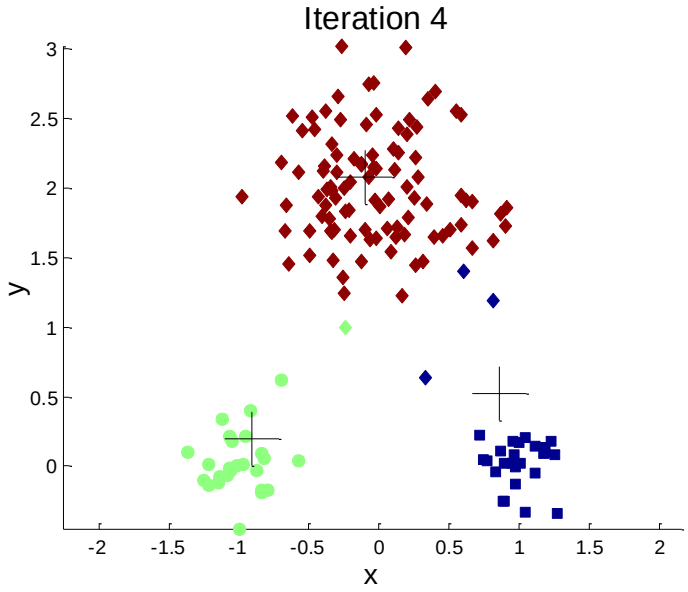


Execution Illustration (Iteration 1, 2, 3 and 4)

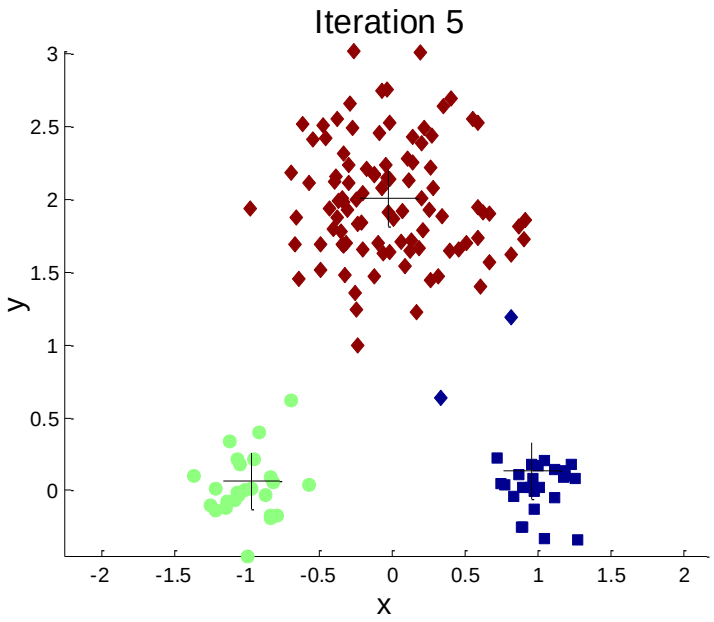


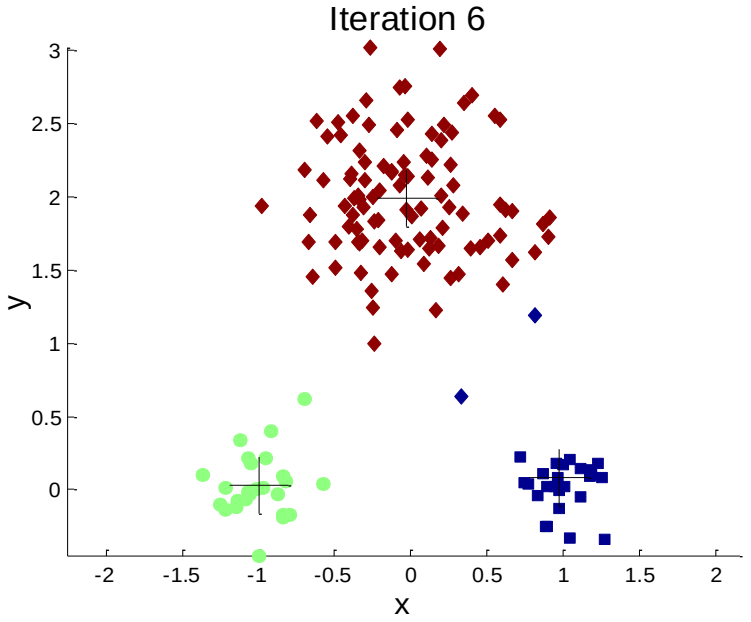






Execution Illustration (Iteration 5 and 6)





Final Result

K-means Clustering as Optimization Problem

- The idea behind *K*-means clustering is that a *good* clustering is one for which the *within-cluster variation* is as small as possible.
- The **within-cluster variation for cluster C_k, $\mathbf{WCV}(C_k)$**, is a measure of the amount by which the observations within a cluster differ from each other.
- Hence we want to solve the problem

$$\underset{C_1,\ldots,C_K}{\text{minimize}} \left\{ \sum_{k=1}^{K} \text{WCV}(C_k) \right\} \qquad (1)$$

  - to partition the observations into *K* clusters such that the total within-cluster variation, summed over all *K* clusters, is as small as possible.

K-means Clustering as Optimization Problem (cont.)

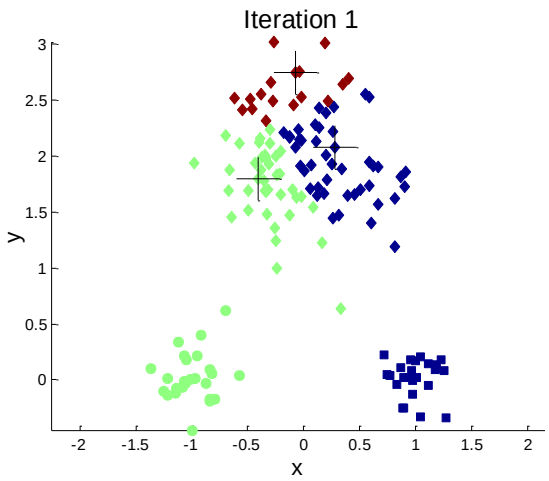
- How to define Within-Cluster Variation?
  - Typically, we use squared Euclidean distance as the dissimilarity measure $d(x_i, x_{i'}) = \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2$

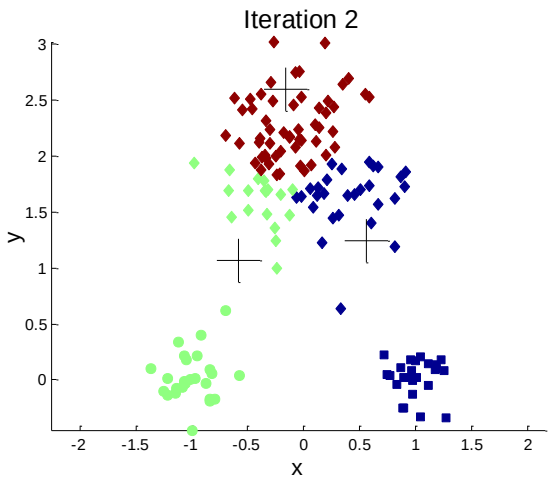
    So, $\mathrm{WCV}(C_k) = \frac{1}{|C_k|} \sum_{i,i' \in C_k} \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2$, (2)

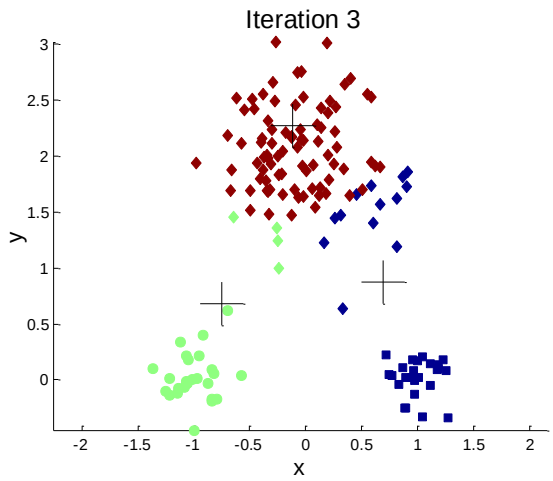
    where $|C_k|$ denotes the number of observations in the *k*th cluster and *d* is the number of features
  - Combining (1) and (2) gives the optimization problem that defines *K*-means clustering,

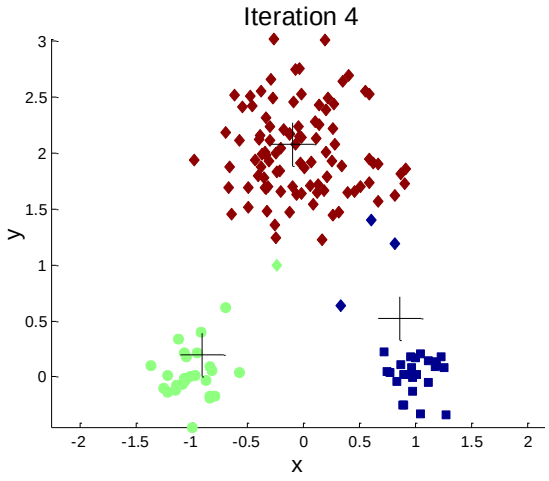
    $$\underset{C_1,\ldots,C_k}{\text{minimize}} \left\{ \sum_{k=1}^{K} \frac{1}{|C_k|} \sum_{i,i' \in C_k} \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2 \right\} \quad (3)$$

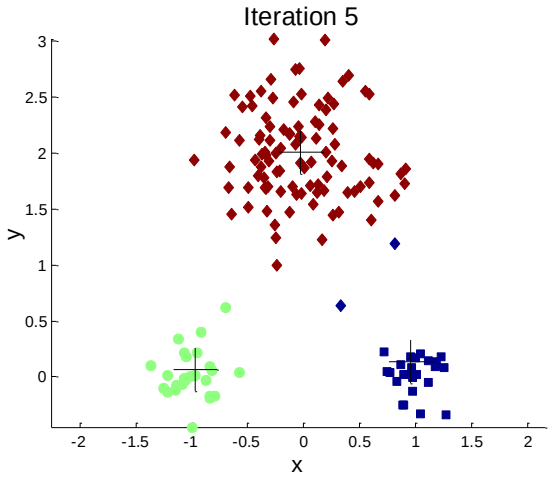
Importance of Choosing Initial Centroids

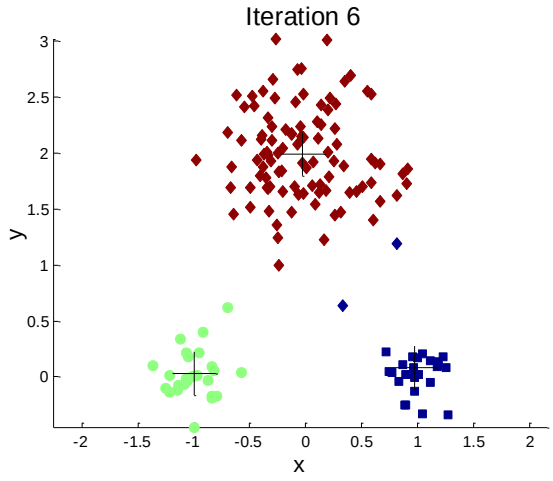




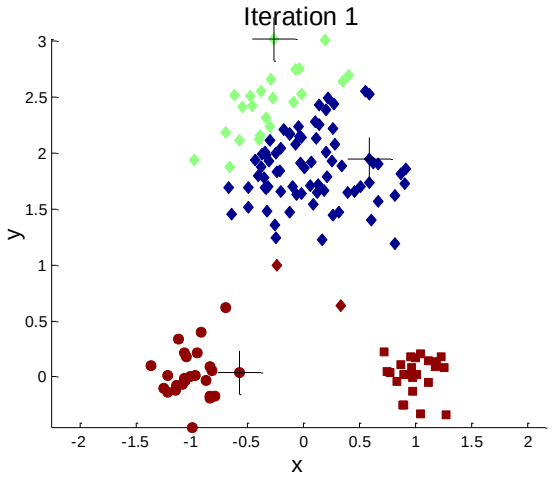


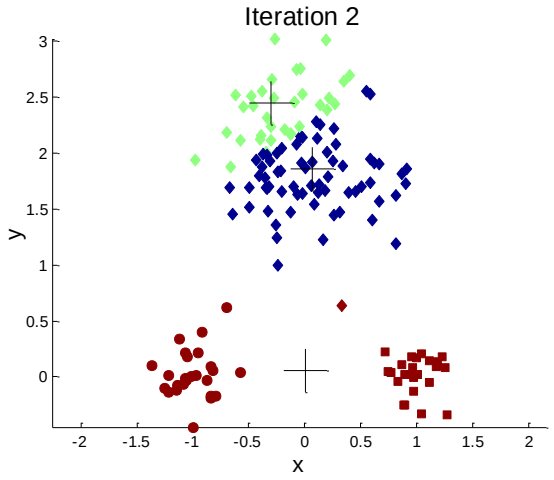


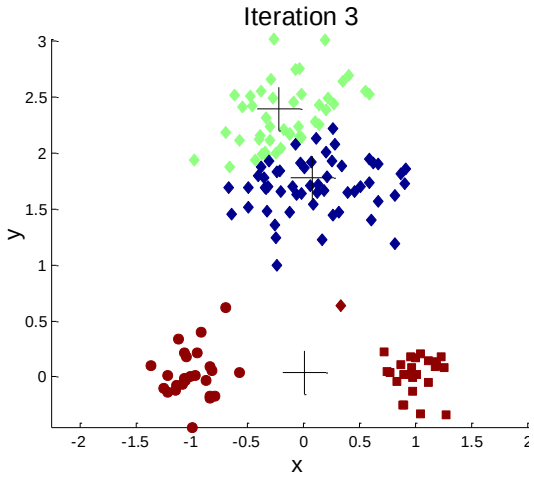


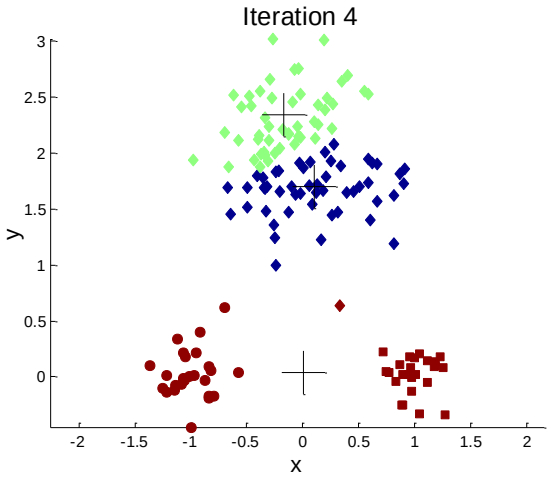


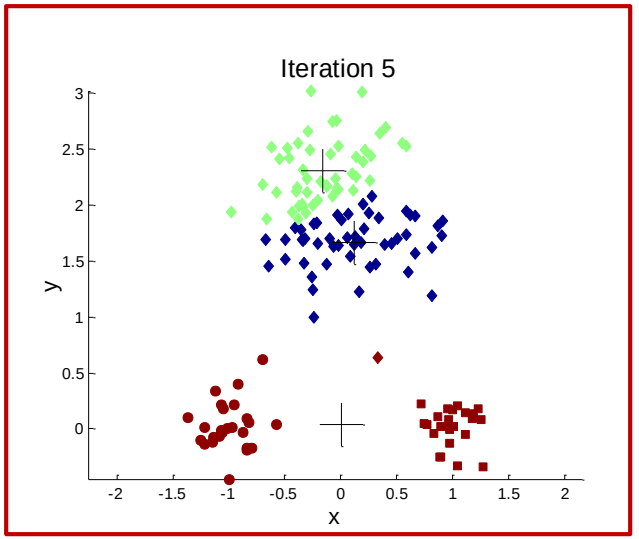
Importance of Choosing Initial Centroids



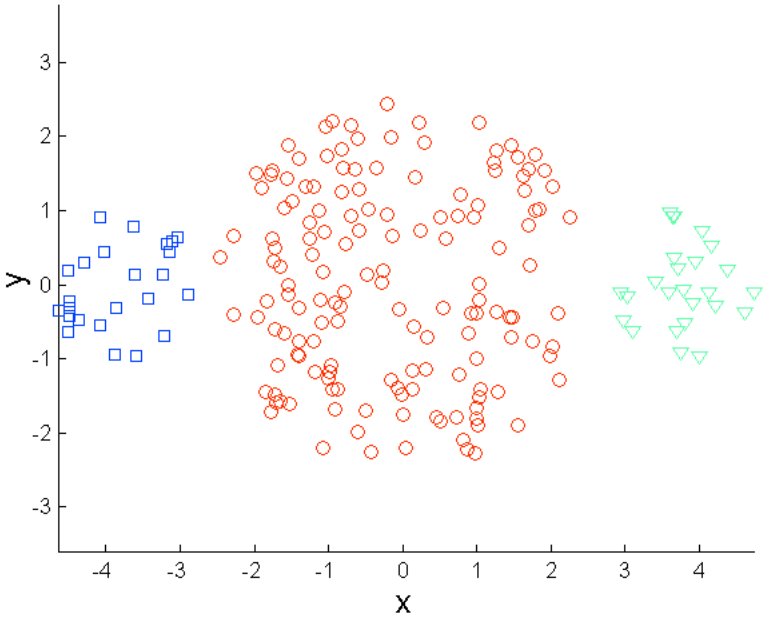






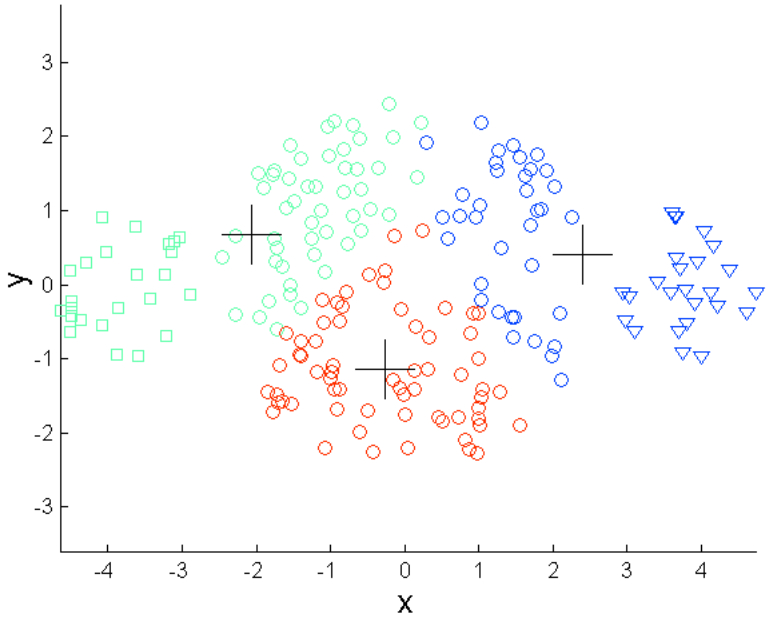


Limitation with Differing Size Data



Original Points

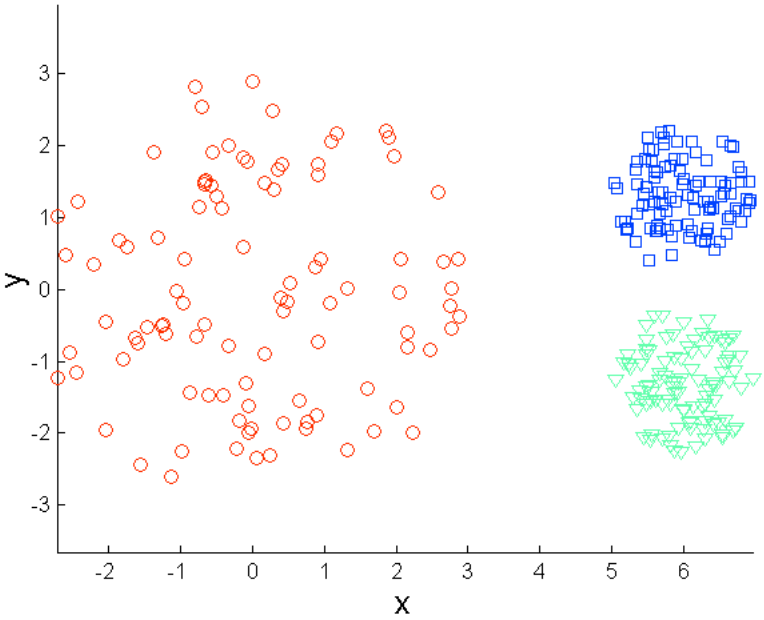
The original dataset is supposed to have one larger cluster and two smaller clusters



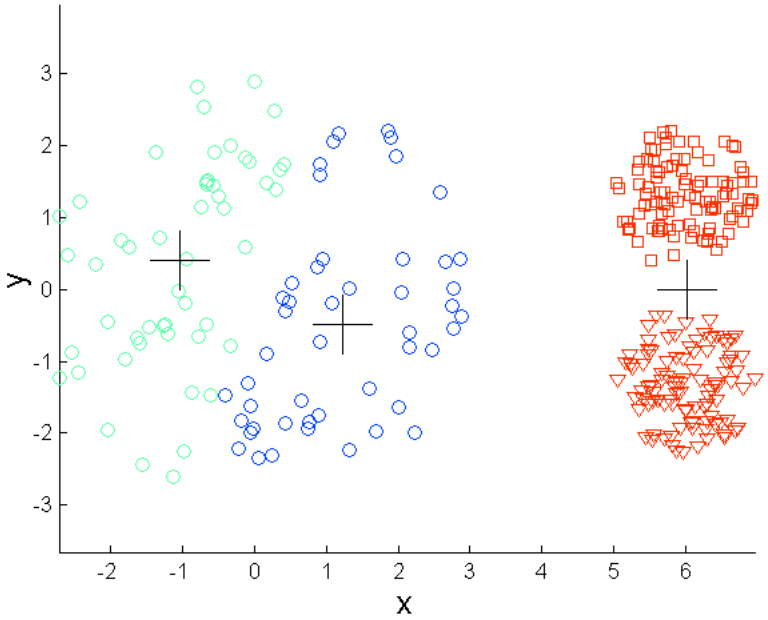
K-means (3 Clusters)

The larger cluster is broken, while one of the smaller clusters is combined with a portion of the larger cluster.

Limitation with Differing Density Data



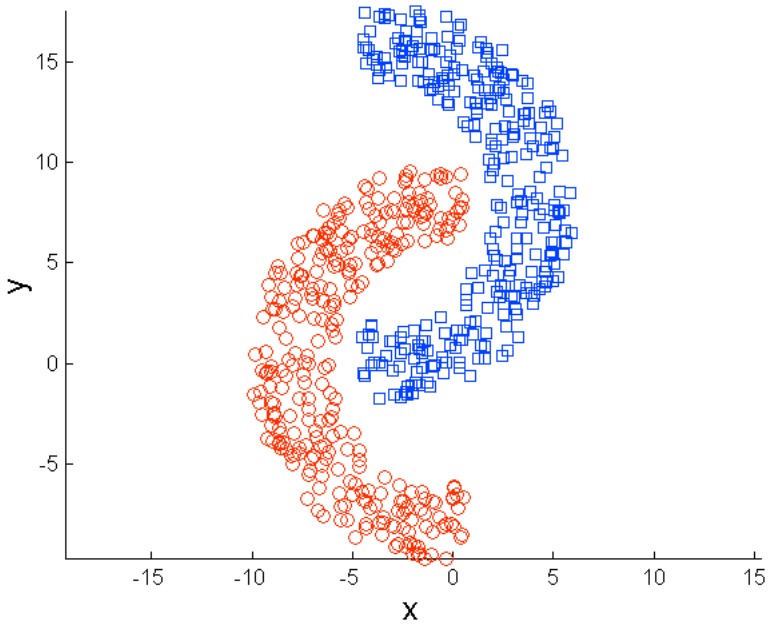
Original Points



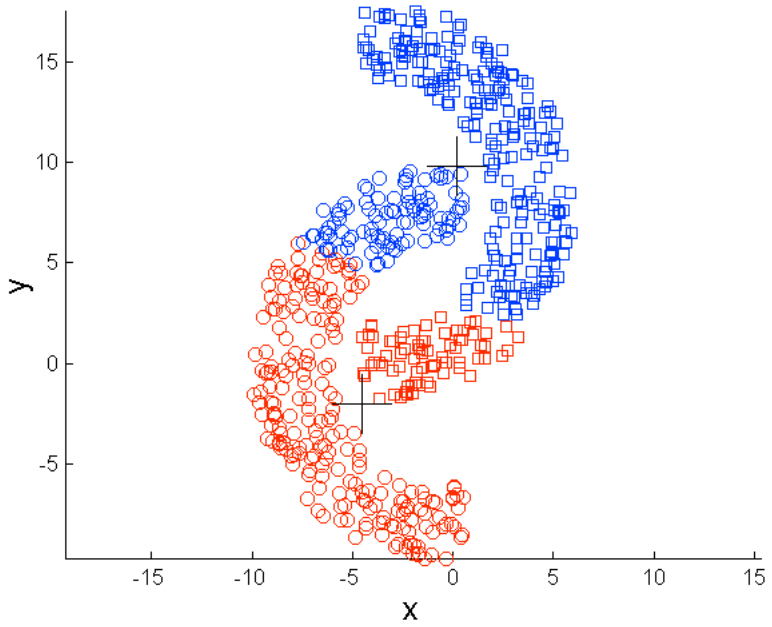
K-means (3 Clusters)

Smaller clusters are much denser than the larger cluster.

Limitation with Non-globular Shapes



Original Points



K-means (2 Clusters)

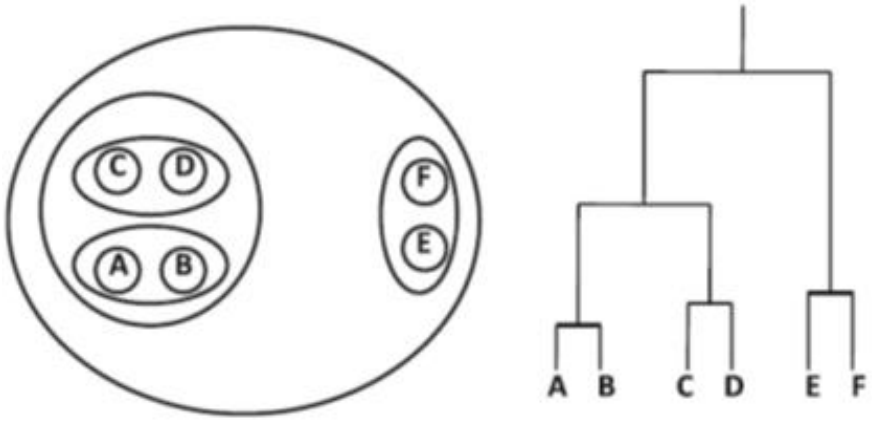
Mix portion of two clusters

Outline

- Supervised vs. Unsupervised Learning
- Cluster Analysis
  - Introduction
  - Similarity/Dissimilarity Metrics
  - Clustering Methods
    - *K*-mean Clustering
    - ☞**Hierarchical Clustering**
  - Evaluation of Clustering

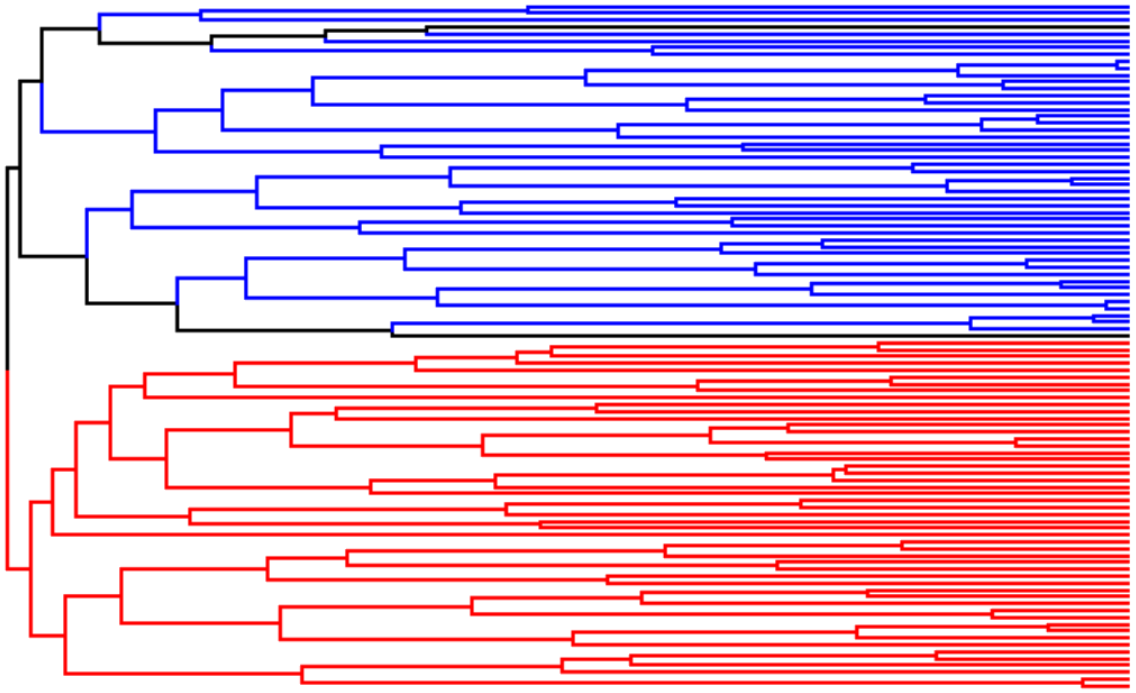
Hierarchical Clustering

- Hierarchical clustering techniques are a second important category of clustering methods
- **Hierarchical clustering** makes the hierarchical decomposition of the data based on group similarities and finds a hierarchy of clusters
- The order of merging naturally creates a hierarchical tree-like structure illustrating the relationship between different clusters, which is referred to as a ***dendrogram***.



Application

- Social media data analysis: US Senator clustering through Twitter – Can we find the party lines through Twitter?



Hierarchal clustering based on Senator's similarity

https://towardsdatascience.com/hierarchical-clustering-and-its-applications-41c1ad4441a6

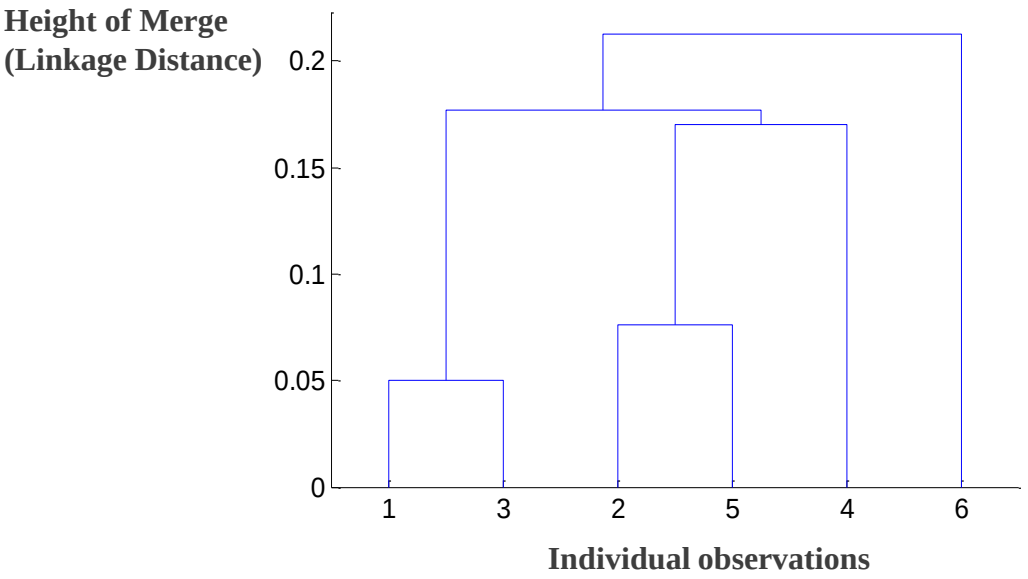
Reds are Republicans, Blues are Democrats, Blacks are independent

Strengths of Hierarchical Clustering

- Do not have to assume any particular number of clusters
  - This method does not require the number of clusters *k* as an input
- Allows to see the grouping before deciding on the number of clusters to extract

Dendrogram

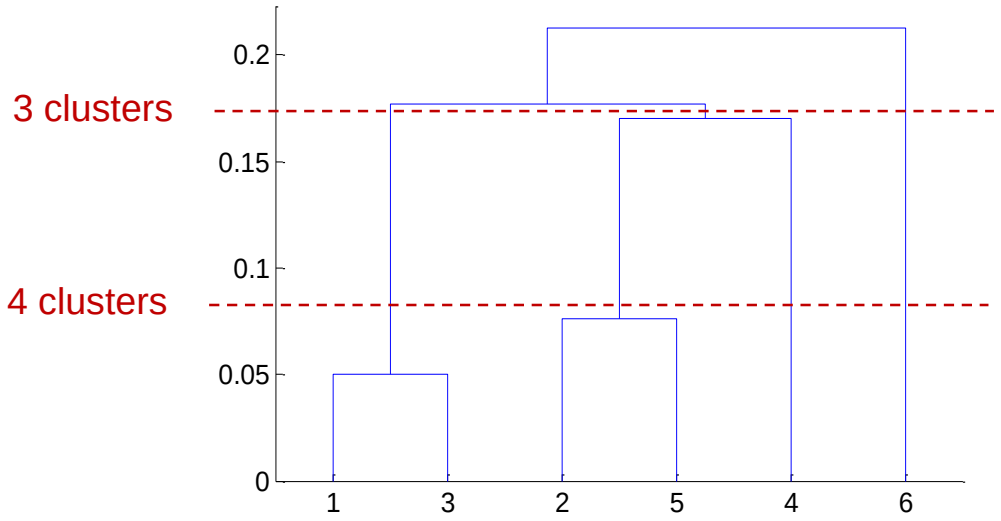
- A **dendrogram** in hierarchical clustering is a tree-like diagram that records the sequences of merges or splits that occur during the hierarchical clustering process.
- It provides a visual summary of the clustering process, displaying a picture of the groups and their proximity with a clear set of relationships outlined.



- **Y-axis**: Height of Merge (Linkage Distance) - The height at which two clusters are joined together represents the distance or dissimilarity between these clusters.
  - When two clusters combine at a low height, it suggests that they are very similar to each other.
  - Conversely, if the merge occurs at a high point on the y-axis, it implies that the clusters are quite different.
- **X-axis**: Data points

Dendrogram

- A dendrogram also gives an idea of where natural clusters may occur.
  - E.g., When consider a relative long distance in 4 clusters (0.07) and in 3 clusters (0.17), the segmentation of the data, yielding 4 clusters, might be good.



In x axis, the individual data points are arranged. y axis represents the distance between merged clusters.

Example Dataset

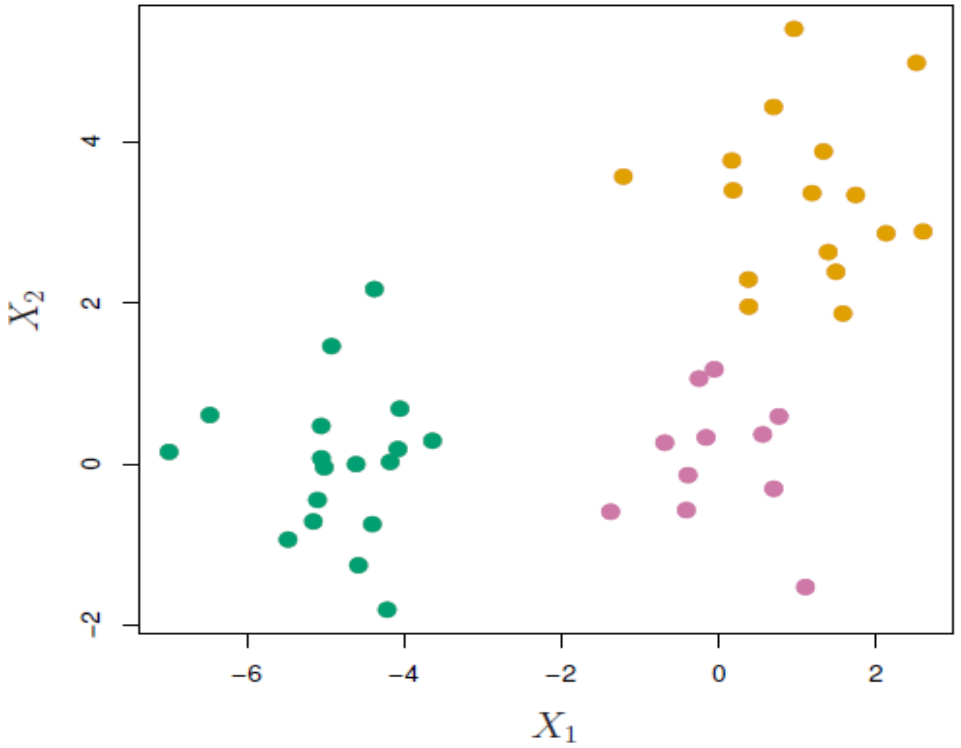
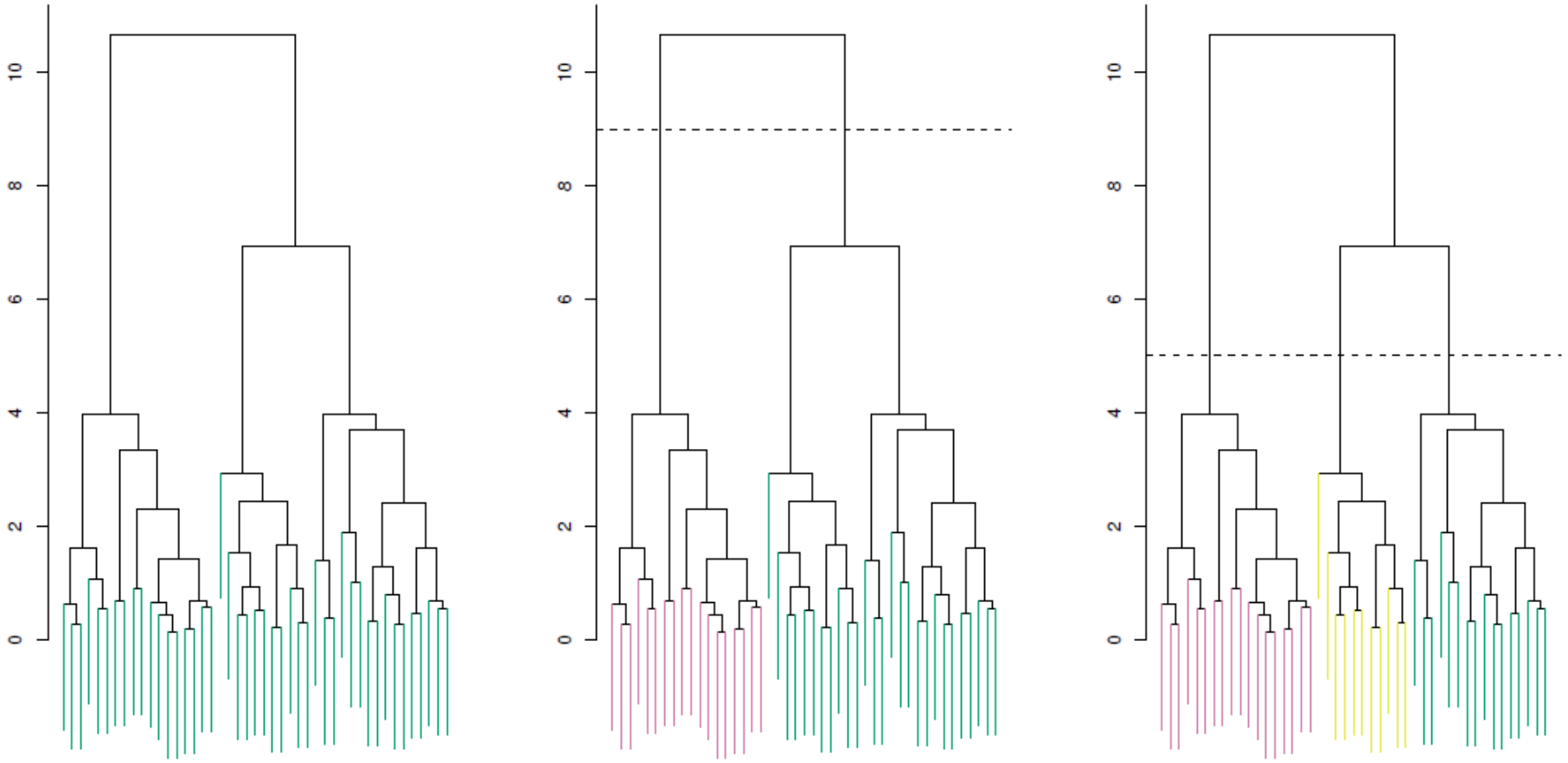


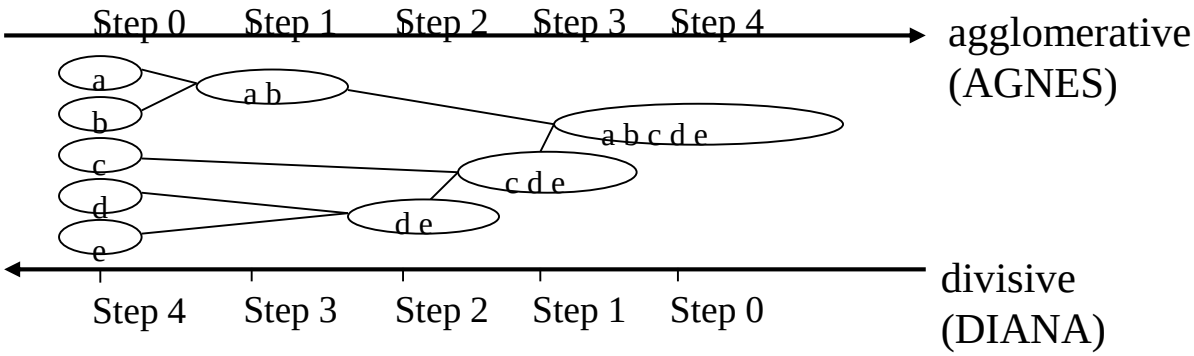
Figure. 45 observations generated in 2-dimensional space. In reality there are three distinct classes, shown in separate colors. However, we will treat these class labels as unknown and will seek to cluster the observations in order to discover the classes from the data.

Example: Number of Clusters in Hierarchical Clustering



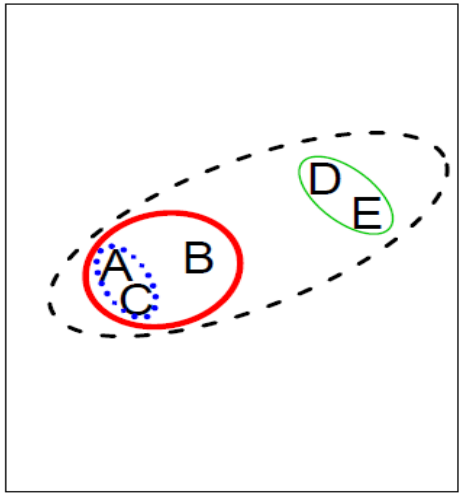
Main Types of Hierarchical Clustering

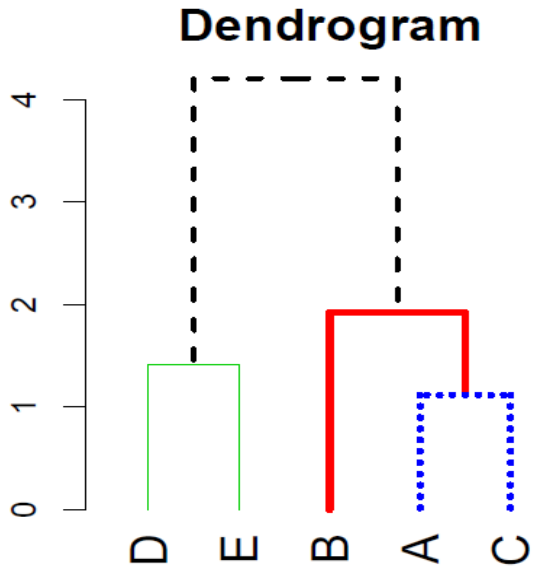
- **Agglomerative**
  - **Start with the points as individual clusters**
  - At each step, merge the closest pair of clusters until only one cluster left
- **Divisive**
  - **Start with one, all-inclusive cluster**
  - At each step, split a cluster until each cluster contains a point (or there are k clusters)



Agglomerative Hierarchical Clustering Algorithm

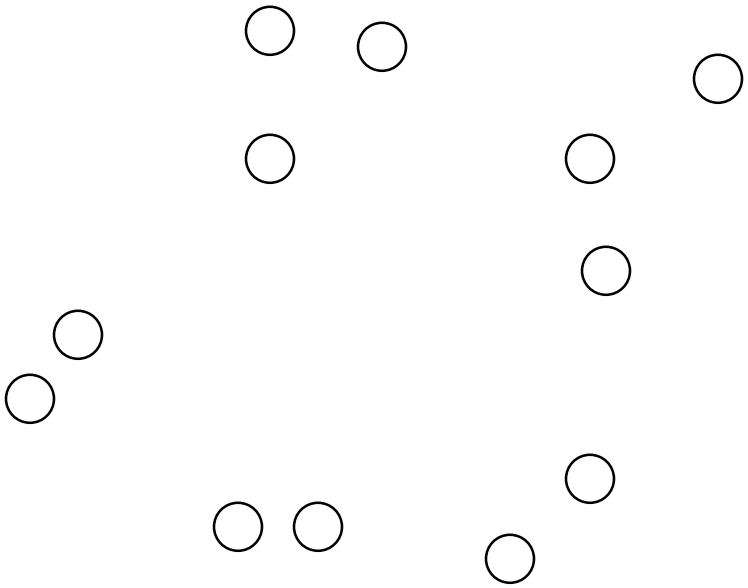
- The *bottom-up* or *agglomerative* clustering is the most common type of hierarchical clustering
- Algorithm
  - Start with each point in its own cluster.
  - Identify the **closest** two clusters and merge them.
  - Repeat.
  - Ends when all points are in a single cluster.





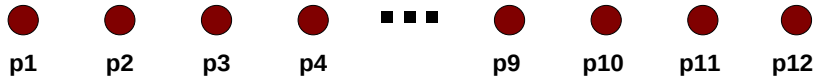
Starting Situation

- Start with clusters of individual points and a proximity matrix



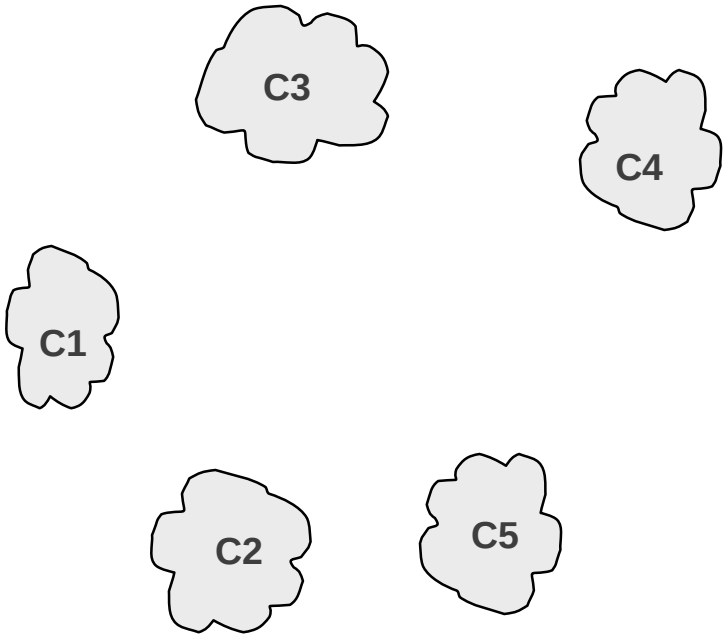
	p1	p2	p3	p4	p5	. . .
p1						
p2						
p3						
p4						
p5						
.						
.						
.						

Proximity Matrix



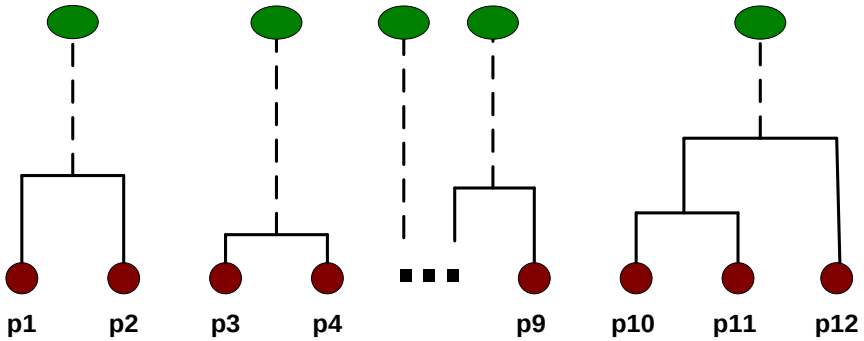
Intermediate Situation

- After some merging steps, we have some clusters



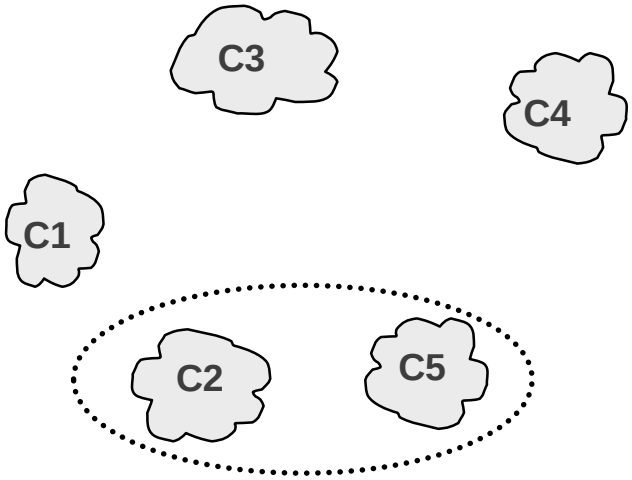
	C1	C2	C3	C4	C5
C1					
C2					
C3					
C4					
C5					

Proximity Matrix



Intermediate Situation

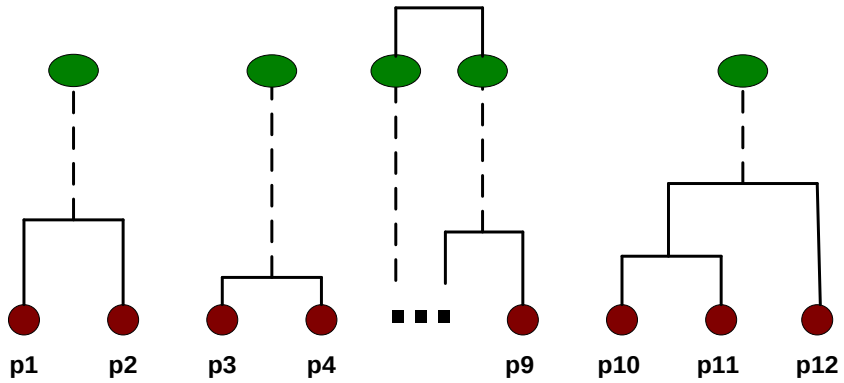
- We want to merge the two closest clusters (C2 and C5) and update the proximity matrix.



	C1	C2	C3	C4	C5
C1					
C2					
C3					
C4					
C5					

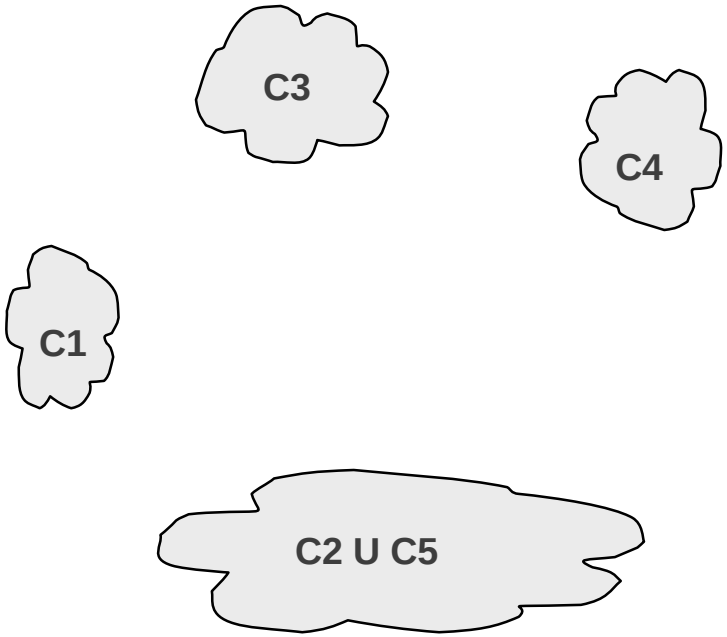
Proximity Matrix

Clusters are merged based on the similarity or distance function.



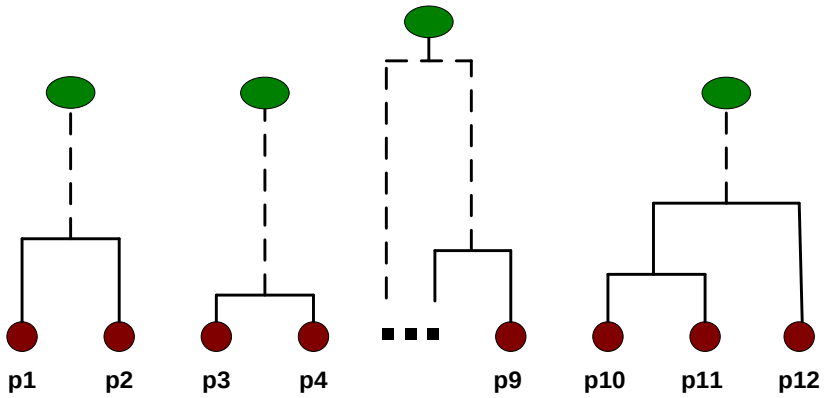
After Merging

- The question is "How do we update the proximity matrix?"



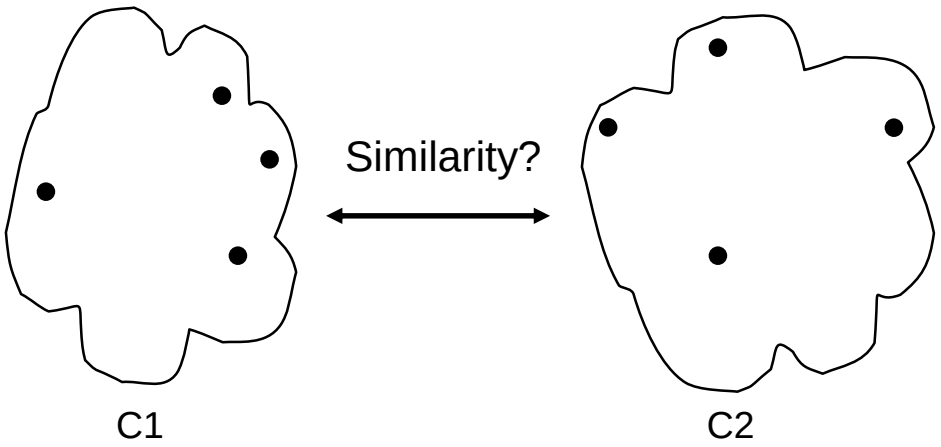
	C1	C2 U C5	C3	C4
C1		?		
C2 U C5	?	?	?	?
C3		?		
C4		?		

Proximity Matrix



How to Define Inter-Cluster Similarity

Key operation is the computation of the proximity of two clusters.



	C1	C2	C3	C4	C5	. . .
C1						
C2						
C3						
C4						
C5						
.						
.						
.						

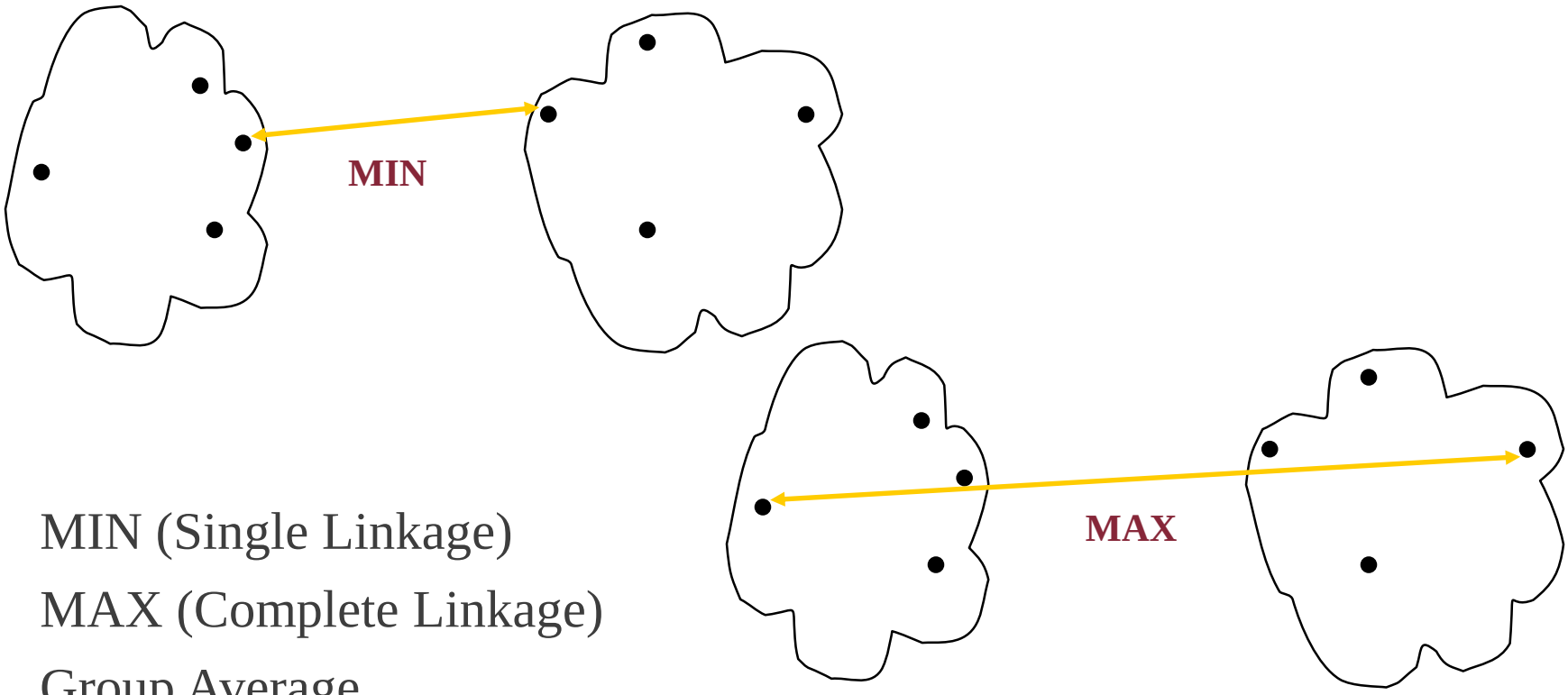
Proximity Matrix

- The distance function for the similarity between two clusters is called the ***linkage*** function.

Types of Linkage

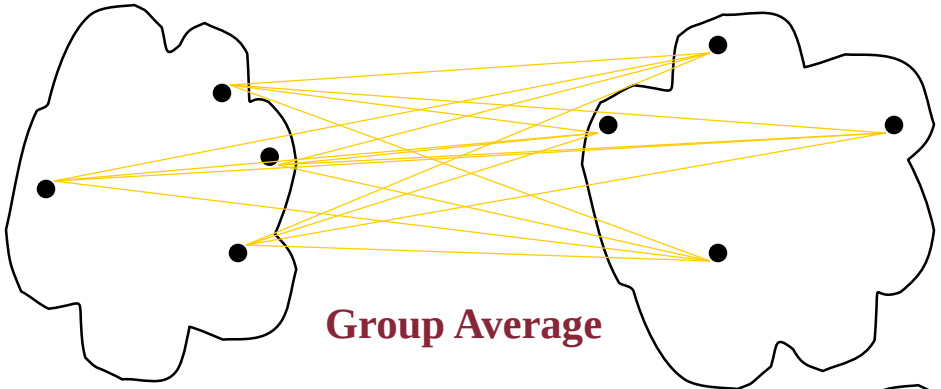
Linkage	Description
Complete	Maximal inter-cluster dissimilarity. Compute all pairwise dissimilarities between the observations in cluster A and the observations in cluster B, and record the largest of these dissimilarities.
Single	Minimal inter-cluster dissimilarity. Compute all pairwise dissimilarities between the observations in cluster A and the observations in cluster B, and record the smallest of these dissimilarities.
Average	Mean inter-cluster dissimilarity. Compute all pairwise dissimilarities between the observations in cluster A and the observations in cluster B, and record the average of these dissimilarities.
Centroid	Dissimilarity between the centroid for cluster A (a mean vector of length p) and the centroid for cluster B. Centroid linkage can result in undesirable inversions.

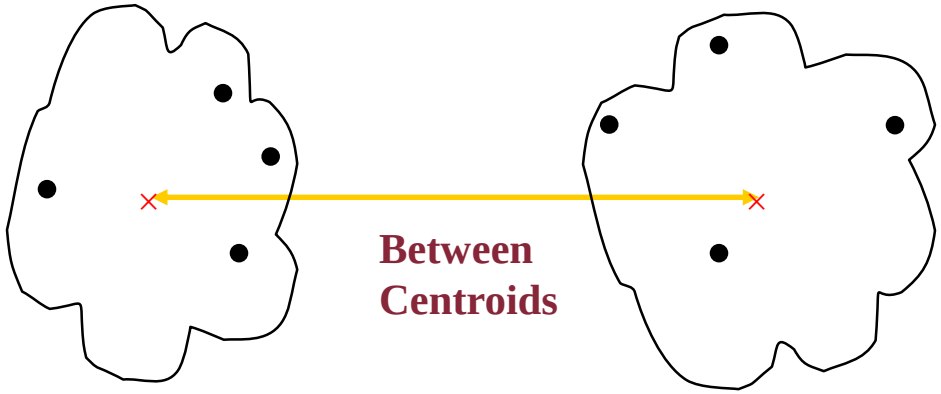
Linkage Functions



- MIN (Single Linkage)
- MAX (Complete Linkage)
- Group Average
- Between Centroids
- Other methods driven by an objective function, e.g., Ward's method

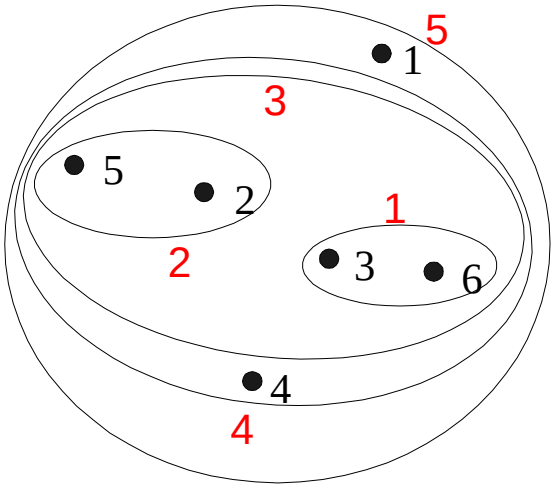
Linkage Functions





- MIN (Single Linkage)
- MAX (Complete Linkage)
- Group Average
- Between Centroids
- Other methods driven by an objective function, e.g., Ward's method

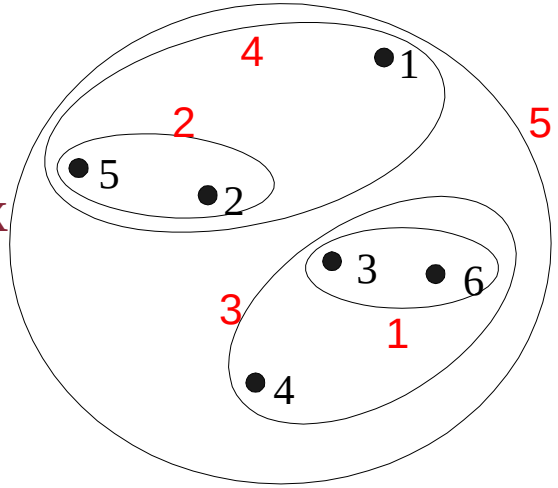
Hierarchical Clustering Results

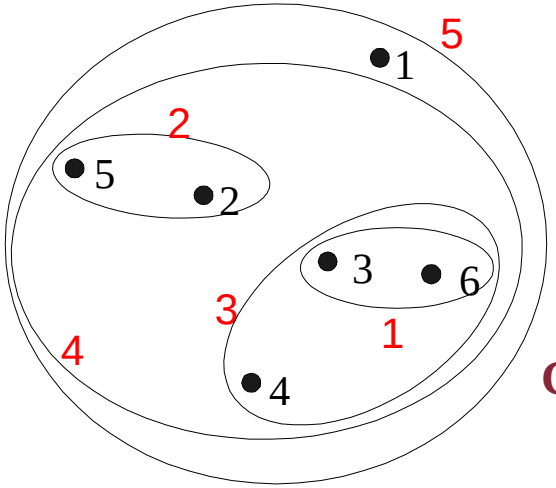


Distance function between clusters :

MIN

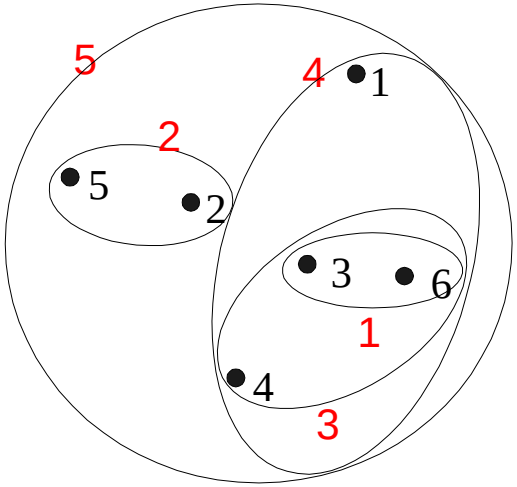
MAX



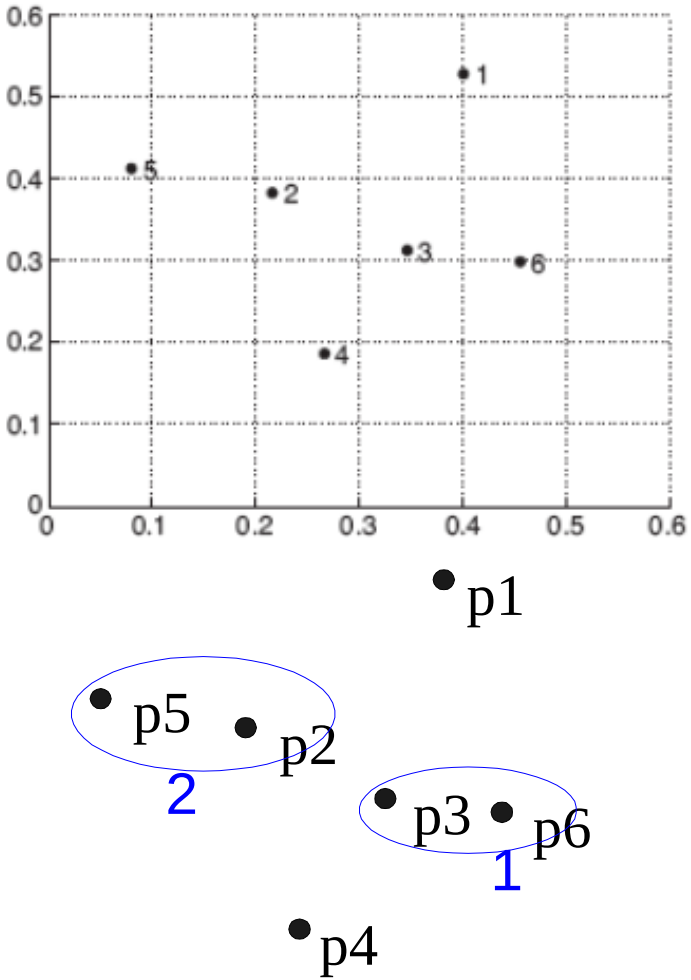


Group Average

Ward's Method



Example of MIN (Single link) Agglomerative Clustering



Point	x Coordinate	y Coordinate
p1	0.40	0.53
p2	0.22	0.38
p3	0.35	0.32
p4	0.26	0.19
p5	0.08	0.41
p6	0.45	0.30

Table 8.3. xy coordinates of 6 points.

	p1	p2	p3	p4	p5	p6
p1	0.00	0.24	0.22	0.37	0.34	0.23
p2	0.24	0.00	0.15	0.20	0.14	0.25
p3	0.22	0.15	0.00	0.15	0.28	0.11
p4	0.37	0.20	0.15	0.00	0.29	0.22
p5	0.34	0.14	0.28	0.29	0.00	0.39
p6	0.23	0.25	0.11	0.22	0.39	0.00

Table 8.4. Euclidean distance matrix for 6 points.

Similarity of two clusters is based on the two most similar (closest) points in the different clusters

MIN Clustering (cont.)

After merging {p2} and {p5},
and {p3} and {p6}

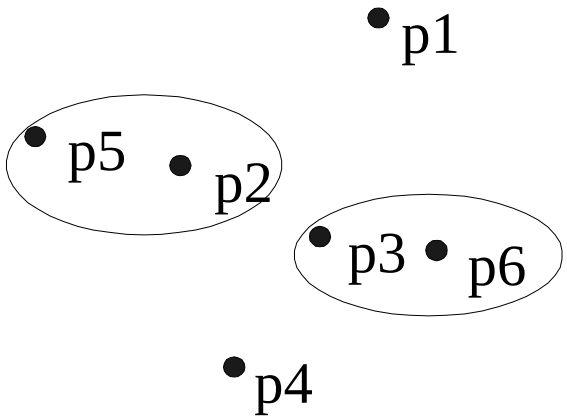
- *dist* (cluster {p3, p6}, cluster {p2,p5})=?
- *dist* ({p3,p6}, {p4}) = ?
- *dist* ({p3,p6}, {p1}) = ?

…

	P1	{P2, P5}	{P3, P6}	P4
P1	0.00	?	?	0.37
{P2, P5}		0.00	?	?
{P3, P6}			0.00	?
P4				0.00

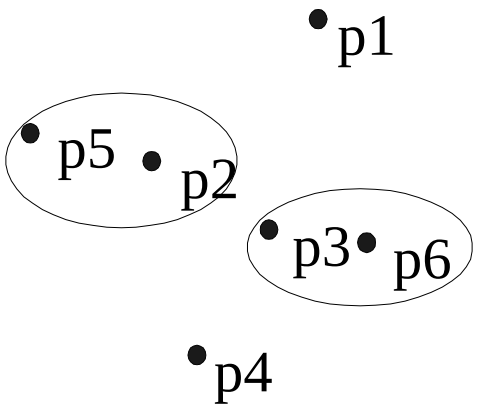
	p1	p2	p3	p4	p5	p6
p1	0.00	0.24	0.22	0.37	0.34	0.23
p2	0.24	0.00	0.15	0.20	0.14	0.25
p3	0.22	0.15	0.00	0.15	0.28	0.11
p4	0.37	0.20	0.15	0.00	0.29	0.22
p5	0.34	0.14	0.28	0.29	0.00	0.39
p6	0.23	0.25	0.11	0.22	0.39	0.00

Table 8.4. Euclidean distance matrix for 6 points.



MIN Clustering (cont.)

- *dist* (cluster {p3, p6}, cluster {p2,p5})
  = min(*dist*(p3,p2), *dist*(p6,p2), *dist*(p3,p5), *dist*(p6,p5))
  = min(0.15, 0.25, 0.28, 0.39)=0.15
- *dist* ({p3,p6}, {p4}) =min(*dist*(p3,p4), *dist*(p6,p4))
  = min(0.15, 0.22)=0.15
- *dist* ({p3,p6}, {p1}) =min(*dist*(p3,p1), *dist*(p6,p1))
  = min(0.22, 0.23)=0.22



	P1	{P2, P5}	{P3, P6}	P4
P1	0.00	0.24	0.22	0.37
{P2, P5}		0.00	0.15	0.24
{P3, P6}			0.00	0.15
P4				0.00

	p1	p2	p3	p4	p5	p6
p1	0.00	0.24	0.22	0.37	0.34	0.23
p2	0.24	0.00	0.15	0.20	0.14	0.25
p3	0.22	0.15	0.00	0.15	0.28	0.11
p4	0.37	0.20	0.15	0.00	0.29	0.22
p5	0.34	0.14	0.28	0.29	0.00	0.39
p6	0.23	0.25	0.11	0.22	0.39	0.00

Table 8.4. Euclidean distance matrix for 6 points.

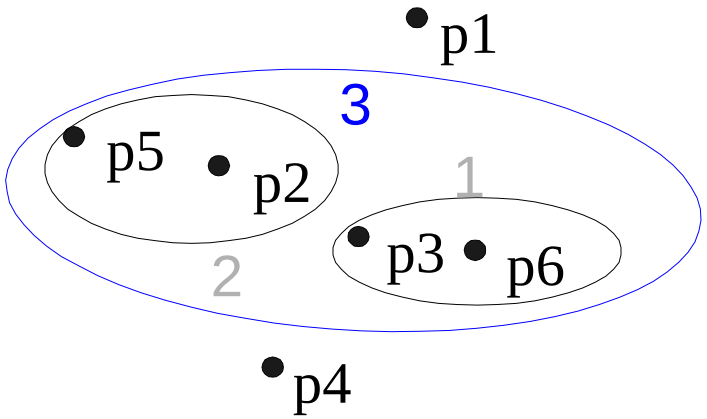
MIN Clustering (cont.)

Which cluster does the cluster {3,6} merge with ?

In Min-based clustering, similarity of two clusters is based on the two most similar (closest) points in the different clusters

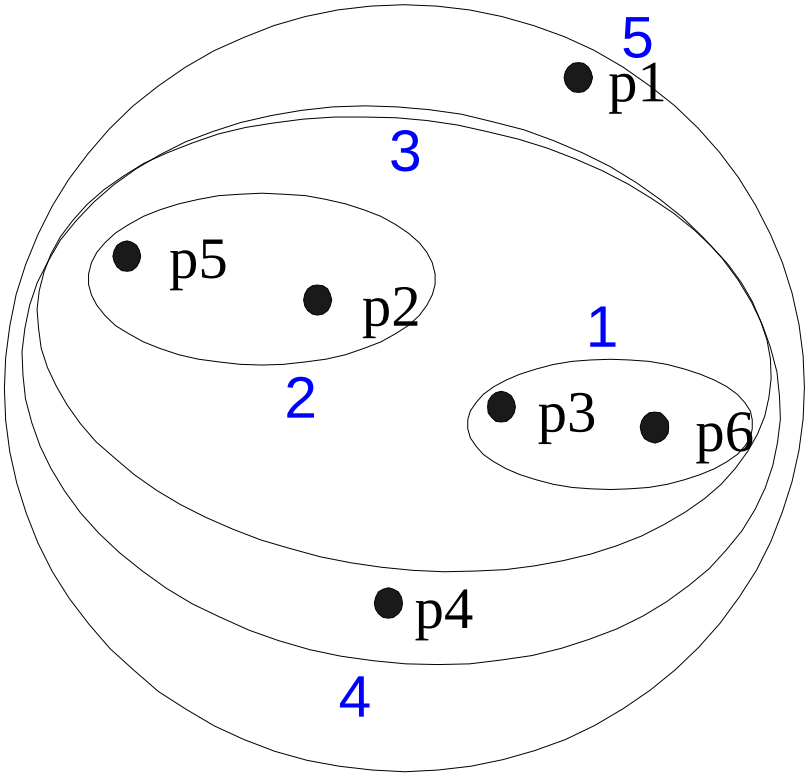
➔ Choose to merge {3,6} and {2,5} or {3,6} and {4}

	P1	{P2, P5}	{P3, P6}	P4
P1	0.00	0.24	0.22	0.37
{P2, P5}	0.24	0.00	0.15	0.24
{P3, P6}	0.22	0.15	0.00	0.15
P4	0.37	0.20	0.15	0.00

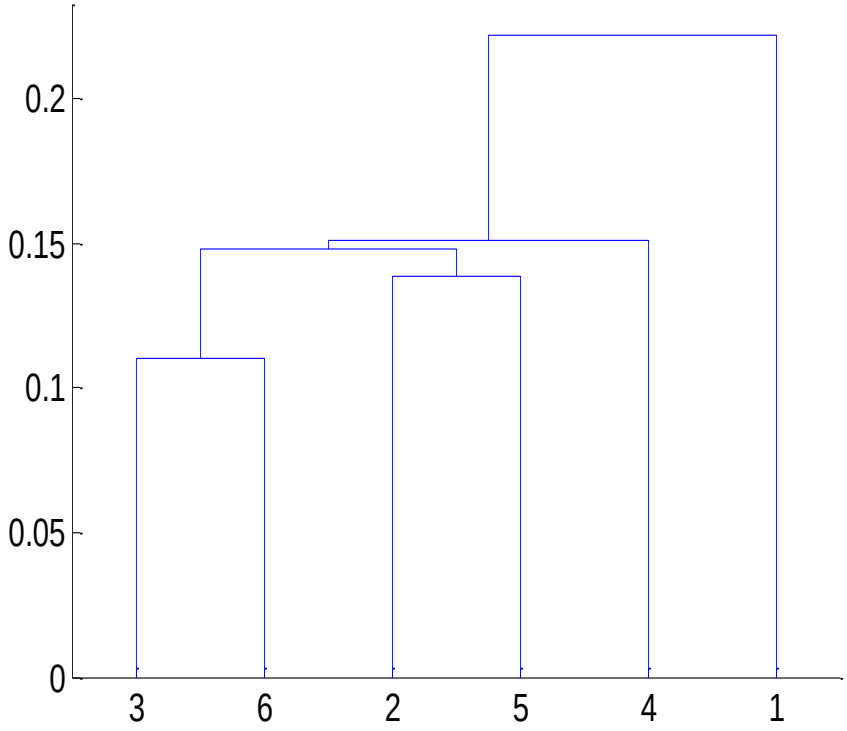


Min clustering

Final Result of MIN Clustering



Hierarchical clustering



Dendrogram

Overall Limitations

- Once a decision is made to combine two clusters, it cannot be undone
- Different schemes have problems with one or more of the following:
  - Sensitivity to noise and outliers
  - Difficulty handling different sized clusters and convex shapes
  - Breaking large clusters
- Computationally expensive
  - Do not scale well: time complexity of at least $O(n^2)$, where n is the number of total objects

Outline

- Supervised vs. Unsupervised Learning
- Cluster Analysis
  - Introduction
  - Similarity/Dissimilarity Metrics
  - Clustering Methods
  - ☞ **Evaluation of Clustering**

Cluster Validation

- For supervised classification we have a variety of measures to evaluate how good our model is. E.g., accuracy, precision, recall, and so on.
- For cluster analysis, the analogous question is how to evaluate the "goodness" of the resulting clusters.
  - But "clusters are in the eye of the beholder"!
- Then why do we want to evaluate them?
  - To compare clustering algorithms
  - To compare two sets of clusters
  - To compare two clusters
  - To avoid finding patterns in noise

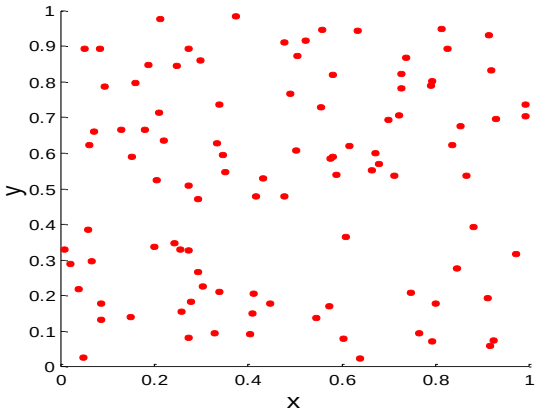
Major Tasks of Clustering Evaluation

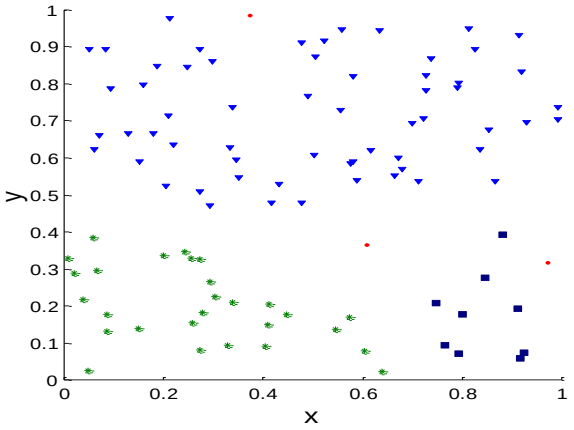
- Assessing clustering tendency
  - Assess whether a nonrandom structure exists in the data
  - Clustering analysis on a data set is meaningful only when there is nonrandom structure in the data
- Determining the number of clusters in a dataset
  - The number of clusters can be regarded as an interesting and important summary statistics of a data set. It is desirable to estimate this number before applying a clustering algorithm.
- Measuring cluster quality
  - Assess how good the resulting clusters are
  - Several measures
    - How well the clusters fits the data set
    - How well the clusters match the ground truth, if such truth is available.

Clusters found in Random Data

- Clusters found in random data are not meaningful!!

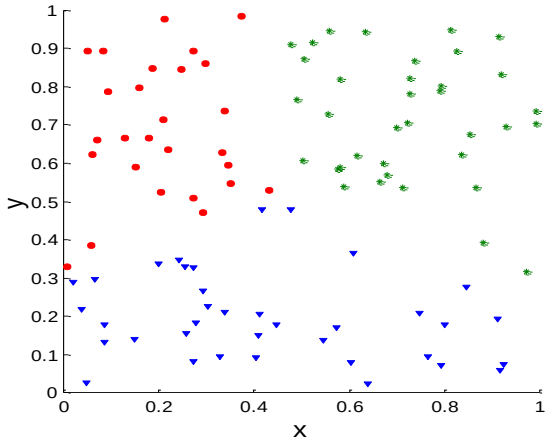
Random Points

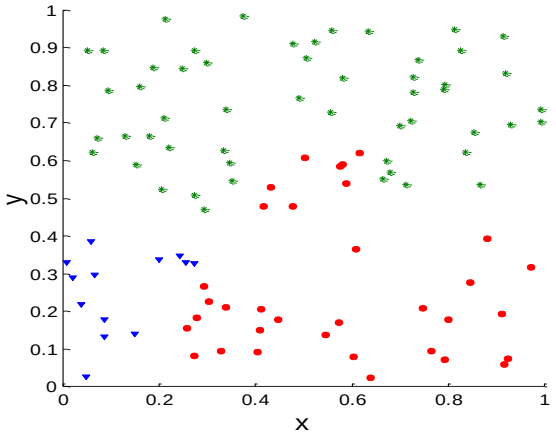




DBSCAN

K-means





Hierarchical - Complete Link

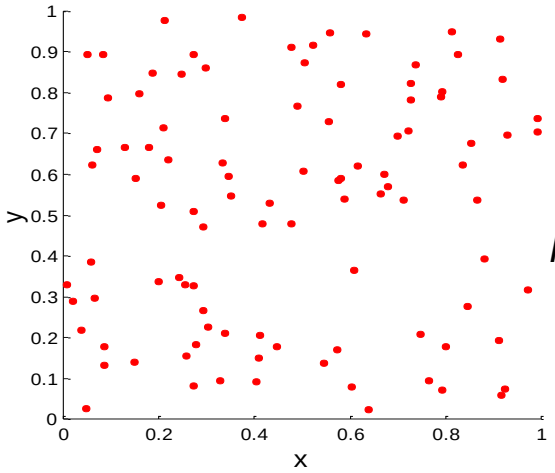
Assessing Clustering Tendency

- Assess if non-random structure exists in the data by measuring the probability that the data is generated by a uniform data distribution
- **Test spatial randomness by statistic test: Hopkins Statistic**
  - Two data sets
    - Sample n points that are randomly distributed across the data space:$\{p_1, \dots, p_n\}$
    - Sample n actual data points: $\{q_1, \dots, q_n\}$ from $\boldsymbol{D}$
  - For both sets of points, find the distance to the nearest neighbor in the original data set $\boldsymbol{D}$.
  - Two types of distances
    - $\boldsymbol{x_i}$ be the shortest distance from the artificially generated point p_i to any point in the dataset $\boldsymbol{D}$, i.e., $\min\{dist(p_i, v)\}$, where $v \in \boldsymbol{D}$
    - $\boldsymbol{y_i}$ be the shortest distance from the set of all distances between q_i and any other data point in $\boldsymbol{D}$, i.e., $\min\{dist(q_i, v)\}$, where $v \in \boldsymbol{D}$ and $q_i \in \boldsymbol{D}$, $\boldsymbol{v \neq q_i}$
  - Calculate **Hopkins Statistic:**

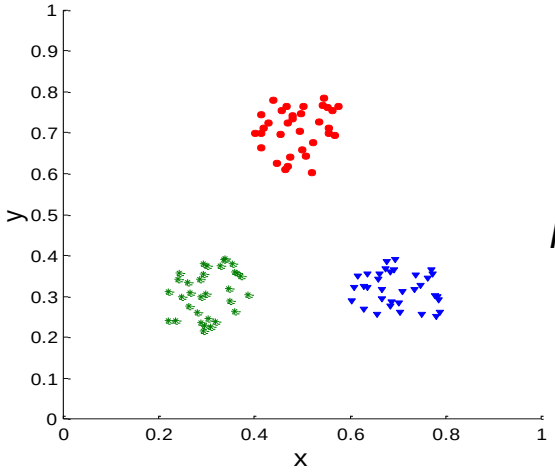
$$\boldsymbol{H = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i + \sum_{i=1}^{n} y_i}}$$

Clustering Tendency Test

- The Hopkins Statistics $H = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i + \sum_{i=1}^{n} y_i}$ is in the range (0,1).
  - If the randomly generated points and the sample of data points ($\in \boldsymbol{D}$) have roughly the same nearest neighbor distance (i.e., $\sum_{i=1}^{n} \boldsymbol{x_i} \approx \sum_{i=1}^{n} \boldsymbol{y_i}$), $\boldsymbol{H}$ is close to 0.5, implying randomness (i.e., unclustered extreme)
  - When data points (D) are clustered, $\sum_{i=1}^{n} \boldsymbol{x_i}$ would be substantially larger, thus $\boldsymbol{H}$ would be close to 0.
  - If $\boldsymbol{H} > 0.5$, than it is unlikely that D has statically significant cluster.



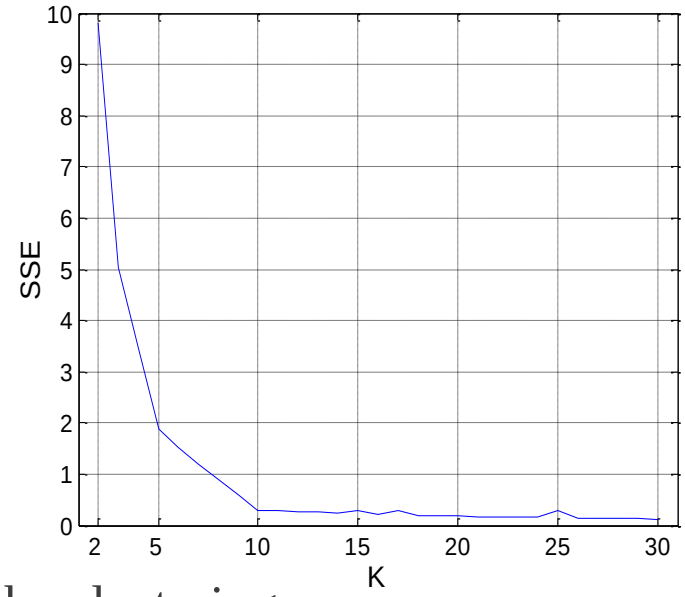
H=0.56



H=0.1

Determine the Number of Clusters

- **Simple method** (by empirical study)
  - # of clusters $\approx \sqrt{n/2}$ for a dataset of n data points
- **Elbow method**
  - Use the turning point in the curve of percentage of variance explained (e.g., SSE, Silhouette cofficient) as a function of the number of clusters



- **Cross validation method**
  - Divide a given data set into *m* parts
  - Use *m* – 1 parts to obtain a clustering model
  - Use the remaining part to test the quality of the clustering
    - E.g., For each point in the test set, find the closest centroid, and use **the sum of squared distance between all points in the test set and the closest centroids** to measure how well the model fits the test set
  - For any k > 0, repeat it *m* times, compare the overall quality measure w.r.t. different *k's*, and find # of clusters that fits the data the best

Evaluation Measures

- Cluster validation is often difficult in real data sets because the problem is defined in an unsupervised way.
- A number of *internal* criteria may be defined to validate the quality of a clustering.
- No *external* validation criteria may be available to evaluate a clustering.
- Alternatively, for real data sets, the class labels, if available, may be used as proxies for the cluster identifiers.

Evaluation Measures

- **Unsupervised**
  - Measures the goodness of a clustering structure without respect to external information
  - E.g., SSE (= SSQ)
  - Also called **internal indices (or intrinsic measures)** because they use only information presented in the data set.
- **Supervised**
  - Measures the extend to which the clustering structure discovered by a clustering algorithm matches some external structure.
  - E.g., entropy
  - Also called **external indices (or extrinsic measures)**
- **Relative**
  - For the purpose of comparison (e.g., of two K-means clustering), uses a supervised or unsupervised evaluation measure.

Unsupervised Measures

- The ground truth is unavailable
- Unsupervised metrics measures the goodness of a clustering structure without respect to external information
- They evaluate the goodness of a clustering by considering how well the clusters are separated, and how compact the clusters are.
- Often the criteria used to validate the quality of the algorithm are borrowed directly from the objective function, which is optimized by a particular clustering model.
- **Silhouette coefficient, Cohesion, Separation**, and so on.

Cohesion and Separation

- Internal measures of cluster validity for partition clustering (e.g., K-Means)
- **Cluster Cohesion**: Measures how closely related are objects in a cluster (compactness, tightness)
  - Cohesion is measured by the within cluster **sum of squares distance (SSQ)**

$$\boldsymbol{Cohesion(C) = WSS = \sum_{i} \sum_{x \in C_i} (x - m_i)^2}$$

, where x is a data point in cluster C_i and m_i is the representative point (mean) for each cluster C_i , i is the number of clusters

- **Cluster Separation**: Measure how distinct or well-separated a cluster is from other clusters (isolation)

$$\boldsymbol{Separation(C) = BSS = \sum_{i} |C_i|(m - m_i)^2}$$

, where m is a representative point (mean) for all clusters, $|C_i|$ is the size of cluster i

Intracluster to Intercluster Distance Ratio

- The cohesion measure, sum of squared distances (with absolute distances) provide no meaningful information to the user for the quality of the underlying clustering
- Intracluster to Intercluster distance ratio, $\boldsymbol{Intra/Inter}$
  - Sample r pairs of data points from the underlying data
    - P : the set of pairs that belong to the same cluster found by the algorithm
    - Q : the remaining pairs.
  - $Intra = \sum_{(x_i, x_j) \in P} dist(x_i, x_j)/|P|$
  - $Inter = \sum_{(x_i, x_j) \in Q} dist(x_i, x_j)/|Q|$
  - Small values of this measure indicate better clustering behavior.

Sihouette Coeffient

- For each $o \in D$,
  - Compute $a(o)$ as the average distance between o and all other points in the cluster to which o belongs
    - Smaller value, the cluster is more compact
  - Compute $b(o)$ as the average distance between o and all other points in the next nearest cluster
- The **silhouette coefficient** of o is defined as

$$s(o) = \frac{b(o) - a(o)}{\max(a(o), b(o))}$$

  - $-1 \leq s(o) \leq 1$
  - If $s(o)$ is negative (i.e., $b(o) < a(o)$), o is closer to the points in another cluster than to the points in the same cluster as o. Bad!
- For a cluster C_i's fitness, use the average silhouette coefficient , $\frac{\sum_{o \in C_i} s(o)}{|C_i|}$

External Validation Criteria

- Measures the goodness of a clustering structure **with external information**
- In the context of real data sets, **class labels** are used for the external information when they are available
  - The major risk with the use of class labels is that these labels are based on application-specific properties of that data set and may not reflect the natural clusters in the underlying data.

Supervised Measures of Cluster Validity

- Classification-oriented measures of cluster validity
  - (Cluster) Purity
  - (Class-based) Entropy, (Class-based) Gini index
  - Precision
  - Recall
  - F-measure
- Similarity-oriented measures of cluster validity
- Others, e.g., Bcubed precision, Bcubed recall

Example: Clustering Validity with Class Labels

Table 5.9. K-means Clustering Results for LA Document Data Set

Cluster	Entertainment	Financial	Foreign	Metro	National	Sports	Entropy	Purity
1	3	5	40	506	96	27	1.2270	0.7474
2	4	7	280	29	39	2	1.1472	0.7756
3	1	1	1	7	4	671	0.1813	0.9796
4	10	162	3	119	73	2	1.7487	0.4390
5	331	22	5	70	13	23	1.3976	0.7134
6	5	358	12	212	48	13	1.5523	0.5525
Total	354	555	341	943	273	738	1.1450	0.7203

- Ideally, each cluster will contain documents from only one class. In reality, each cluster contains documents from many classes.
- In this example, many clusters contain documents primarily from one class.
- In particular, cluster 3, which contains mostly documents from the Sports section, is exceptionally good , both in terms of purity and entropy.