EE5907: Pattern Recognition &
EE5026: Machine Learning for Data Analytics

Dr. WANG Si

Email: si.wang@nus.edu.sg

Office: E1a-04-01

Outlines

- Pattern Representation Learning
  - Unsupervised Representation Learning (week 7)
  - Supervised Representation Learning (week 8)
  - Unified Framework: Graph Embedding (week 8)
- Patter Recognition Models
  - Clustering (K-means, Agglomerative)(week 9)
  - Gaussian Mixture Model and Boosting (week 10)
  - Support Vector Machines (week 11)
- Deep Learning (week 12)
- Revision and Q&A (week 13)

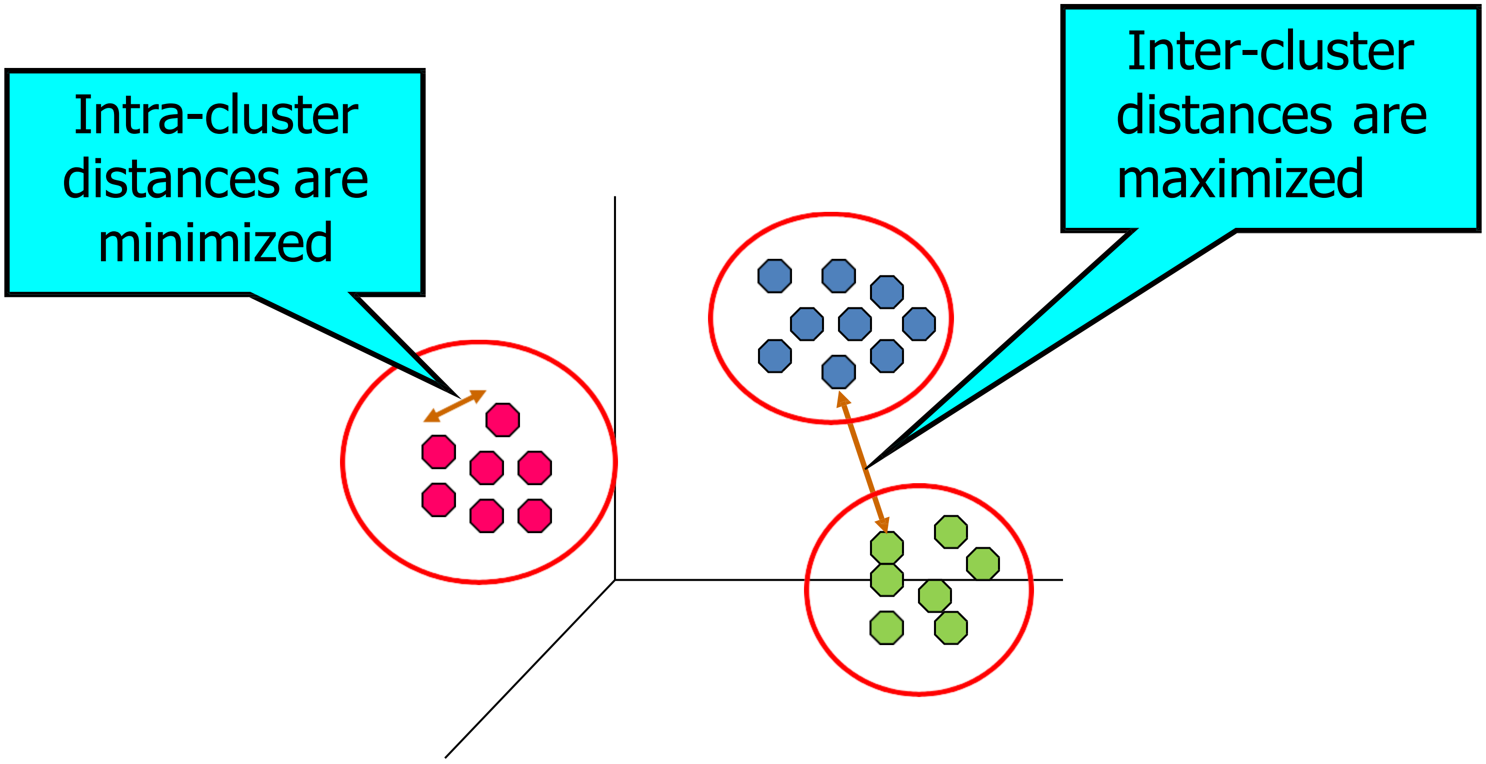
Some slides are from Shuicheng Yan, Henry Lin

Clustering

- We will discuss:
  - Definition and types of Clustering, and types of clusters
  - K-means clustering
  - Agglomerative clustering
- At the end of this lecture, you should be able to:
  - Identify type of clustering and cluster
  - Perform K-means and Agglomerative clustering
  - Understand strengths and limitations

What is Clustering?

- Process of grouping a set of objects into clusters based on similarity.



Types of Clustering

- Hard Clustering
  - Each data point belongs to a cluster completely or not.
  - Similarity measures, such as cosine similarity, Euclidean distance, etc.
  - e.g., K-means, Hierarchical, etc.
- Soft Clustering
  - Each data point belongs to each cluster with a probability or degree of membership.
  - e.g., Gaussian mixture model, Fuzzy C-means, etc.

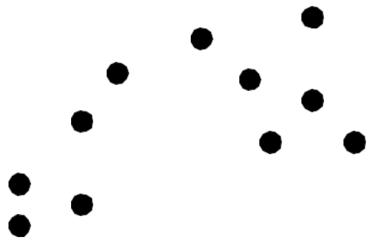
Data Points	Hard	Soft	
	Clusters	P(C1)	P(C2)
A	C1	0.91	0.09
B	C2	0.3	0.7
C	C2	0.17	0.83
D	C1	1	0

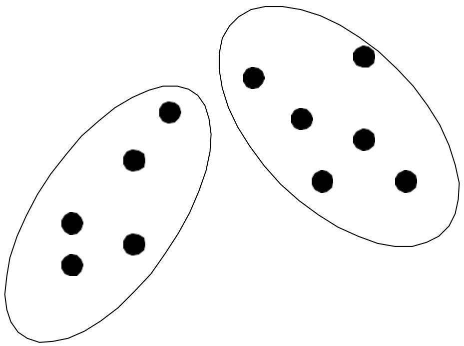
Hard Clustering

- Partition-Based
  - Divide data into non-overlapping clusters such that each data point belongs to exactly one cluster
  - e.g., **K-Means**, K-Medoids, CLARA, etc.
- Hierarchical
  - Create a nested hierarchy of clusters but assigns each data point to only one cluster at any given level
  - e.g., **Agglomerative**, Divisive, etc.
- Density-Based
  - Clusters are formed based on regions of high density, and points that don't fit are considered noise.
  - e.g., DBSCAN, OPTICS, etc.
- Graph-Based
  - Clusters are identified using graph structures, where nodes represent data points, and edges represent similarity.
  - e.g., Spectral Clustering, Min-Cut Clustering, etc.

Partitional Clustering

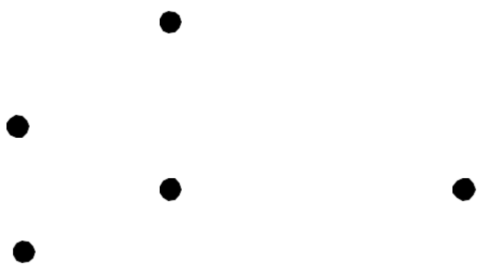
A division of data objects into **non-overlapping** subsets (clusters) such that each data object is in exactly one subset

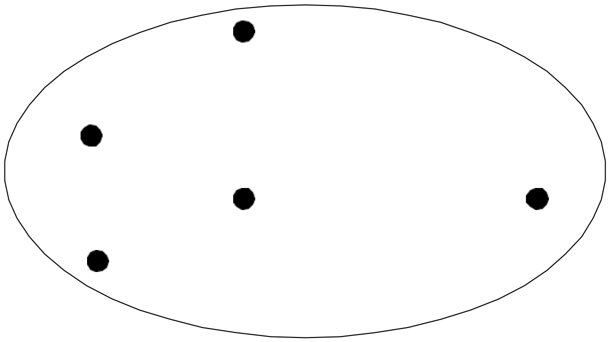




Original Points

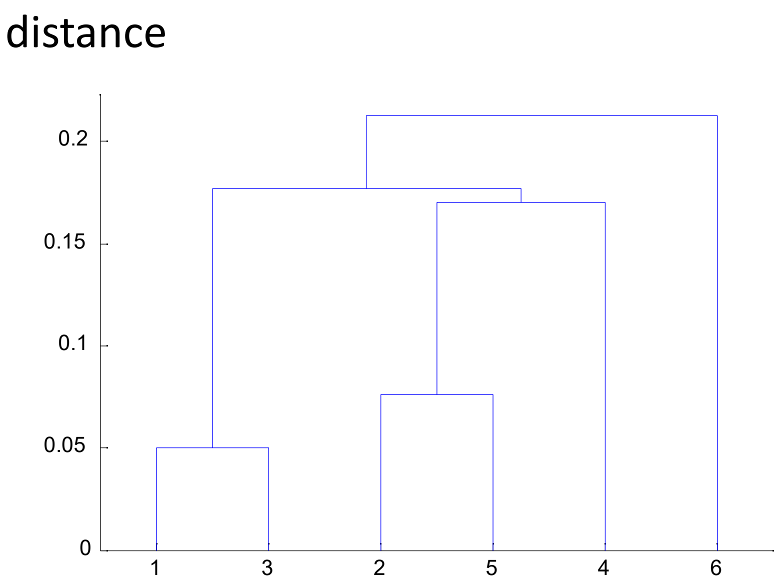
A Partitional Clustering



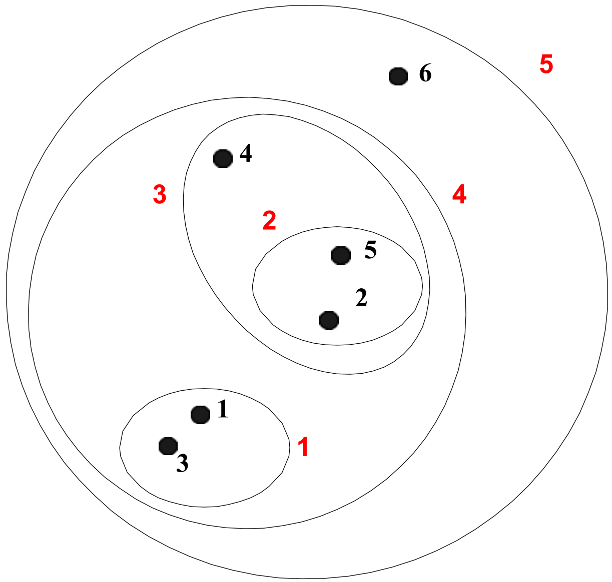


Hierarchical Clustering

A set of nested clusters organized as a hierarchical tree

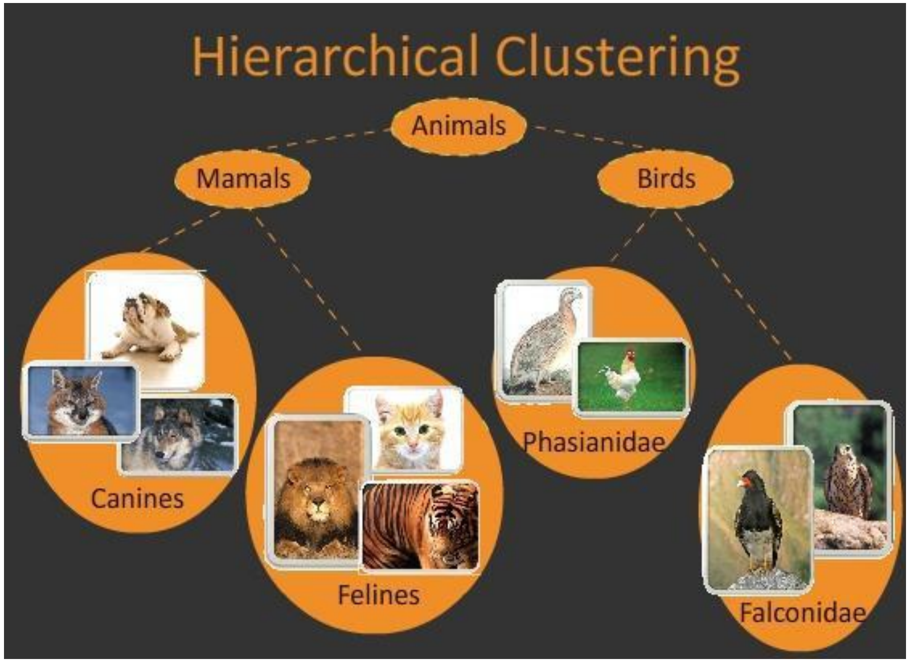


Dendrogram



Cluster number is determined as you like

Partitional Clustering vs. Hierarchical Clustering



Notion of a Cluster can be Ambiguous

How many clusters?

Two Clusters

Four Clusters

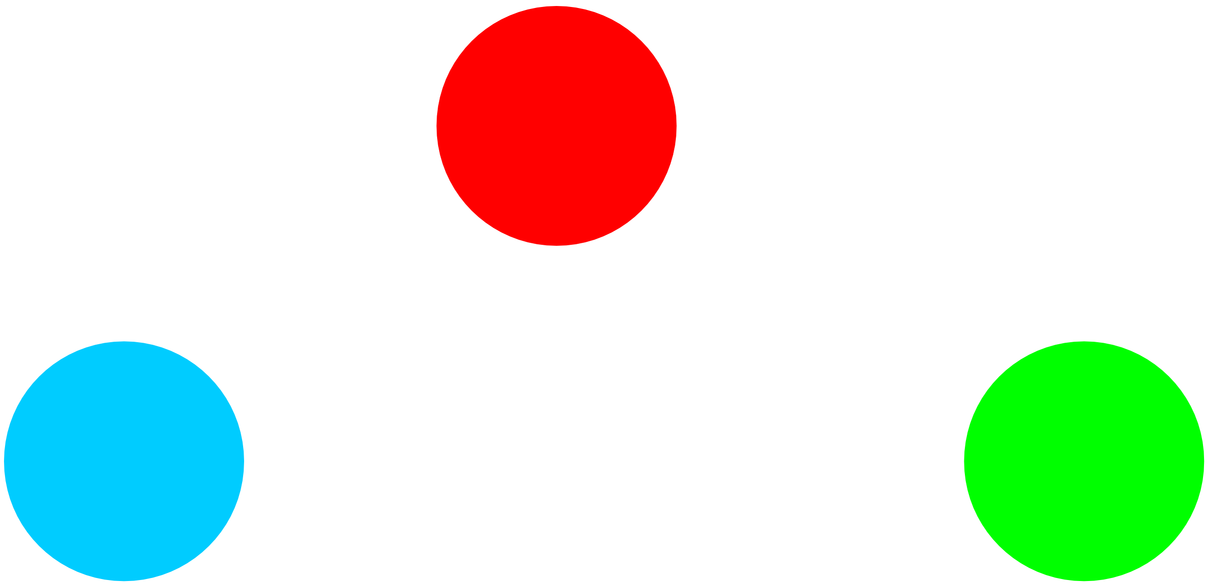
Six Clusters

Types of Clusters

- Well-separated clusters
- Center-based clusters
- Contiguous clusters
- Density-based clusters

Types of Clusters: Well-Separated

- Well-Separated Cluster:
  - 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 Cluster:
  - A cluster is a set of objects such that a point 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 the most "representative" point of a cluster

4 center-based clusters

How many well-separated clusters?

Types of Clusters: Contiguity-Based

- Contiguous Cluster (Nearest neighbor)
  - A cluster is a set of points such that a point in a cluster is closer (or more similar) to **one or more** other points in the cluster than to any point not in the cluster.

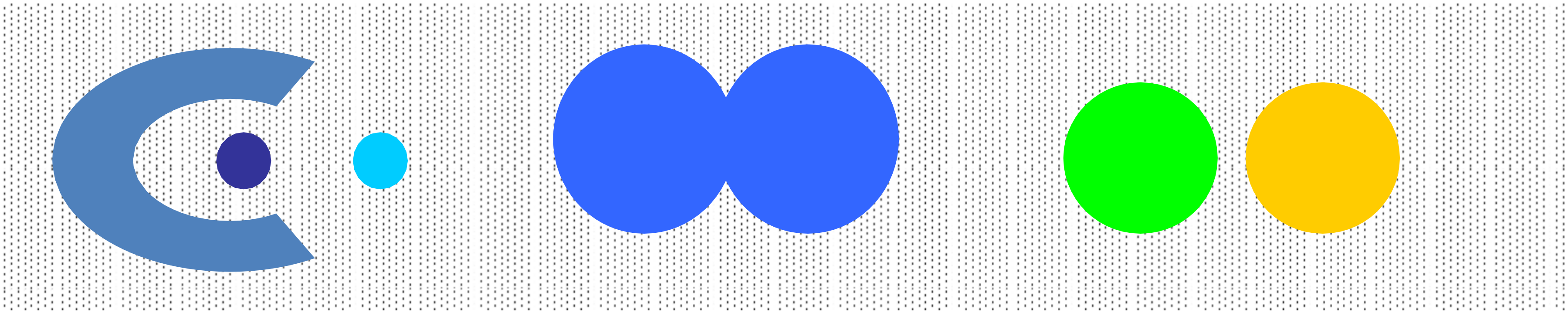
8 contiguous clusters

How many well-separated clusters?

How many center-based clusters?

Types of Clusters: Density-Based

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



6 density-based clusters

Partitional Clustering

K-means

K-means Clustering

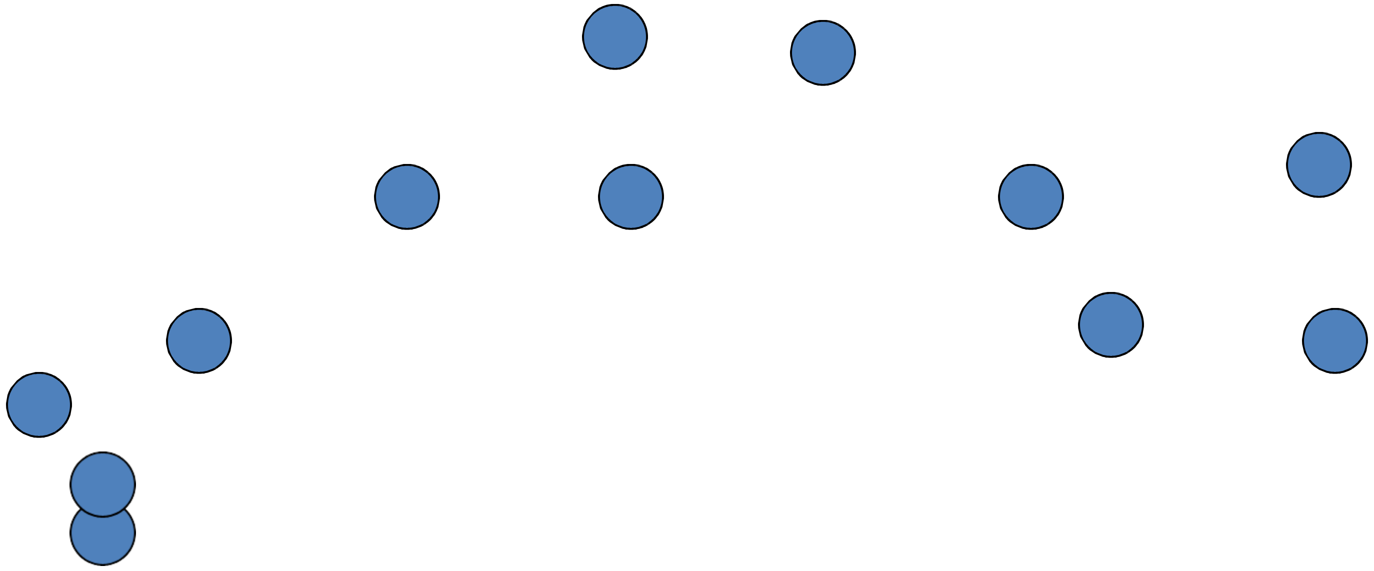
What type of clusters?

- Each cluster is associated with a centroid (center point)
- Each point is assigned to the cluster with the closest centroid

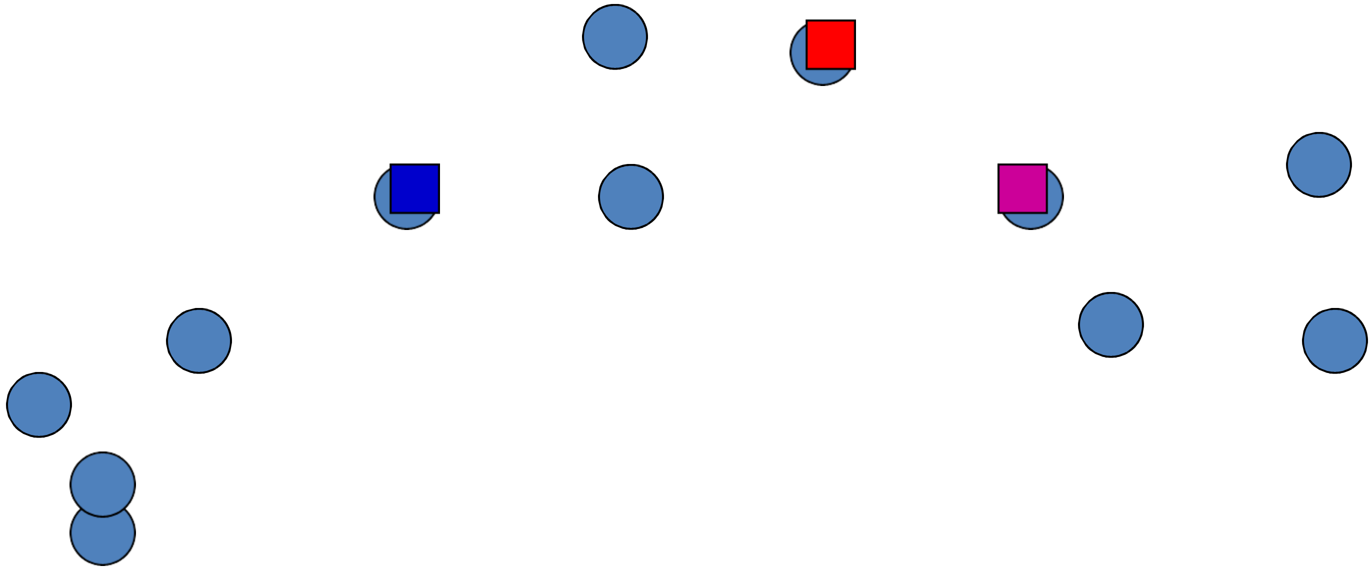
```
Select K points as the initial centroids.
repeat
    Form K clusters by assigning all points to the closest centroid.
    Recompute the centroid of each cluster.
until The centroids don't change
```

- Complexity is O(n * K * d * I)
  - n = number of points, K = number of clusters,
    I = number of iterations, d = number of features

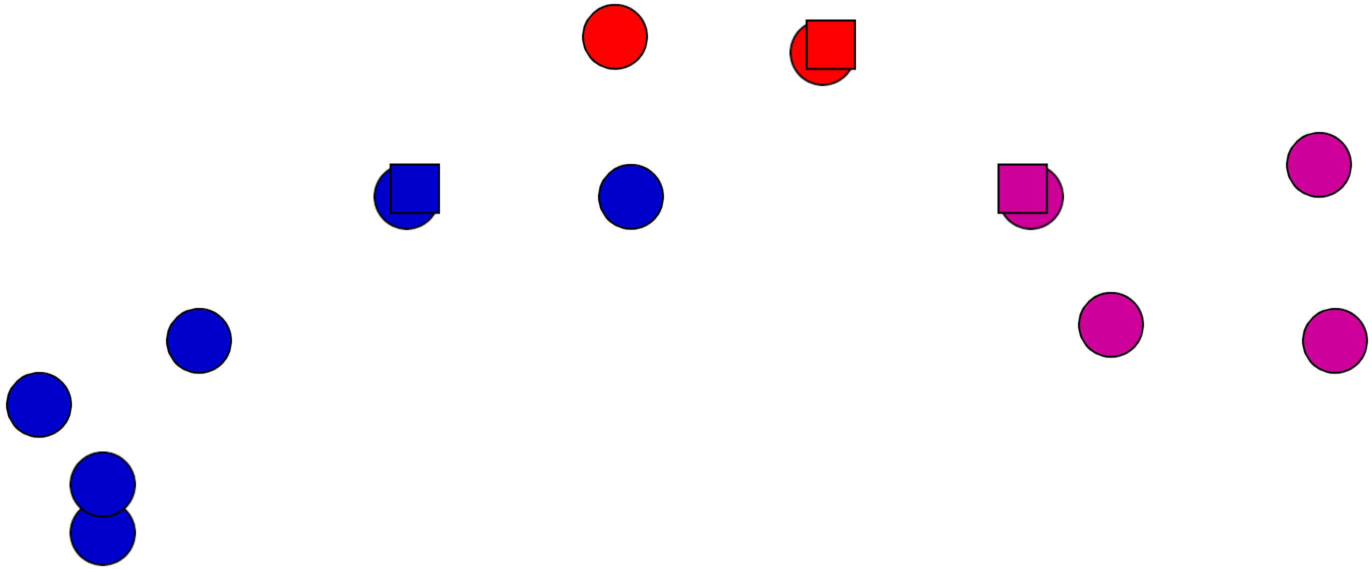
K-means: an example



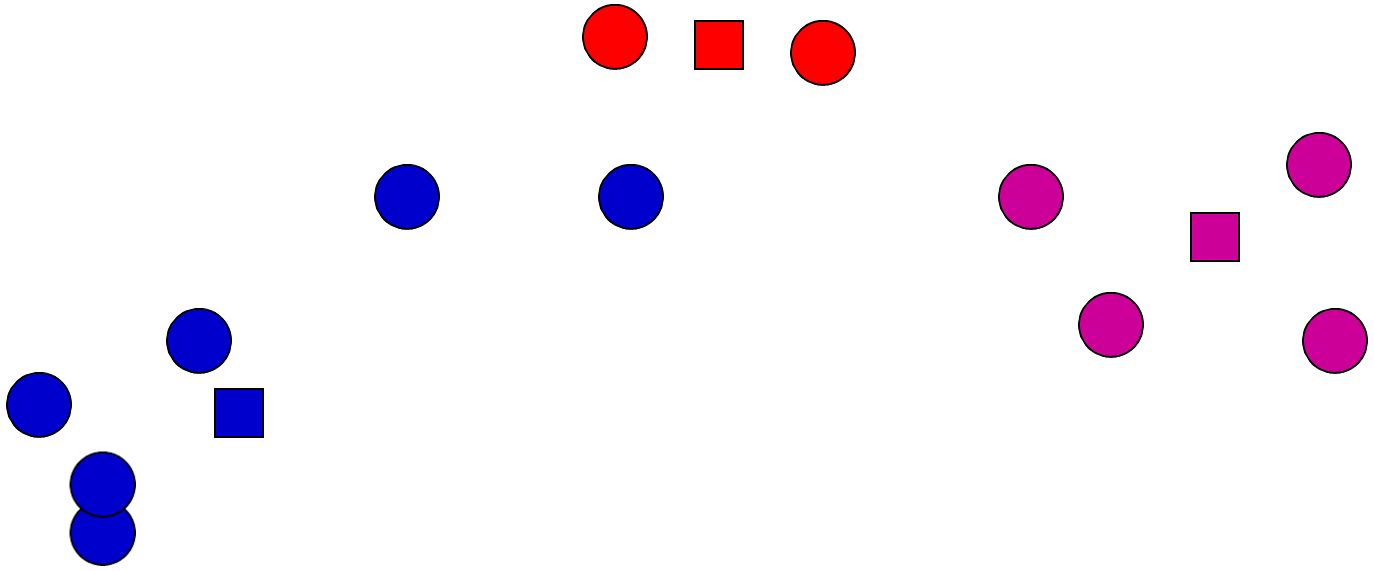
K-means: Initialize centers randomly



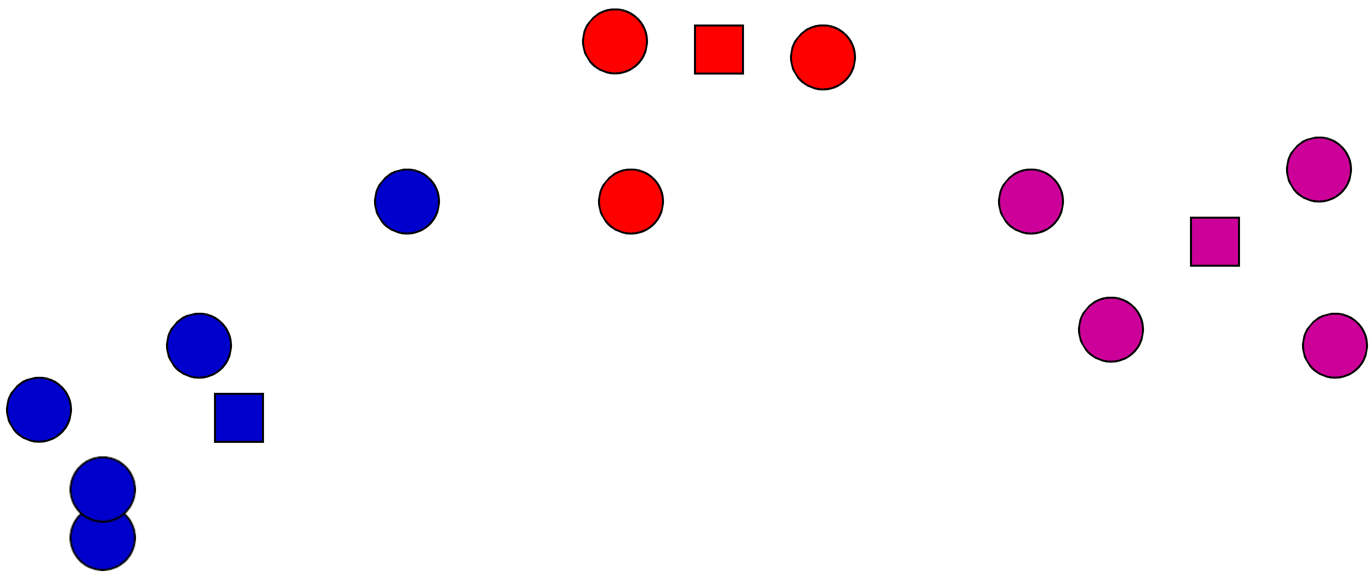
K-means: assign points to nearest center



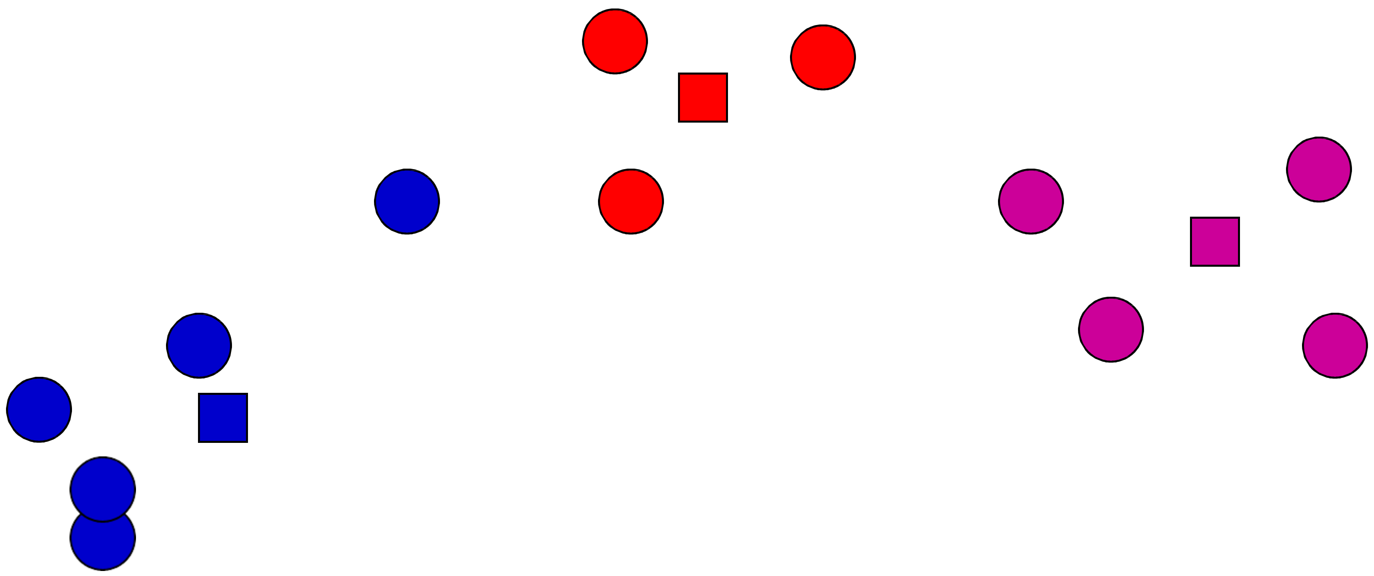
K-means: readjust centers



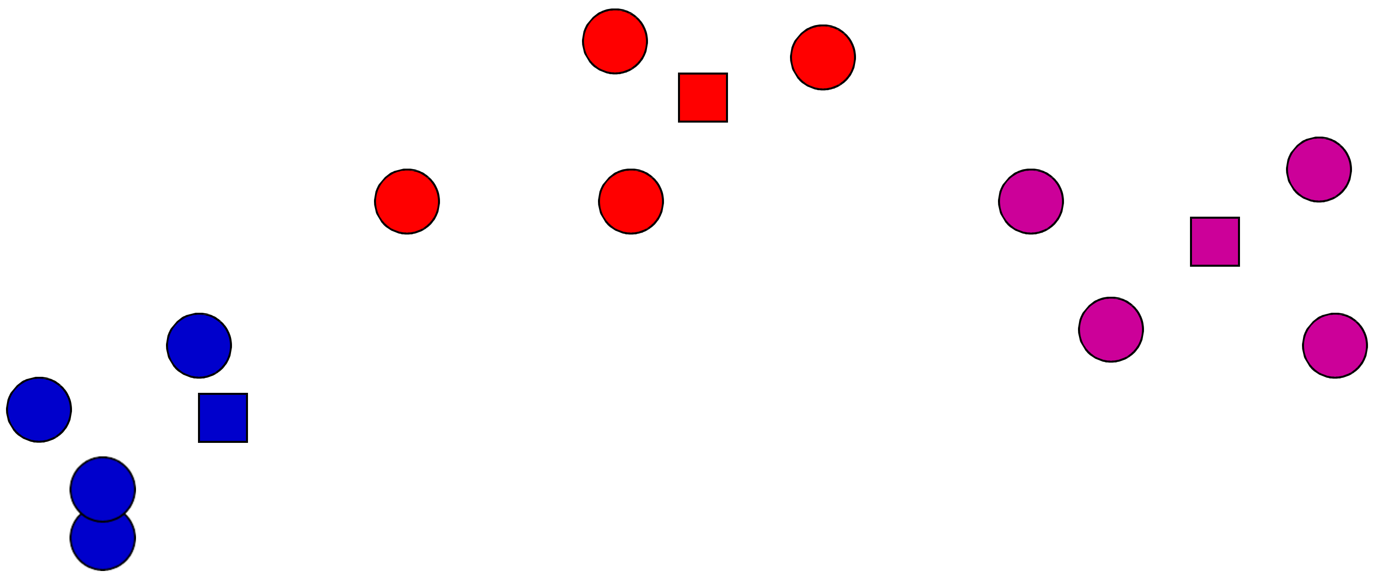
K-means: assign points to nearest center



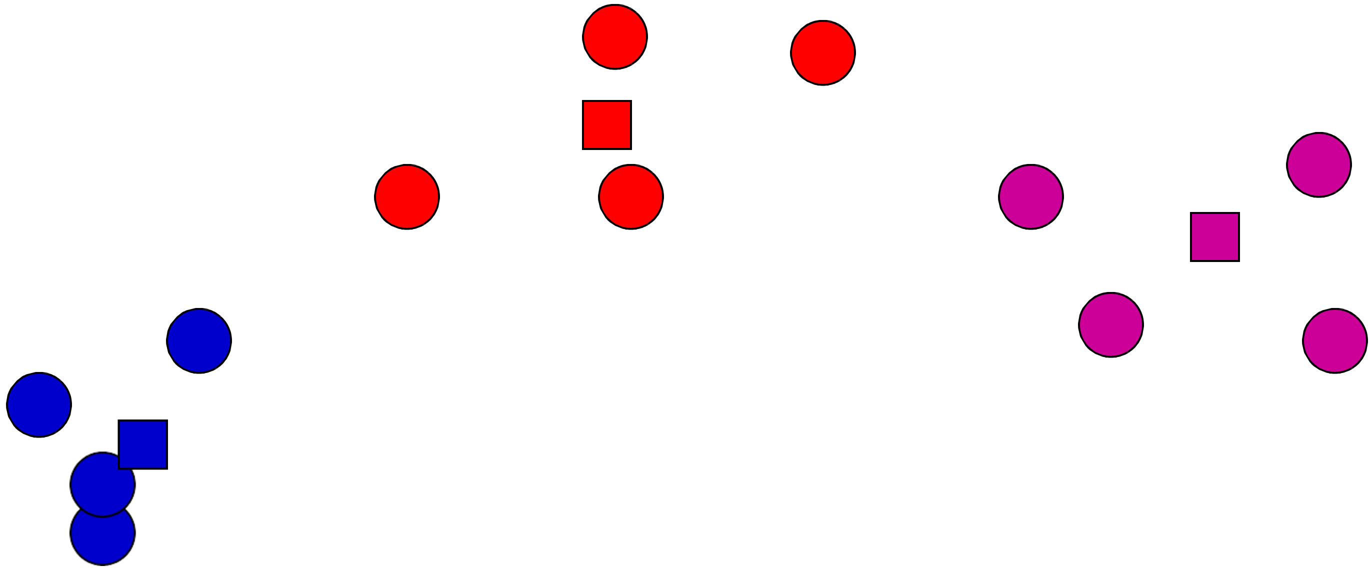
K-means: readjust centers



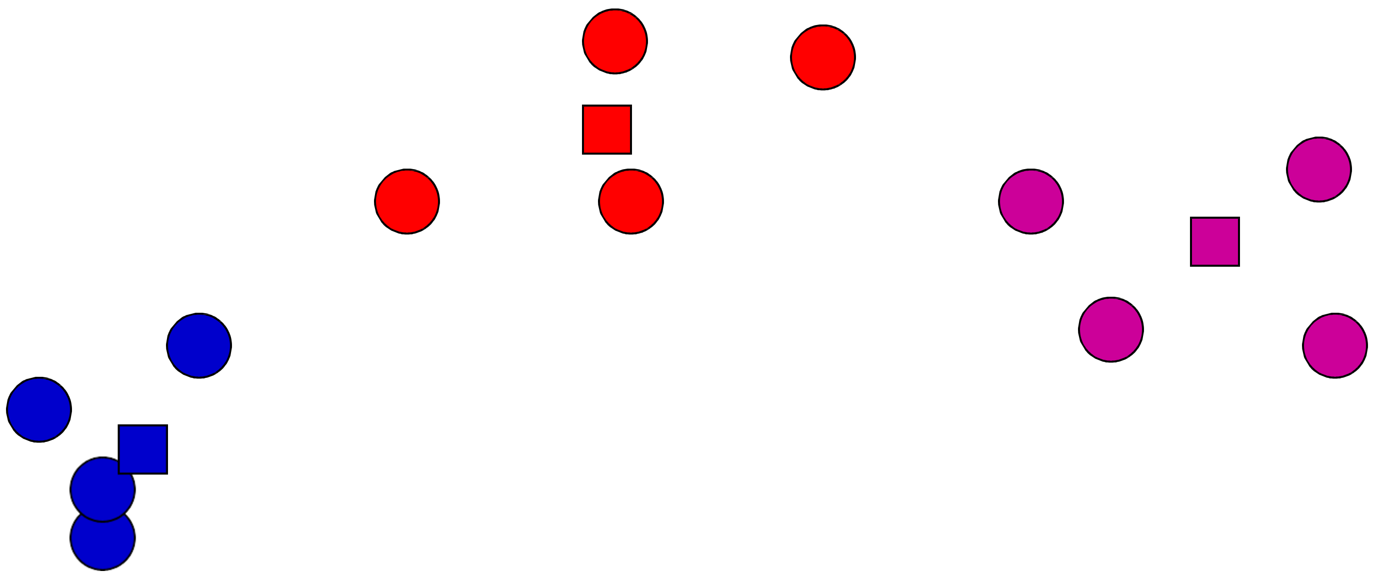
K-means: assign points to nearest center



K-means: readjust centers



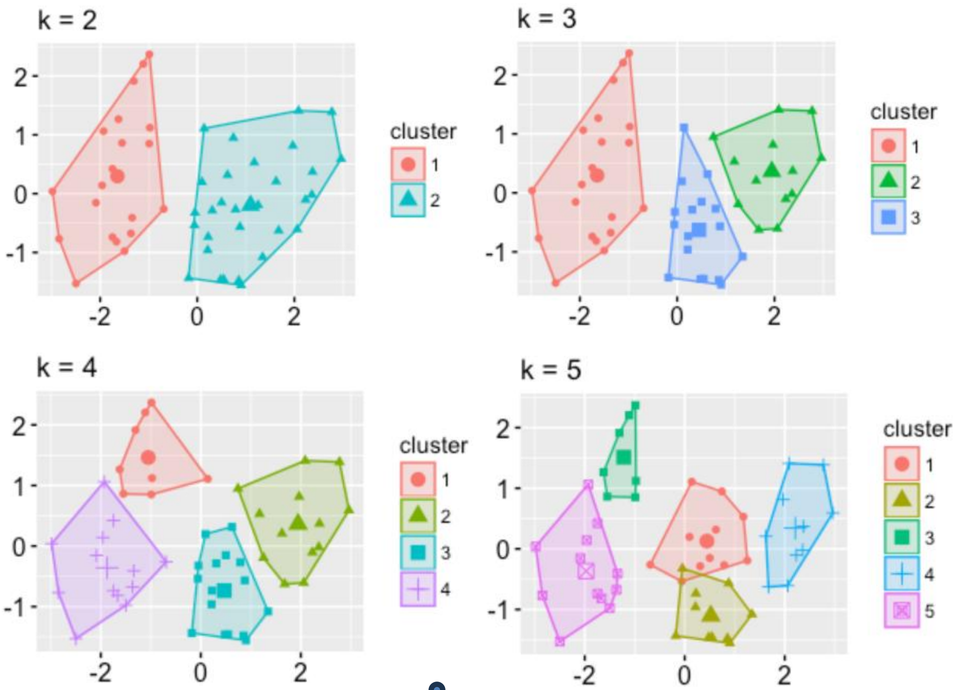
K-means: assign points to nearest center



No changes: Done

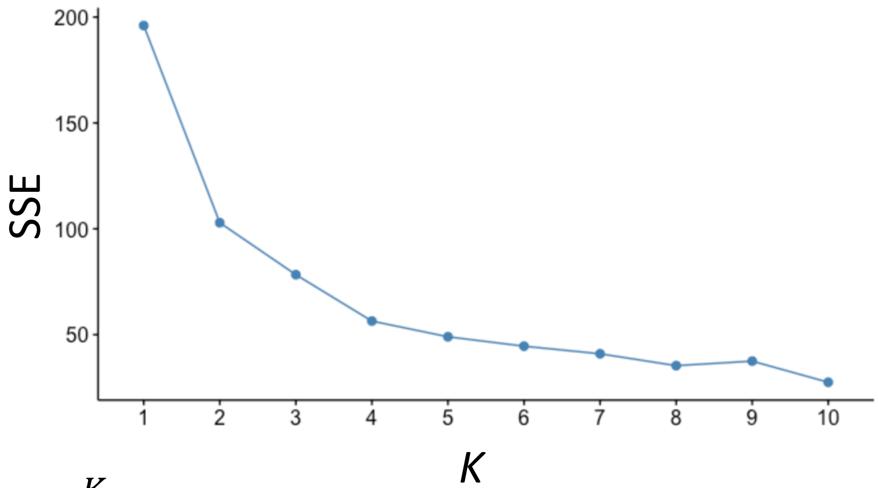
Limitations of K-Means

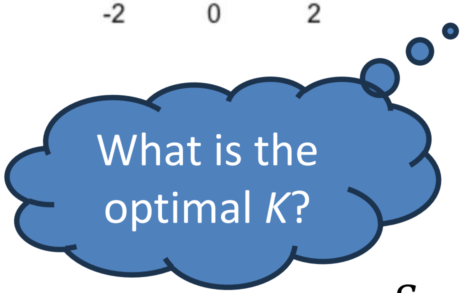
- Need to predefine *K*



Elbow Method:

1. K-means for different *K*
2. For each *K*, calculate SSE
3. Plot the curve of SSE over *K*
4. “bend” or “knee”



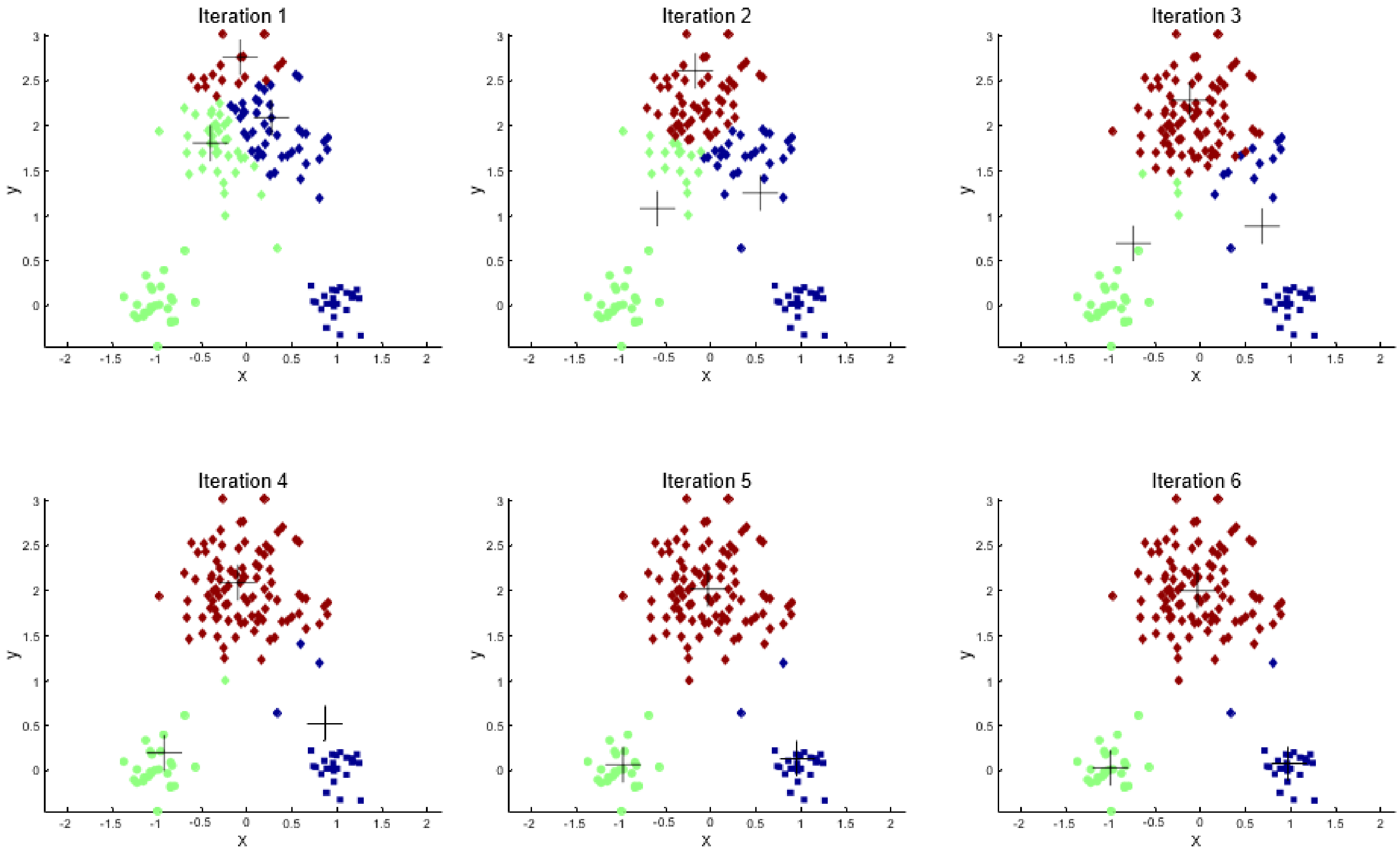


$$Sum\ of\ Squared\ Error\ (SSE) = \sum_{i=1}^{K} \sum_{\boldsymbol{x} \in C_i} dist^2(\boldsymbol{x}, \boldsymbol{\mu}_i)$$

Other solutions: Average Silhouette, Gap Statistic

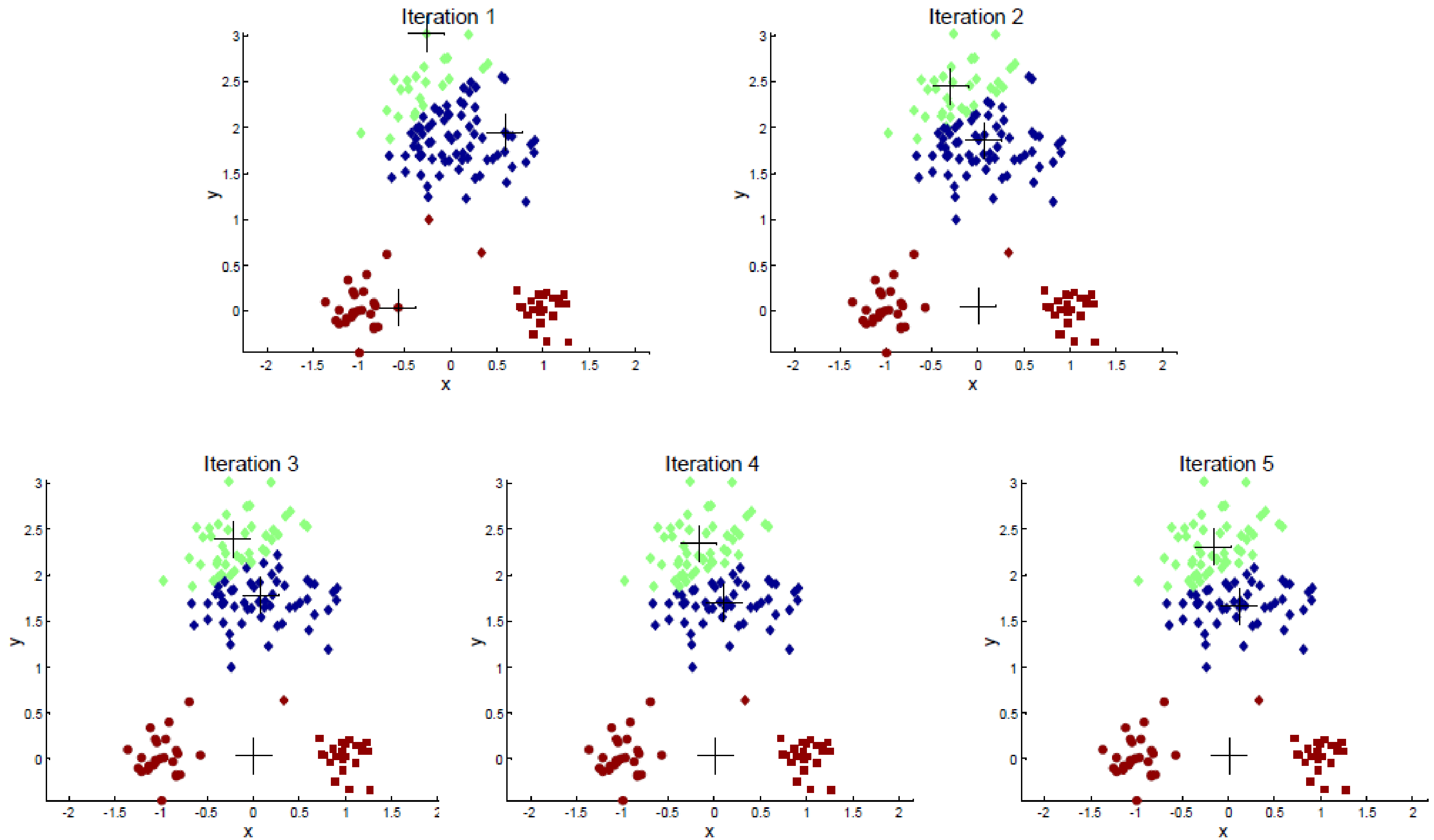
Limitations of K-Means

- Sensitive to initial centroid selection



Limitations of K-Means

- Sensitive to initial centroid selection



Limitations of K-Means

- Sensitive to initial centroid selection

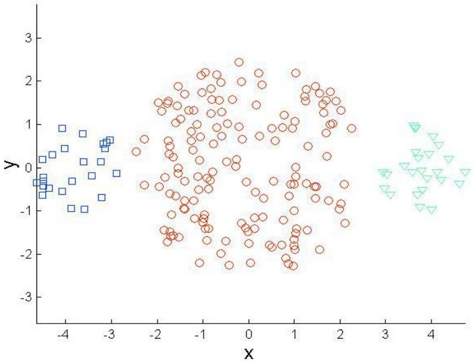
<u>Possible Solutions:</u>

- Multiple runs with different random initializations
  - Helps, but probability is not on your side
- Use hierarchical clustering to determine initial centroids
- Select more than *K* initial centroids and then select among these initial centroids
  - Select most widely separated
- 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

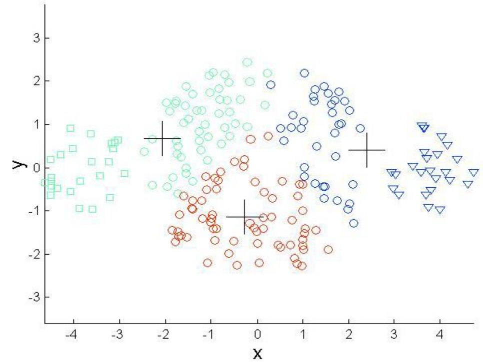
K-Means++: Improved initialization for K-Means

Limitations of K-Means

- Assume spherical, equal-sized, similar-density clusters

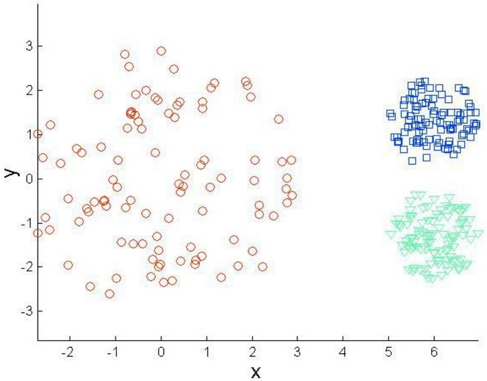


Original Points

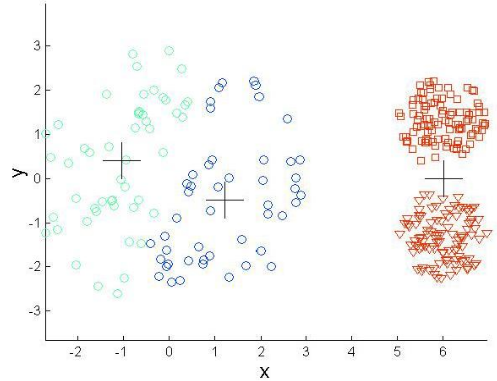


K-means (3 Clusters)

Differ in size!



Original Points



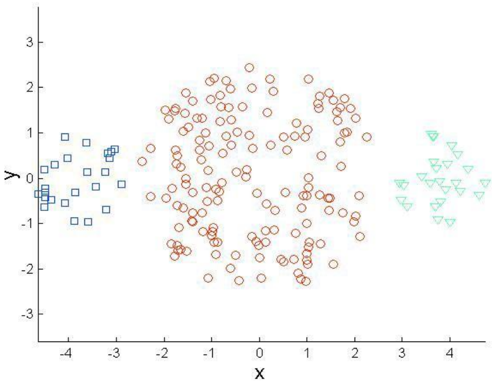
K-means (3 Clusters)

Differ in Density!

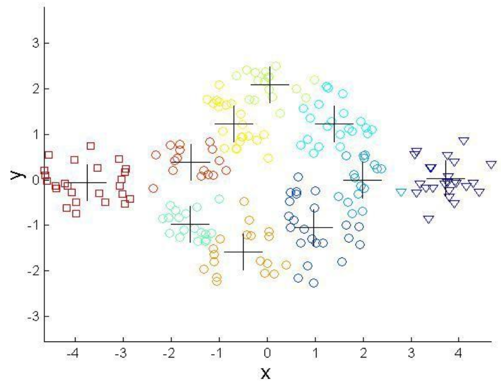
Limitations of K-Means

- Assume spherical, equal-sized, similar-density clusters

  <u>Possible Solution:</u> Use large *K*, then postprocess

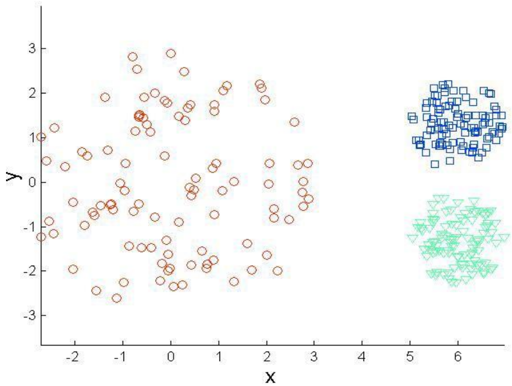


Original Points

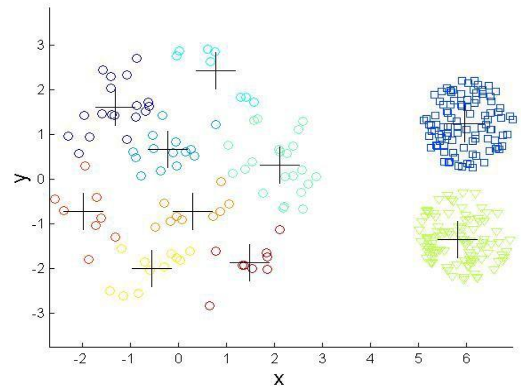


K-means (10 Clusters)

Differ in size



Original Points

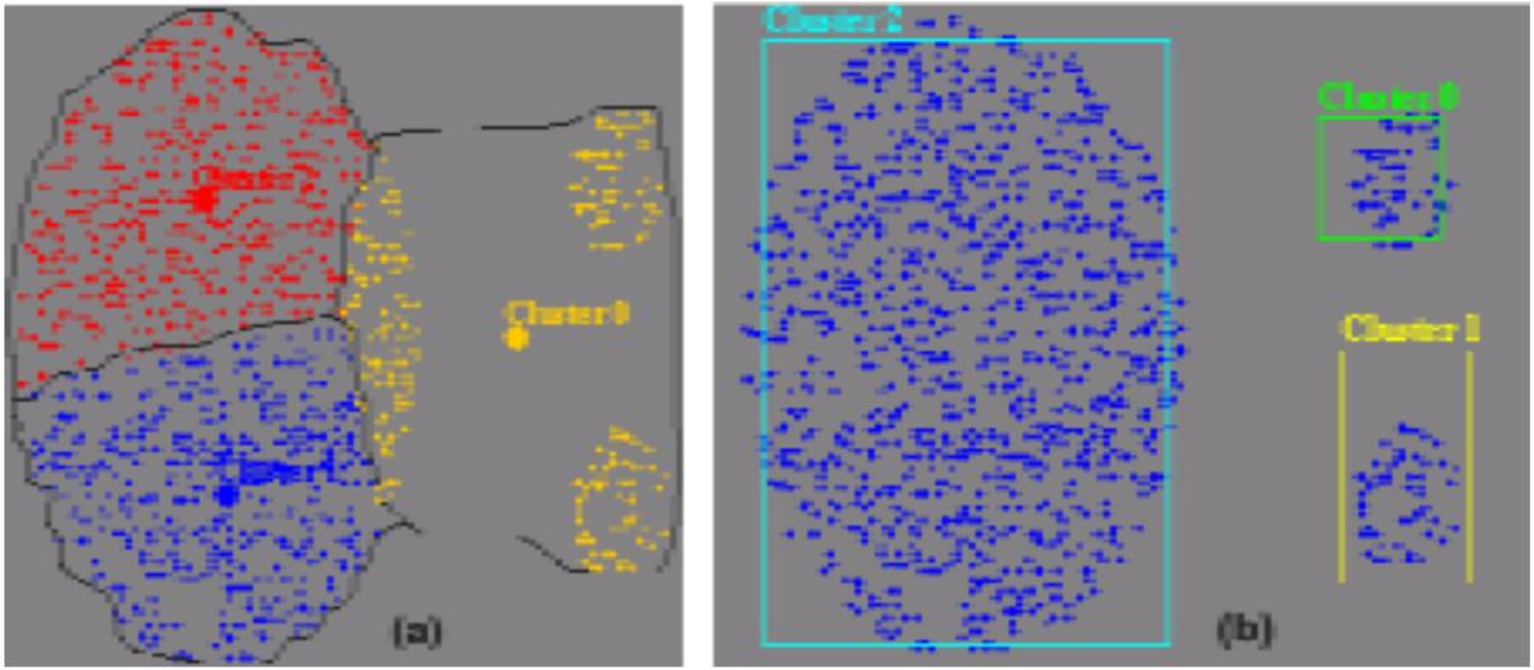


K-means (10 Clusters)

Differ in Density

Limitations of K-Means

- Sensitive to outliers and noisy data



<u>Possible Solution</u>: preprocessing

K-Medoids: Use actual data points (medoids) as cluster centers

Example

Four data points are $x_1 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$, $x_2 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$, $x_3 = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$, $x_4 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$ and the task is to partition the data into two clusters. i) Perform k-means with two initial centroids as $\{x_1, x_2\}$, please list the iterative centroids until converged; and ii) perform k-means with two initial centroids as $\{x_2, x_4\}$, please list the iterative centroids until converged.

Example

Four data points are $x_1 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$, $x_2 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$, $x_3 = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$, $x_4 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$ and the task is to partition the data into two clusters. i) Perform k-means with two initial centroids as $\{x_1, x_2\}$, please list the iterative centroids until converged; and ii) perform k-means with two initial centroids as $\{x_2, x_4\}$, please list the iterative centroids until converged.

(i)

Initialization:

$$\boldsymbol{\mu}_1 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \boldsymbol{\mu}_2 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$$

1st iteration:

$$dist(\boldsymbol{x}_1, \boldsymbol{\mu}_1) = 0, dist(\boldsymbol{x}_1, \boldsymbol{\mu}_2) = 1, \Rightarrow \boldsymbol{x}_1 \epsilon C_1$$

$$dist(\boldsymbol{x}_2, \boldsymbol{\mu}_1) = 1, dist(\boldsymbol{x}_2, \boldsymbol{\mu}_2) = 0, \Rightarrow \boldsymbol{x}_2 \epsilon C_2$$

$$dist(\boldsymbol{x}_3, \boldsymbol{\mu}_1) = 1, dist(\boldsymbol{x}_3, \boldsymbol{\mu}_2) = 1.41, \Rightarrow \boldsymbol{x}_3 \epsilon C_1$$

$$dist(\boldsymbol{x}_4, \boldsymbol{\mu}_1) = 1.41, dist(\boldsymbol{x}_4, \boldsymbol{\mu}_2) = 1, \Rightarrow \boldsymbol{x}_4 \epsilon C_2$$

$$\boldsymbol{\mu}_1 = \frac{1}{2}(\boldsymbol{x}_1 + \boldsymbol{x}_3) = \begin{pmatrix} 0 \\ 0.5 \end{pmatrix}, \boldsymbol{\mu}_2 = \frac{1}{2}(\boldsymbol{x}_2 + \boldsymbol{x}_4) = \begin{pmatrix} 1 \\ 0.5 \end{pmatrix}$$

2nd iteration:

$$dist(\boldsymbol{x}_1, \boldsymbol{\mu}_1) = 0.5, dist(\boldsymbol{x}_1, \boldsymbol{\mu}_2) = 1.12, \Rightarrow \boldsymbol{x}_1 \epsilon C_1$$

$$dist(\boldsymbol{x}_2, \boldsymbol{\mu}_1) = 1.12, dist(\boldsymbol{x}_2, \boldsymbol{\mu}_2) = 0.5, \Rightarrow \boldsymbol{x}_2 \epsilon C_2$$

$$dist(\boldsymbol{x}_3, \boldsymbol{\mu}_1) = 0.5, dist(\boldsymbol{x}_3, \boldsymbol{\mu}_2) = 1.12, \Rightarrow \boldsymbol{x}_3 \epsilon C_1$$

$$dist(\boldsymbol{x}_4, \boldsymbol{\mu}_1) = 1.12, dist(\boldsymbol{x}_4, \boldsymbol{\mu}_2) = 0.5, \Rightarrow \boldsymbol{x}_4 \epsilon C_2$$

Cluster assignment is unchanged, stop!

Example

Four data points are $x_1 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$, $x_2 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$, $x_3 = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$, $x_4 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$ and the task is to partition the data into two clusters. i) Perform k-means with two initial centroids as $\{x_1, x_2\}$, please list the iterative centroids until converged; and ii) perform k-means with two initial centroids as $\{x_2, x_4\}$, please list the iterative centroids until converged.

(ii)

Initialization:

$$\boldsymbol{\mu}_1 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \boldsymbol{\mu}_2 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

1st iteration:

$$dist(\boldsymbol{x}_1, \boldsymbol{\mu}_1) = 1, dist(\boldsymbol{x}_1, \boldsymbol{\mu}_2) = 1.41, \Rightarrow \boldsymbol{x}_1 \epsilon C_1$$

$$dist(\boldsymbol{x}_2, \boldsymbol{\mu}_1) = 0, dist(\boldsymbol{x}_2, \boldsymbol{\mu}_2) = 1, \Rightarrow \boldsymbol{x}_2 \epsilon C_1$$

$$dist(\boldsymbol{x}_3, \boldsymbol{\mu}_1) = 1.41, dist(\boldsymbol{x}_3, \boldsymbol{\mu}_2) = 1, \Rightarrow \boldsymbol{x}_3 \epsilon C_2$$

$$dist(\boldsymbol{x}_4, \boldsymbol{\mu}_1) = 1, dist(\boldsymbol{x}_4, \boldsymbol{\mu}_2) = 0, \Rightarrow \boldsymbol{x}_4 \epsilon C_2$$

$$\boldsymbol{\mu}_1 = \frac{1}{2}(\boldsymbol{x}_1 + \boldsymbol{x}_2) = \begin{pmatrix} 0.5 \\ 0 \end{pmatrix}, \boldsymbol{\mu}_2 = \frac{1}{2}(\boldsymbol{x}_3 + \boldsymbol{x}_4) = \begin{pmatrix} 0.5 \\ 1 \end{pmatrix}$$

2nd iteration:

$$dist(\boldsymbol{x}_1, \boldsymbol{\mu}_1) = 0.5, dist(\boldsymbol{x}_1, \boldsymbol{\mu}_2) = 1.12, \Rightarrow \boldsymbol{x}_1 \epsilon C_1$$

$$dist(\boldsymbol{x}_2, \boldsymbol{\mu}_1) = 0.5, dist(\boldsymbol{x}_2, \boldsymbol{\mu}_2) = 1.12, \Rightarrow \boldsymbol{x}_2 \epsilon C_1$$

$$dist(\boldsymbol{x}_3, \boldsymbol{\mu}_1) = 1.12, dist(\boldsymbol{x}_3, \boldsymbol{\mu}_2) = 0.5, \Rightarrow \boldsymbol{x}_3 \epsilon C_2$$

$$dist(\boldsymbol{x}_4, \boldsymbol{\mu}_1) = 1.12, dist(\boldsymbol{x}_4, \boldsymbol{\mu}_2) = 0.5, \Rightarrow \boldsymbol{x}_4 \epsilon C_2$$

Cluster assignment is unchanged, stop!

Hierarchical Clustering

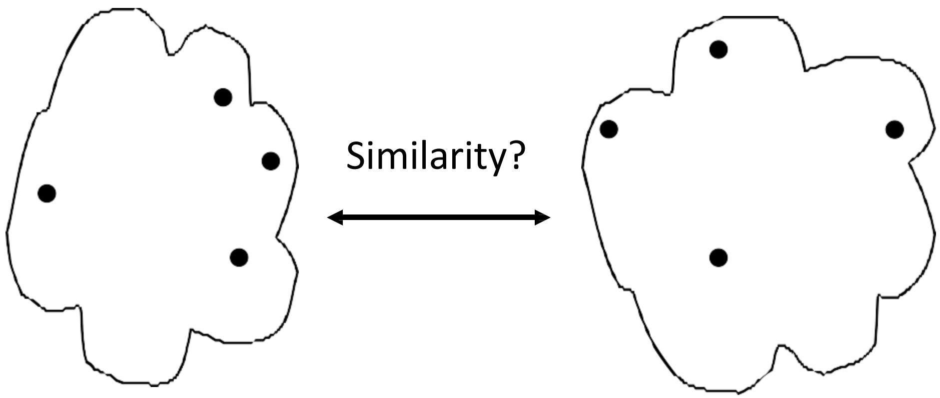
Agglomerative

Agglomerative Clustering Algorithm

- More popular hierarchical clustering technique
- Basic algorithm is straightforward
  1. Let each data point be a cluster
  2. Compute the proximity/distance matrix
  3. **Repeat**
  4. Merge the two closest clusters
  5. Update the proximity/distance matrix
  6. **Until** only a single cluster remains

- $O(N^2)$ space since it uses the proximity matrix.
  – N is the number of points.
- $O(N^3)$ time
  – There are N-1 steps and at each step, the proximity matrix of size N^2 must be updated and searched

How to Define Inter-Cluster Similarity

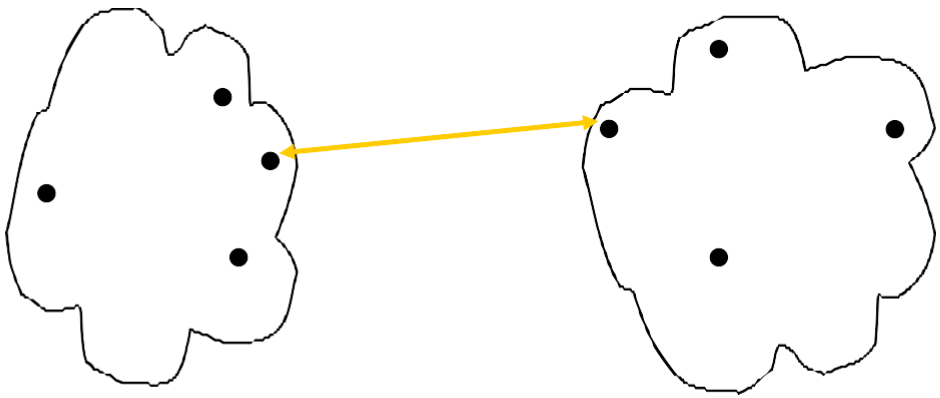


	p1	p2	p3	p4	p5	. . .
p1						
p2						
p3						
p4						
p5						
.						
.						
.						

Proximity Matrix

- MIN
- MAX
- Group Average
- Distance Between Centroids

How to Define Inter-Cluster Similarity



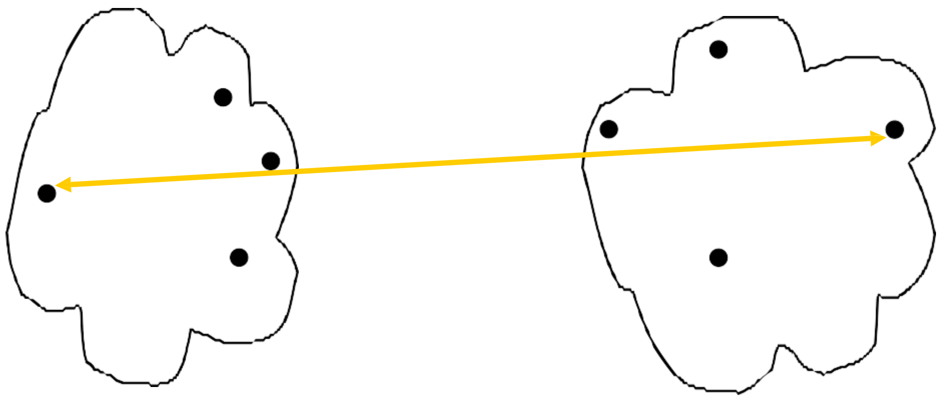
	p1	p2	p3	p4	p5	. . .
p1						
p2						
p3						
p4						
p5						
.						
.						
.						

Proximity Matrix

- MIN
- MAX
- Group Average
- Distance Between Centroids

What type of clusters?

How to Define Inter-Cluster Similarity



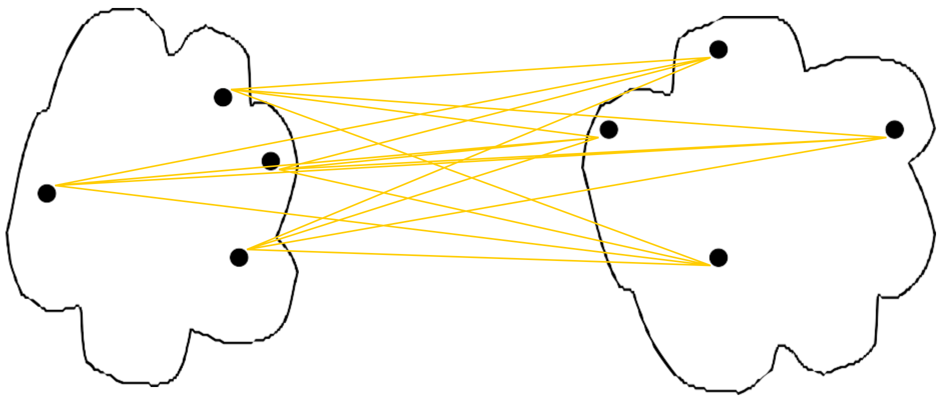
	p1	p2	p3	p4	p5	. . .
p1						
p2						
p3						
p4						
p5						
.						
.						
.						

Proximity Matrix

- MIN
- MAX
- Group Average
- Distance Between Centroids

What type of clusters?

How to Define Inter-Cluster Similarity



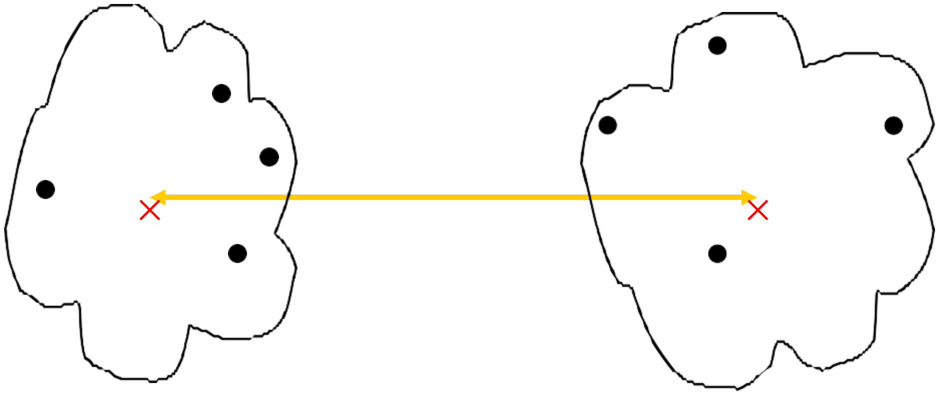
	p1	p2	p3	p4	p5	. . .
p1						
p2						
p3						
p4						
p5						
.						
.						
.						

Proximity Matrix

- MIN
- MAX
- Group Average
- Distance Between Centroids

What type of clusters?

How to Define Inter-Cluster Similarity



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

Proximity Matrix

- MIN
- MAX
- Group Average
- Distance Between Centroids

What type of clusters?

Cluster Similarity: MIN or Single Link

- Similarity of two clusters is based on the two most similar (closest) points in the different clusters
  - Determined by one pair of points

	I1	I2	I3	I4	I5
I1	1.00	0.90	0.10	0.65	0.20
I2		1.00	0.70	0.60	0.50
I3			1.00	0.40	0.30
I4				1.00	0.80
I5					1.00

Proximity (similarity) Matrix

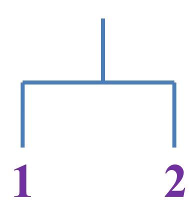
Cluster Similarity: MIN or Single Link

- Similarity of two clusters is based on the two most similar (closest) points in the different clusters
  - Determined by one pair of points

	I1	I2	I3	I4	I5
I1	1.00	0.90	0.10	0.65	0.20
I2		1.00	0.70	0.60	0.50
I3			1.00	0.40	0.30
I4				1.00	0.80
I5					1.00

Proximity (similarity) Matrix

	(I1,I2)	I3	I4	I5
(I1,I2)	1.00	0.70	0.65	0.50
I3		1.00	0.40	0.30
I4			1.00	0.80
I5				1.00

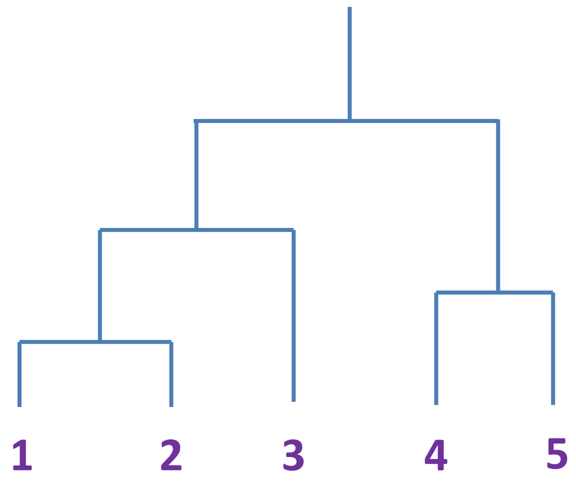


Cluster Similarity: MIN or Single Link

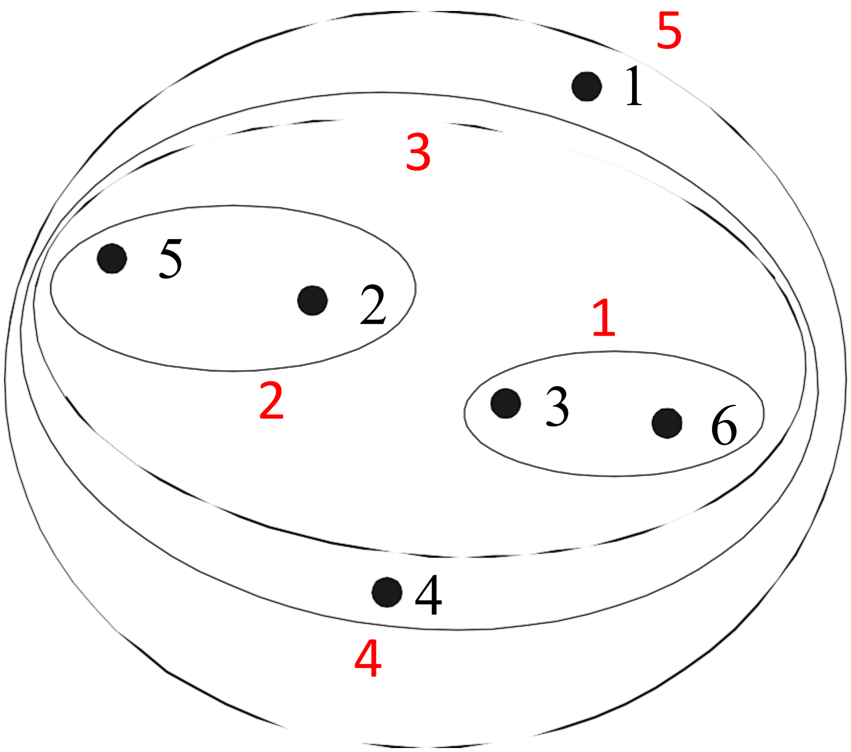
	(I1,I2)	I3	I4	I5
(I1,I2)	1.00	0.70	0.65	0.50
I3		1.00	0.40	0.30
I4			1.00	0.80
I5				1.00

	(I1,I2)	I3	(I4,I5)
(I1,I2)	1.00	0.70	0.65
I3		1.00	0.40
(I4,I5)			1.00

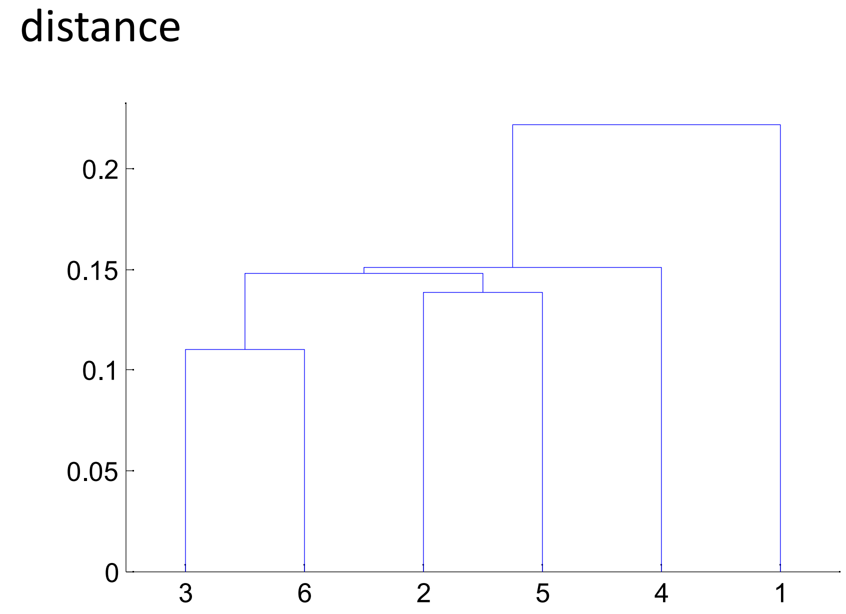
	(I1,I2,I3)	(I4,I5)
(I1,I2,I3)	1.00	0.65
(I4,I5)		1.00



Agglomerative Clustering: MIN



Nested Clusters



Dendrogram

Cluster Similarity: MAX or Complete Linkage

- Similarity of two clusters is based on the two least similar (most distant) points in the different clusters
  - Determined by all pairs of points in the two clusters

	I1	I2	I3	I4	I5
I1	1.00	0.90	0.10	0.65	0.20
I2		1.00	0.70	0.60	0.50
I3			1.00	0.40	0.30
I4				1.00	0.80
I5					1.00

Proximity (Similarity) Matrix

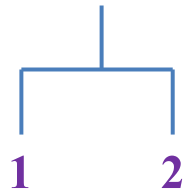
Cluster Similarity: MAX or Complete Linkage

- Similarity of two clusters is based on the two least similar (most distant) points in the different clusters
  - Determined by all pairs of points in the two clusters

	I1	I2	I3	I4	I5
I1	1.00	0.90	0.10	0.65	0.20
I2		1.00	0.70	0.60	0.50
I3			1.00	0.40	0.30
I4				1.00	0.80
I5					1.00

Proximity (Similarity) Matrix

	(I1,I2)	I3	I4	I5
(I1,I2)	1.00	0.10	0.60	0.20
I3		1.00	0.40	0.30
I4			1.00	0.80
I5				1.00

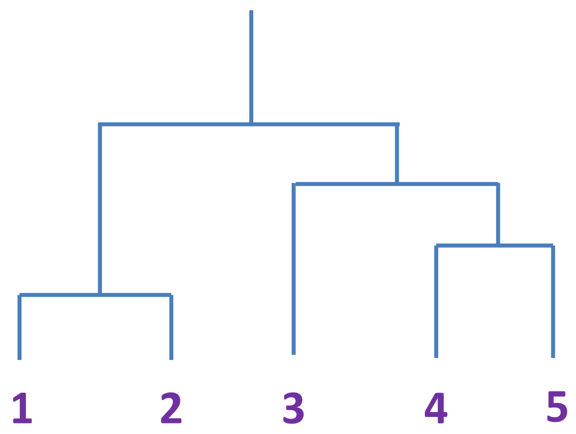


Cluster Similarity: MAX or Complete Linkage

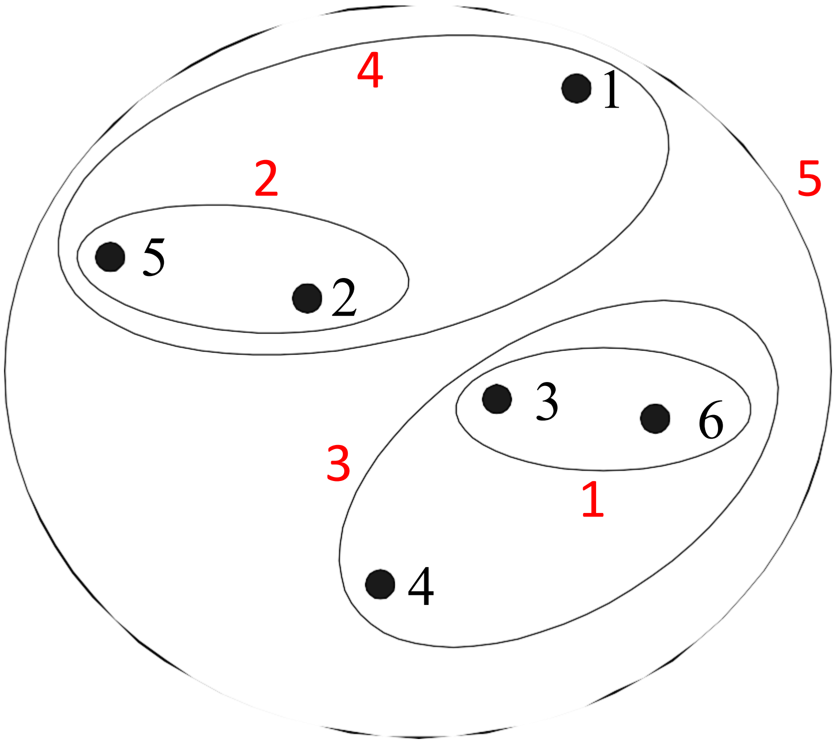
	(I1,I2)	I3	I4	I5
(I1,I2)	1.00	0.10	0.60	0.20
I3		1.00	0.40	0.30
I4			1.00	0.80
I5				1.00

	(I1,I2)	I3	(I4,I5)
(I1,I2)	1.00	0.10	0.20
I3		1.00	0.30
(I4,I5)			1.00

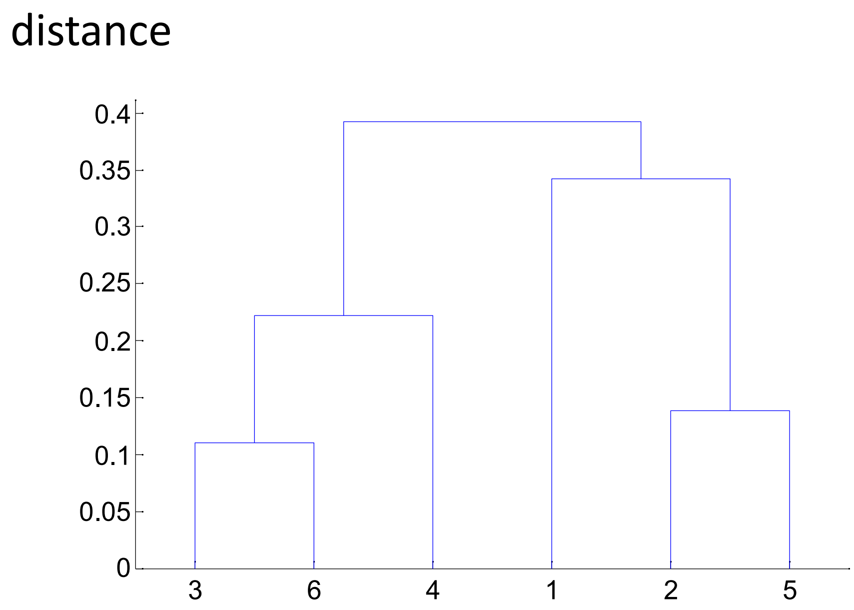
	(I1,I2)	(I3,I4,I5)
(I1,I2)	1.00	0.10
(I3,I4,I5)		1.00



Agglomerative Clustering: MAX



Nested Clusters



Dendrogram

Cluster Similarity: Group Average

- Proximity of two clusters is the average of pairwise proximity between points in the two clusters.

$$\mathbf{proximity(Cluster_i, Cluster_j)} = \frac{\sum\limits_{\substack{p_i \in Cluster_i \\ p_j \in Cluster_j}} \mathbf{proximity(p_i, p_j)}}{|\mathbf{Cluster_i}| * |\mathbf{Cluster_j}|}$$

	I1	I2	I3	I4	I5
I1	1.00	0.90	0.10	0.65	0.20
I2		1.00	0.70	0.60	0.50
I3			1.00	0.40	0.30
I4				1.00	0.80
I5					1.00

Proximity (Similarity) Matrix

Cluster Similarity: Group Average

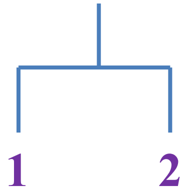
- Proximity of two clusters is the average of pairwise proximity between points in the two clusters.

$$\text{proximity}(\text{Cluster}_i, \text{Cluster}_j) = \frac{\sum\limits_{\substack{p_i \in \text{Cluster}_i \\ p_j \in \text{Cluster}_j}} \text{proximity}(p_i, p_j)}{|\text{Cluster}_i| * |\text{Cluster}_j|}$$

	I1	I2	I3	I4	I5
I1	1.00	0.90	0.10	0.65	0.20
I2		1.00	0.70	0.60	0.50
I3			1.00	0.40	0.30
I4				1.00	0.80
I5					1.00

Proximity (Similarity) Matrix

	(I1,I2)	I3	I4	I5
(I1,I2)	1.00	0.40	0.625	0.35
I3		1.00	0.40	0.30
I4			1.00	0.80
I5				1.00

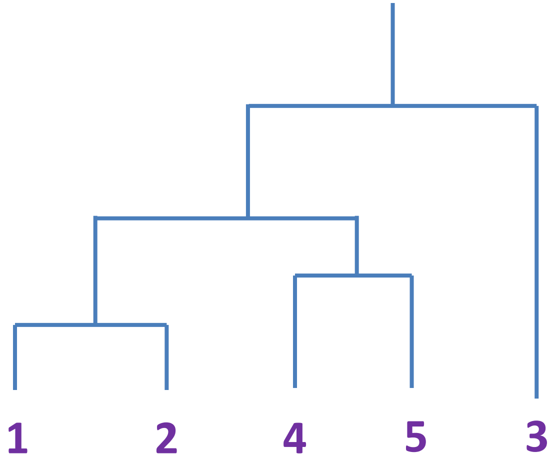


Cluster Similarity: Group Average

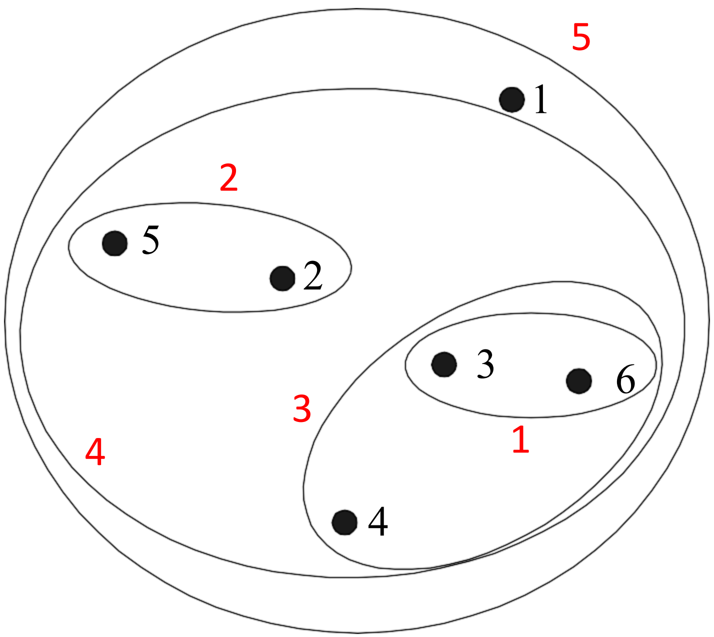
	(I1,I2)	I3	I4	I5
(I1,I2)	1.00	0.40	0.625	0.35
I3		1.00	0.40	0.30
I4			1.00	0.80
I5				1.00

	(I1,I2)	I3	(I4,I5)
(I1,I2)	1.00	0.40	0.4875
I3		1.00	0.35
(I4,I5)			1.00

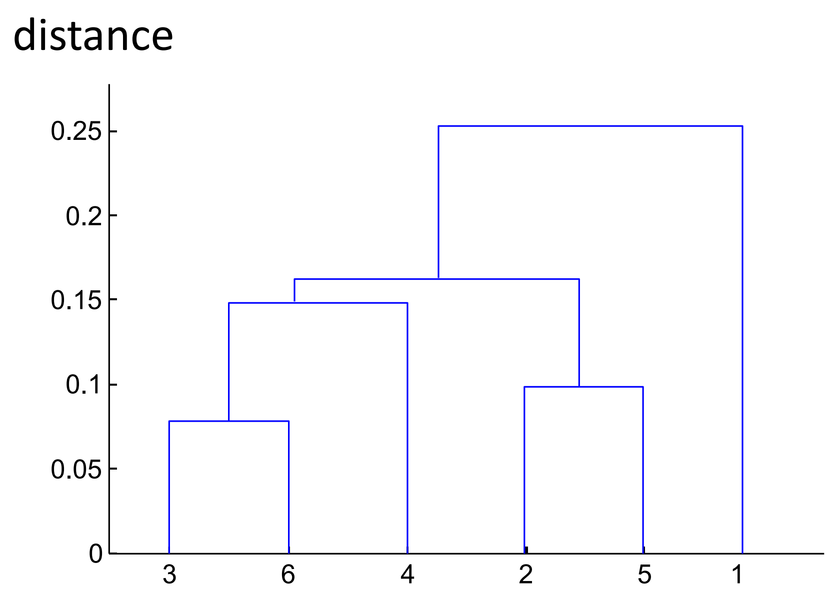
	(I1,I2,I4,I5)	I3
(I1,I2,I4,I5)	1.00	0.375
I3		1.00



Agglomerative Clustering: Group Average



Nested Clusters



Dendrogram

Cluster Similarity: Distance Between Centroids

- Centroid linkage

Suppose I1=(1,1), I2=(1,2), I3=(4,4), I4=(5,4), I5=(5,5)

	I1	I2	I3	I4	I5
I1	0.00	1.00	4.24	5.00	5.66
I2		0.00	3.61	4.47	5.00
I3			0.00	1.00	1.41
I4				0.00	1.00
I5					0.00

Distance Matrix

Cluster Similarity: Distance Between Centroids

- Centroid linkage

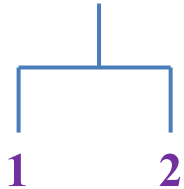
Suppose I1=(1,1), I2=(1,2), I3=(4,4), I4=(5,4), I5=(5,5)

	I1	I2	I3	I4	I5
I1	0.00	1.00	4.24	5.00	5.66
I2		0.00	3.61	4.47	5.00
I3			0.00	1.00	1.41
I4				0.00	1.00
I5					0.00

Distance Matrix

Centroid of (I1, I2)=(1,1.5)

	(I1,I2)	I3	I4	I5
(I1,I2)	0.00	3.91	4.72	5.32
I3		0.00	1.00	1.41
I4			0.00	1.00
I5				0.00



Cluster Similarity: Distance Between Centroids

Suppose I1=(1,1), I2=(1,2), I3=(4,4), I4=(5,4), I5=(5,5)

Centroid of (I1, I2)=(1,1.5)

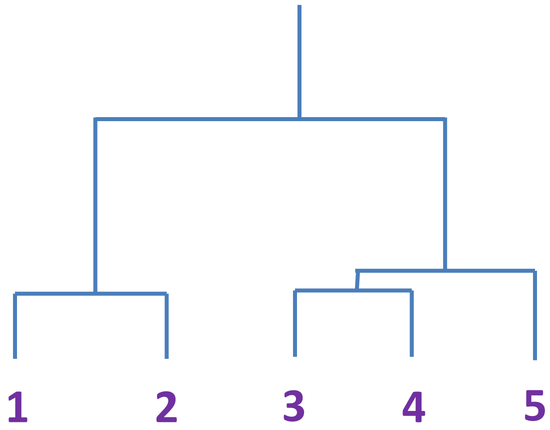
	(I1,I2)	I3	I4	I5
(I1,I2)	0.00	3.91	4.72	5.32
I3		0.00	1.00	1.41
I4			0.00	1.00
I5				0.00

Centroid of (I3, I4)=(4.5,4)

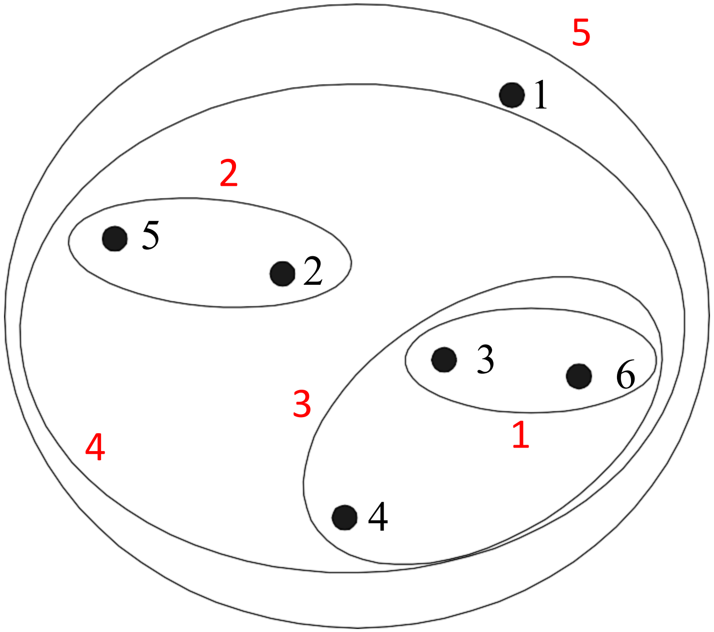
	(I1,I2)	(I3,I4)	I5
(I1,I2)	0.00	4.30	5.32
(I3,I4)		0.00	1.12
I5			0.00

Centroid of (I3, I4, I5)=(4.67,4.33)

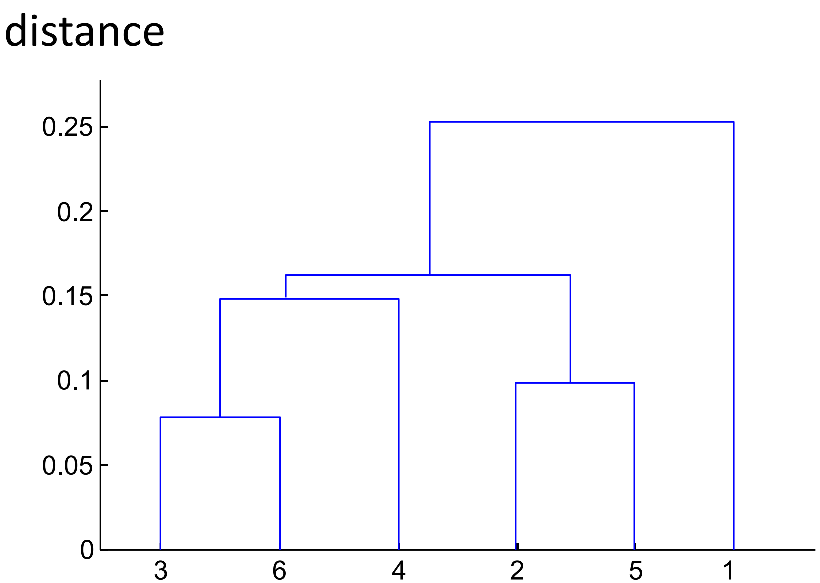
	(I1,I2)	(I3,I4,I5)
(I1,I2)	0.00	4.63
(I3,I4,I5)		0.00



Agglomerative Clustering: Centroid



Nested Clusters



Dendrogram

Limitations of Agglomerative Clustering

- No objective function is directly minimized
- Different schemes have problems with one or more of the following:
  - Sensitivity to noise and outliers
  - Difficulty handling different-sized clusters

Linkage	Cluster Shape	Sensitivity to Noise/Outliers	Handling Different-sized clusters
Single	elongated	low	poor
Complete	compact, spherical	high	moderate
Average	moderately compact	moderate	moderate
Centroid	spherical	high	poor

Measures of Cluster Validity

- Numerical measures that are applied to judge various aspects of cluster validity, are classified into the following **two** types.
  - External Index: Used to measure the extent to which cluster labels match externally supplied class labels.
    - Entropy: $-\sum_{i=1}^{c} p_i \times log_2(p_i)$
    - Accuracy with major class label as cluster label
  - Internal Index: Used to measure the goodness of a clustering structure *without* respect to external information.
    - Sum of Squared Error (SSE)

Example

6 data points are

$$x_1 = \begin{pmatrix} 1.5 \\ 4 \end{pmatrix}, x_2 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, x_3 = \begin{pmatrix} 1 \\ -1 \end{pmatrix}, x_4 = \begin{pmatrix} 0.5 \\ -2.1 \end{pmatrix}, x_5 = \begin{pmatrix} -1 \\ 0.9 \end{pmatrix}, x_6 = \begin{pmatrix} 2 \\ -1.5 \end{pmatrix}$$

Please list the nested clusters by using agglomerative clustering based on MIN, MAX, Group Average, and Centroid methods for inter-cluster Euclidean distance measure respectively. Note that, for each method, list all the intermediate Euclidean distance matrices.

Example

6 data points are

$$x_1 = \begin{pmatrix} 1.5 \\ 4 \end{pmatrix}, x_2 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, x_3 = \begin{pmatrix} 1 \\ -1 \end{pmatrix}, x_4 = \begin{pmatrix} 0.5 \\ -2.1 \end{pmatrix}, x_5 = \begin{pmatrix} -1 \\ 0.9 \end{pmatrix}, x_6 = \begin{pmatrix} 2 \\ -1.5 \end{pmatrix}$$

Please list the nested clusters by using agglomerative clustering based on MIN, MAX, Group Average, and Centroid methods for inter-cluster Euclidean distance measure respectively. Note that, for each method, list all the intermediate Euclidean distance matrices.

	x1	x2	x3	x4	x5	x6
x1	0	4.27	5.02	6.18	3.98	5.52
x2		0	1.41	2.16	1.35	2.5
x3			0	1.21	2.76	1.12
x4				0	3.35	1.62
x5					0	3.84
x6						0

Example

MIN:

	x1	x2	x3	x4	x5	x6
x1	0	4.27	5.02	6.18	3.98	5.52
x2		0	1.41	2.16	1.35	2.5
x3			0	1.21	2.76	1.12
x4				0	3.35	1.62
x5					0	3.84
x6						0

	x1	x2	(x3,x6)	x4	x5
x1	0	4.27	5.02	6.18	3.98
x2		0	1.41	2.16	1.35
(x3,x6)			0	1.21	2.76
x4				0	3.35
x5					0

	x1	x2	(x3,x4,x6)	x5
x1	0	4.27	5.02	3.98
x2		0	1.41	1.35
(x3,x4,x6)			0	2.76
x5				0

	x1	(x2,x5)	(x3,x4,x6)
x1	0	3.98	5.02
(x2,x5)		0	1.41
(x3,x4,x6)			0

	x1	(x2,x3,x4,x5,x6)
x1	0	3.98
(x2,x3,x4,x5,x6)		0

Example

MAX:

	x1	x2	x3	x4	x5	x6
x1	0	4.27	5.02	6.18	3.98	5.52
x2		0	1.41	2.16	1.35	2.5
x3			0	1.21	2.76	1.12
x4				0	3.35	1.62
x5					0	3.84
x6						0

	x1	x2	(x3,x6)	x4	x5
x1	0	4.27	5.52	6.18	3.98
x2		0	2.5	2.16	1.35
(x3,x6)			0	1.62	3.84
x4				0	3.35
x5					0

	x1	(x2,x5)	(x3,x6)	x4
x1	0	4.27	5.52	6.18
(x2,x5)		0	3.84	3.35
(x3,x6)			0	1.62
x4				0

	x1	(x2,x5)	(x3,x4,x6)
x1	0	4.27	6.18
(x2,x5)		0	3.84
(x3,x4,x6)			0

	x1	(x2,x3,x4,x5,x6)
x1	0	6.18
(x2,x3,x4,x5,x6)		0

Example

Group Average:

	x1	x2	x3	x4	x5	x6
x1	0	4.27	5.02	6.18	3.98	5.52
x2		0	1.41	2.16	1.35	2.5
x3			0	1.21	2.76	1.12
x4				0	3.35	1.62
x5					0	3.84
x6						0

	x1	x2	(x3,x6)	x4	x5
x1	0	4.27	5.27	6.18	3.98
x2		0	1.96	2.16	1.35
(x3,x6)			0	1.42	3.30
x4				0	3.35
x5					0

	x1	(x2,x5)	(x3,x6)	x4
x1	0	4.13	5.27	6.18
(x2,x5)		0	2.63	2.76
(x3,x6)			0	1.42
x4				0

	x1	(x2,x5)	(x3,x4,x6)
x1	0	4.13	5.57
(x2,x5)		0	2.66
(x3,x4,x6)			0

	x1	(x2,x3,x4,x5,x6)
x1	0	4.99
(x2,x3,x4,x5,x6)		0

Example

$$x_1 = \begin{pmatrix}1.5\\4\end{pmatrix}, x_2 = \begin{pmatrix}0\\0\end{pmatrix}, x_3 = \begin{pmatrix}1\\-1\end{pmatrix}, x_4 = \begin{pmatrix}0.5\\-2.1\end{pmatrix}, x_5 = \begin{pmatrix}-1\\0.9\end{pmatrix}, x_6 = \begin{pmatrix}2\\-1.5\end{pmatrix}$$

Centroid:

	x1	x2	x3	x4	x5	x6
x1	0	4.27	5.02	6.18	3.98	5.52
x2		0	1.41	2.16	1.35	2.5
x3			0	1.21	2.76	1.12
x4				0	3.35	1.62
x5					0	3.84
x6						0

Centroid of (x3,x6)= $\begin{pmatrix}1.5\\-1.25\end{pmatrix}$

	x1	x2	(x3,x6)	x4	x5
x1	0	4.27	5.25	6.18	3.98
x2		0	1.95	2.16	1.35
(x3,x6)			0	1.31	3.30
x4				0	3.35
x5					0

	x1	x2	(x3,x4,x6)	x5
x1	0	4.27	5.54	3.98
x2		0	1.93	1.35
(x3,x4,x6)			0	3.26
x5				0

Centroid of (x3,x4,x6)= $\begin{pmatrix}1.17\\-1.53\end{pmatrix}$

Centroid of (x2,x5)= $\begin{pmatrix}-0.5\\0.45\end{pmatrix}$

	x1	(x2,x5)	(x3,x4,x6)
x1	0	4.07	5.54
(x2,x5)		0	2.59
(x3,x4,x6)			0

	x1	(x2,x3,x4,x5,x6)
x1	0	4.84
(x2,x3,x4,x5,x6)		0

Centroid of (x2,x3,x4,x5,x6)= $\begin{pmatrix}0.5\\-0.74\end{pmatrix}$