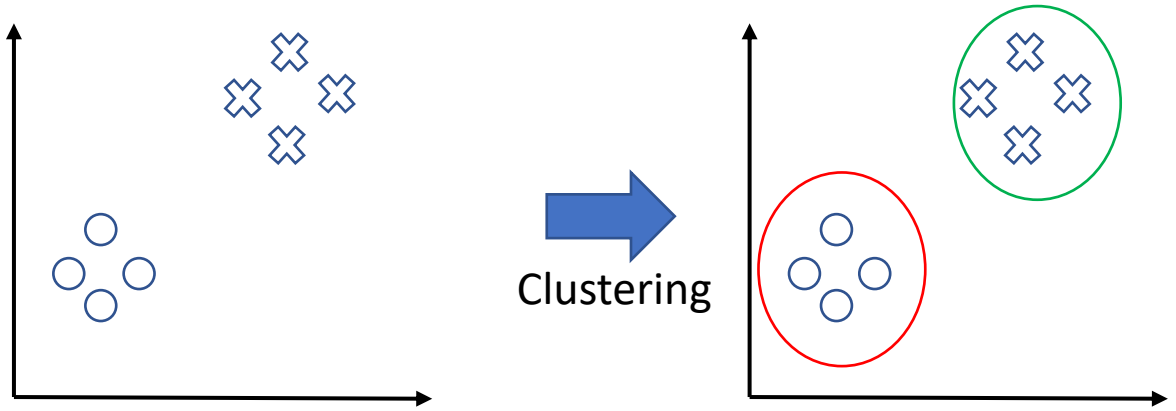
Clustering

Hongchang Gao

Spring 2024

What is clustering?

- Unsupervised Learning
  - Given: only samples, NO labels
  - Clustering: find meaningful groups of samples s.t.
    - Samples in the same group are “similar”
    - Samples in different groups are “dissimilar”



Know NOTHING about labels

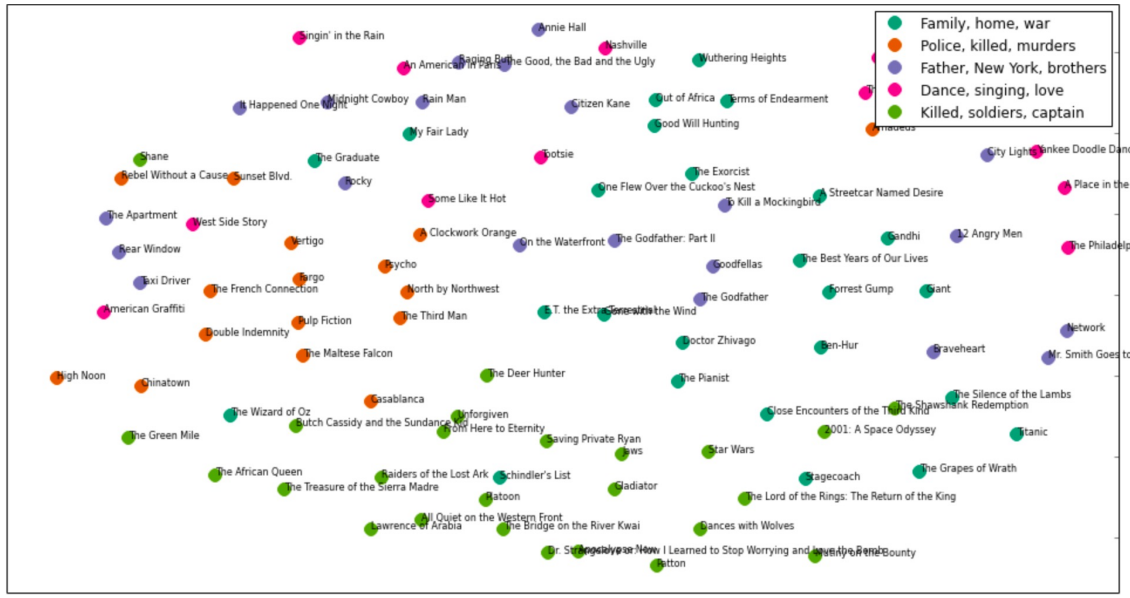
Examples

- Image Segmentation

http://people.cs.uchicago.edu/ pff/segment

Examples

- Topic discovery



athletics	cricket	football	rugby	tennis
olymp	wicket	chelsea	wale	seed
athlet	cricket	arsen	ireland	63
indoor	test	leagu	robinson	open
athen	pakistan	club	england	64
kenteri	seri	unit	nation	76
thanou	bowl	liverpool	franc	australian
greek	india	mourinho	rugbi	roddick
iaaf	south	football	six	75
drug	onedai	manag	scotland	hewitt
race	africa	manchest	itali	beat

Clustering

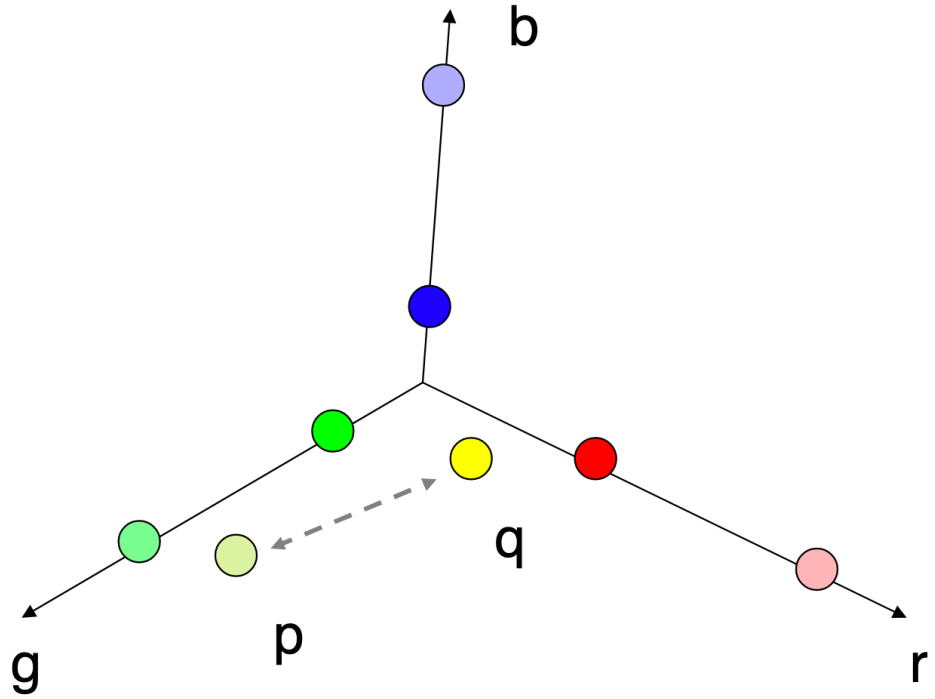
- How to measure the similarity between different samples?
  - Euclidean distance

$$d = \|\mathbf{x} - \mathbf{y}\|_2 = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots + (x_d - y_d)^2}$$

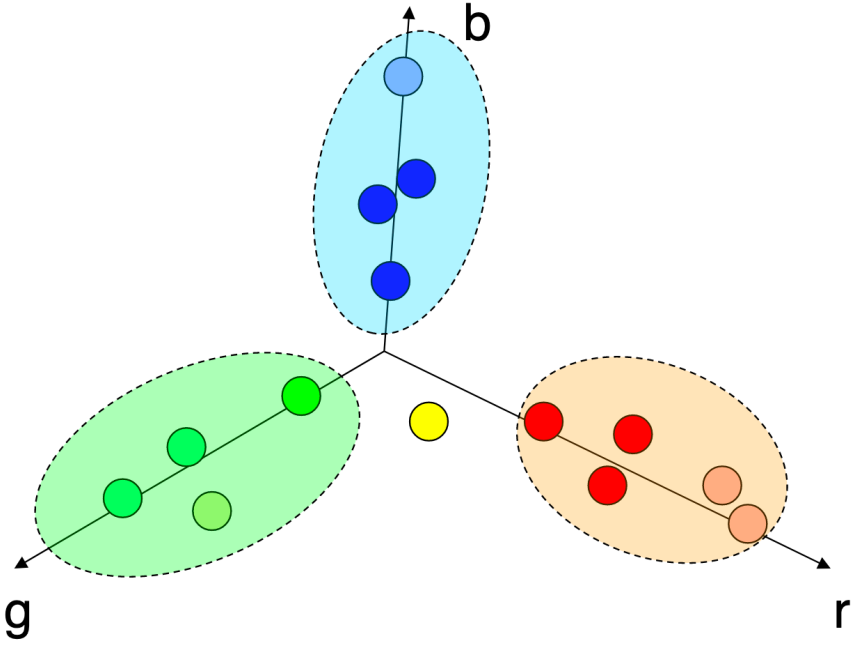
    - Large distance, small similarity
    - Small distance, large similarity

Clustering

$$d = \|\mathbf{x} - \mathbf{y}\|_2 = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots + (x_d - y_d)^2}$$



Compute the similarity between different samples

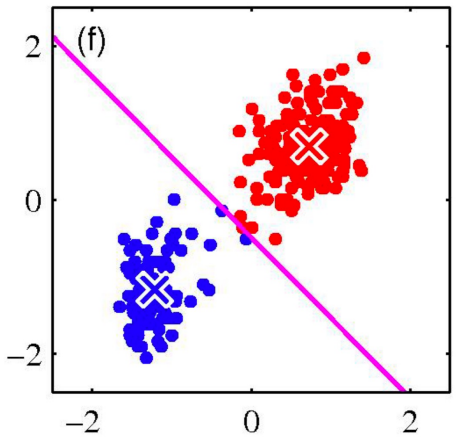


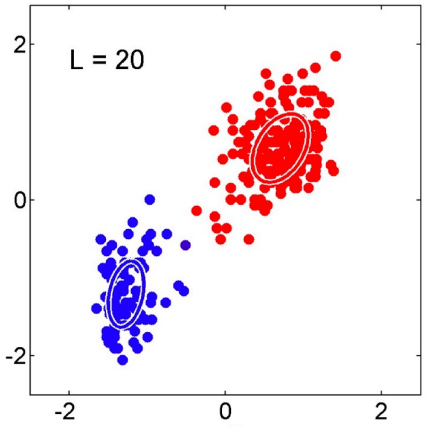
Cluster samples together with high similarity

Fig ref: https://web.eecs.umich.edu/~jjcorso/t/598F14/files/lecture_1006_clustering.pdf

Clustering

- 1. Partition methods
  - Construct various partitions and then evaluate them by some criterion
    - K-means
    - Gaussian mixture model
    - Spectral clustering





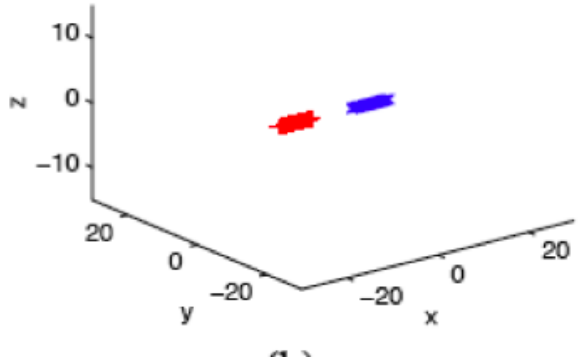
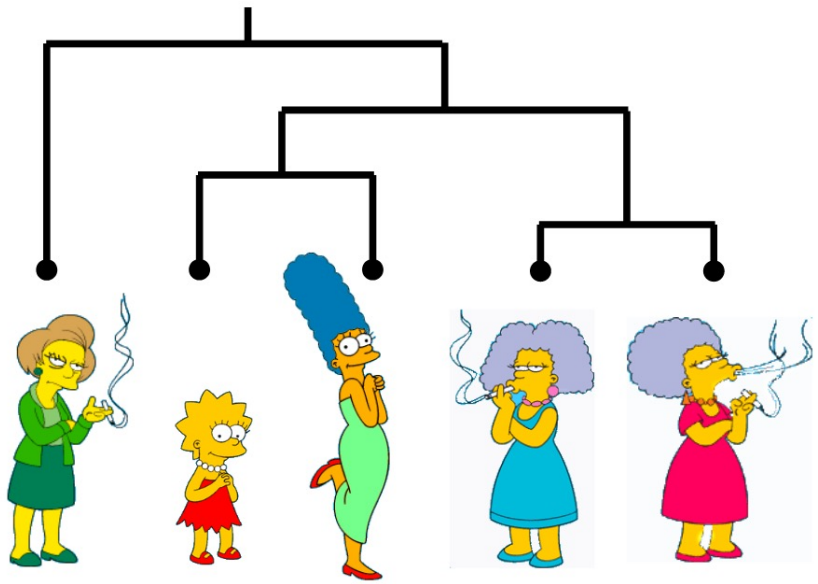


Fig: https://www.cs.cmu.edu/~epxing/Class/10701/slides/clustering.pdf

Clustering

- 2. Hierarchical methods
  - Create a hierarchical decomposition of the set of objects using some criterion
    - Bottom up – agglomerative
    - Top down – divisive



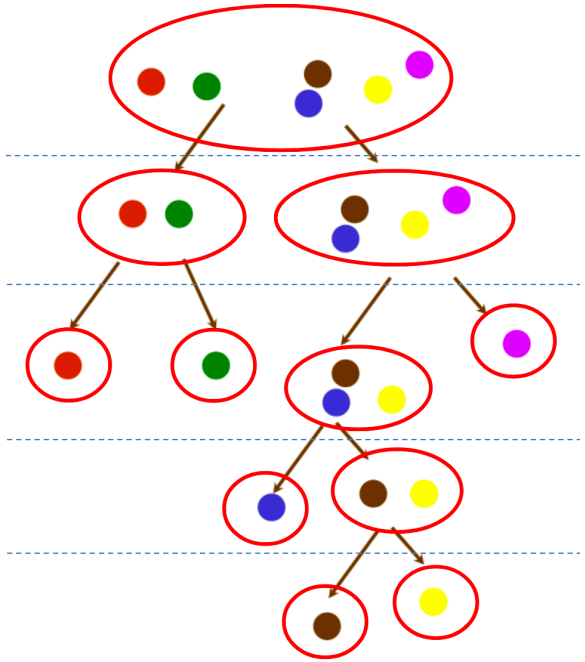


Fig: https://www.cs.cmu.edu/~epxing/Class/10701/slides/clustering.pdf

K-Means

- Given a dataset $\{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n\}$, K-Means partitions it into K clusters:
  - Each cluster has a cluster center, called centroid
  - K is specified by the user

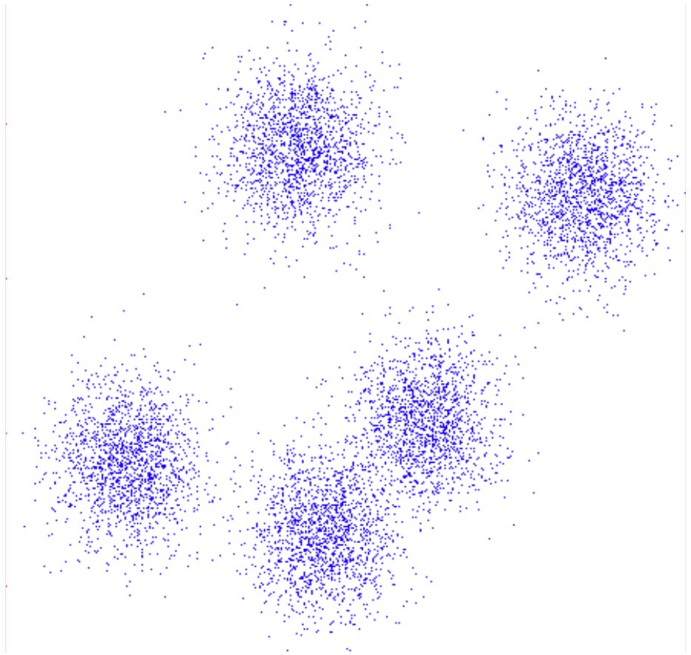
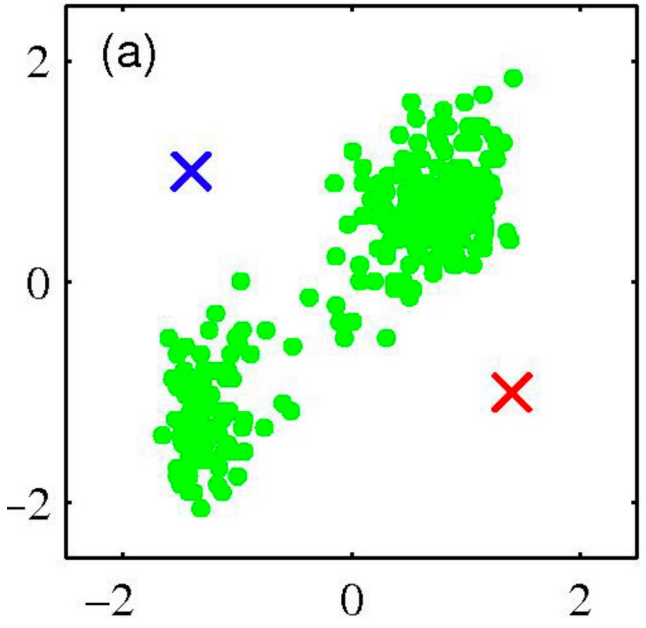


Fig: https://people.csail.mit.edu/dsontag/courses/ml12/slides/lecture14.pdf

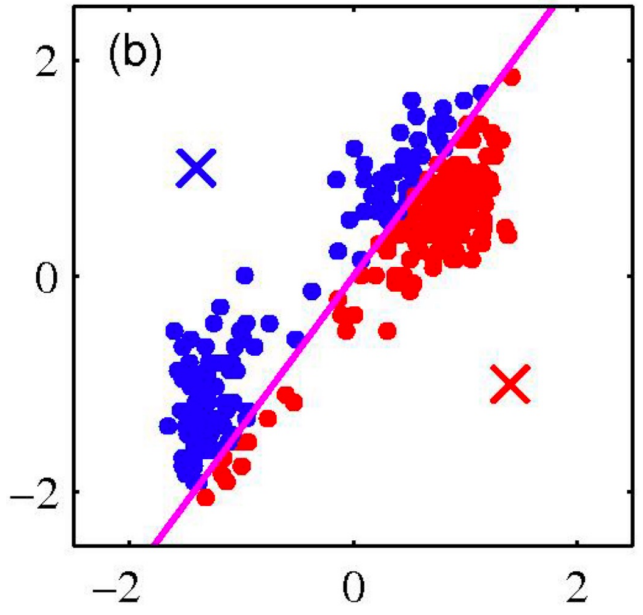
K-Means

- 1. Randomly initialize the cluster centroid $\mu_1, \mu_2, \cdots, \mu_K$
- 2. Repeat until no change in μ_i
  - 2.1 Classify N samples in terms of the nearest cluster centroid
  - 2.2 Re-compute the cluster centroid



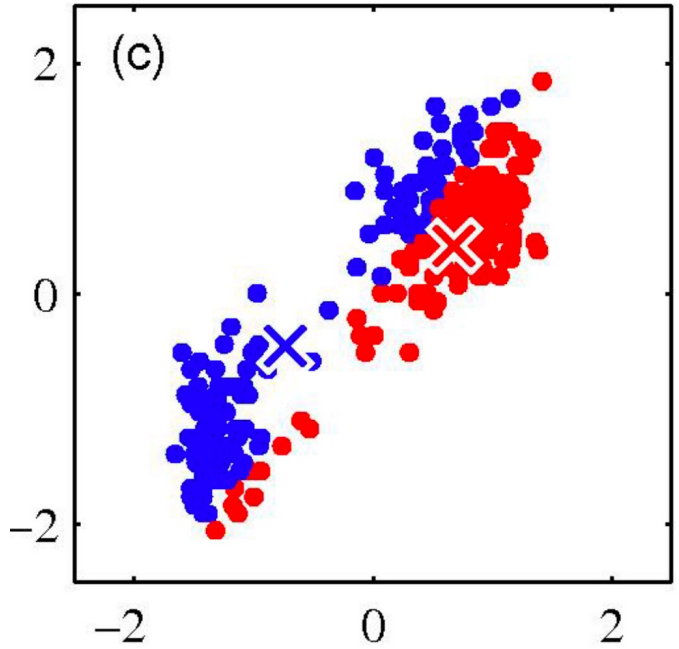
K-Means

- 1. Randomly initialize the cluster centroid $\mu_1, \mu_2, \cdots, \mu_K$
- 2. Repeat until no change in μ_i
  - 2.1 Classify N samples in terms of the nearest cluster centroid
  - 2.2 Re-compute the cluster centroid



$$\|\mathbf{x}_i - \mu_1\|_2$$

$$\|\mathbf{x}_i - \mu_2\|_2$$

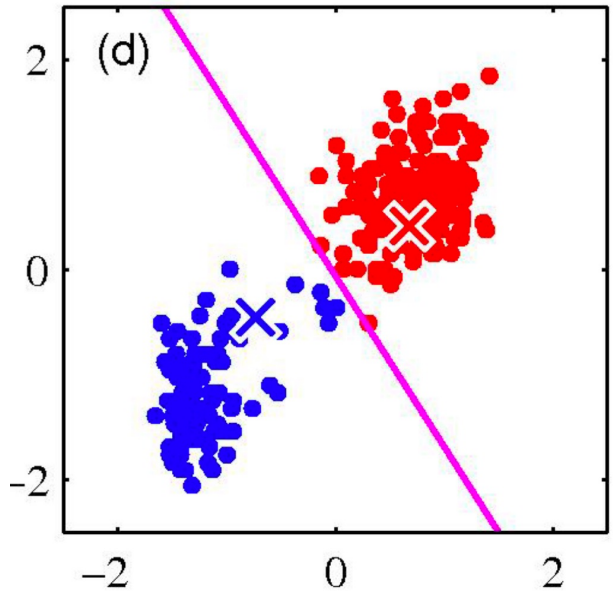


$$\mu_1 = \frac{1}{|C_1|} \sum_{\mathbf{x}_i \in C_1} \mathbf{x}_i$$

$$\mu_2 = \frac{1}{|C_2|} \sum_{\mathbf{x}_i \in C_2} \mathbf{x}_i$$

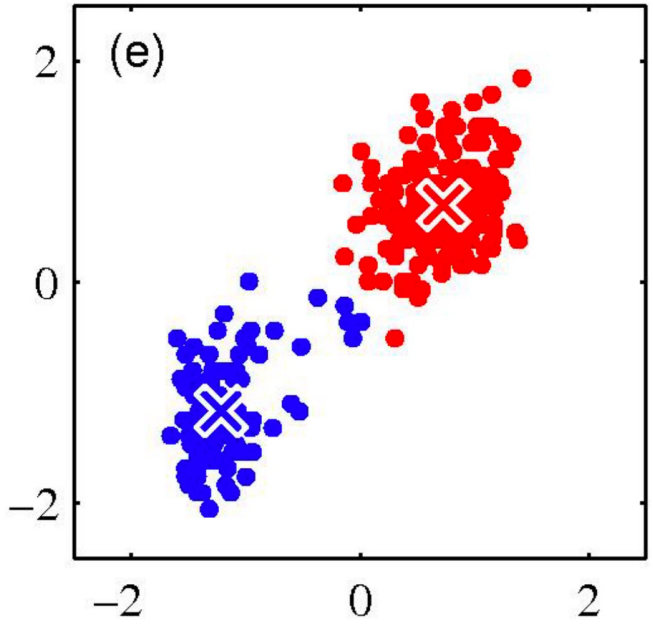
K-Means

- 1. Randomly initialize the cluster centroid $\mu_1, \mu_2, \cdots, \mu_K$
- 2. Repeat until no change in μ_i
  - 2.1 Classify N samples in terms of the nearest cluster centroid
  - 2.2 Re-compute the cluster centroid



$$\|\mathbf{x}_i - \mu_1\|_2$$

$$\|\mathbf{x}_i - \mu_2\|_2$$

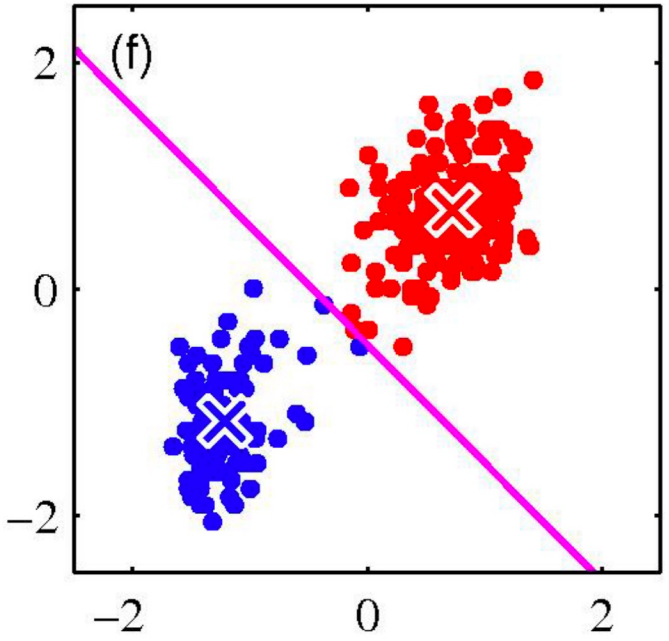


$$\mu_1 = \frac{1}{|C_1|} \sum_{\mathbf{x}_i \in C_1} \mathbf{x}_i$$

$$\mu_2 = \frac{1}{|C_2|} \sum_{\mathbf{x}_i \in C_2} \mathbf{x}_i$$

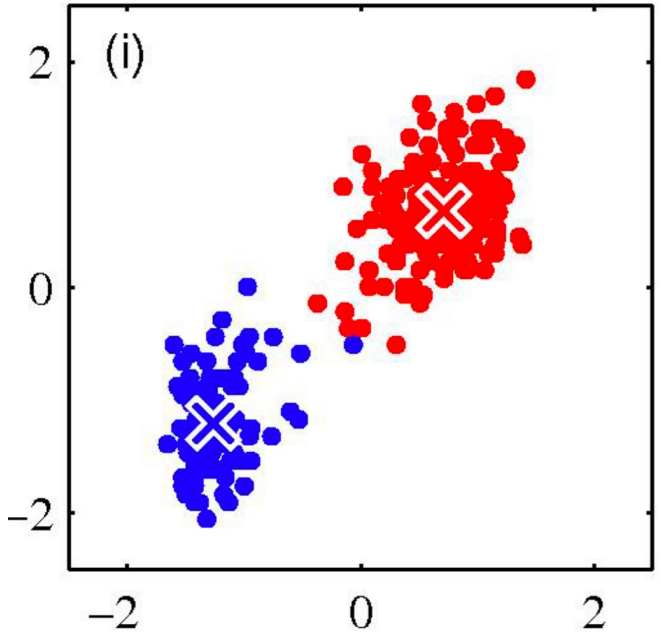
K-Means

- 1. Randomly initialize the cluster centroid $\mu_1, \mu_2, \cdots, \mu_K$
- 2. Repeat until no change in μ_i
  - 2.1 Classify N samples in terms of the nearest cluster centroid
  - 2.2 Re-compute the cluster centroid



$$\|\mathbf{x}_i - \mu_1\|_2$$

$$\|\mathbf{x}_i - \mu_2\|_2$$



$$\mu_1 = \frac{1}{|C_1|} \sum_{\mathbf{x}_i \in C_1} \mathbf{x}_i$$

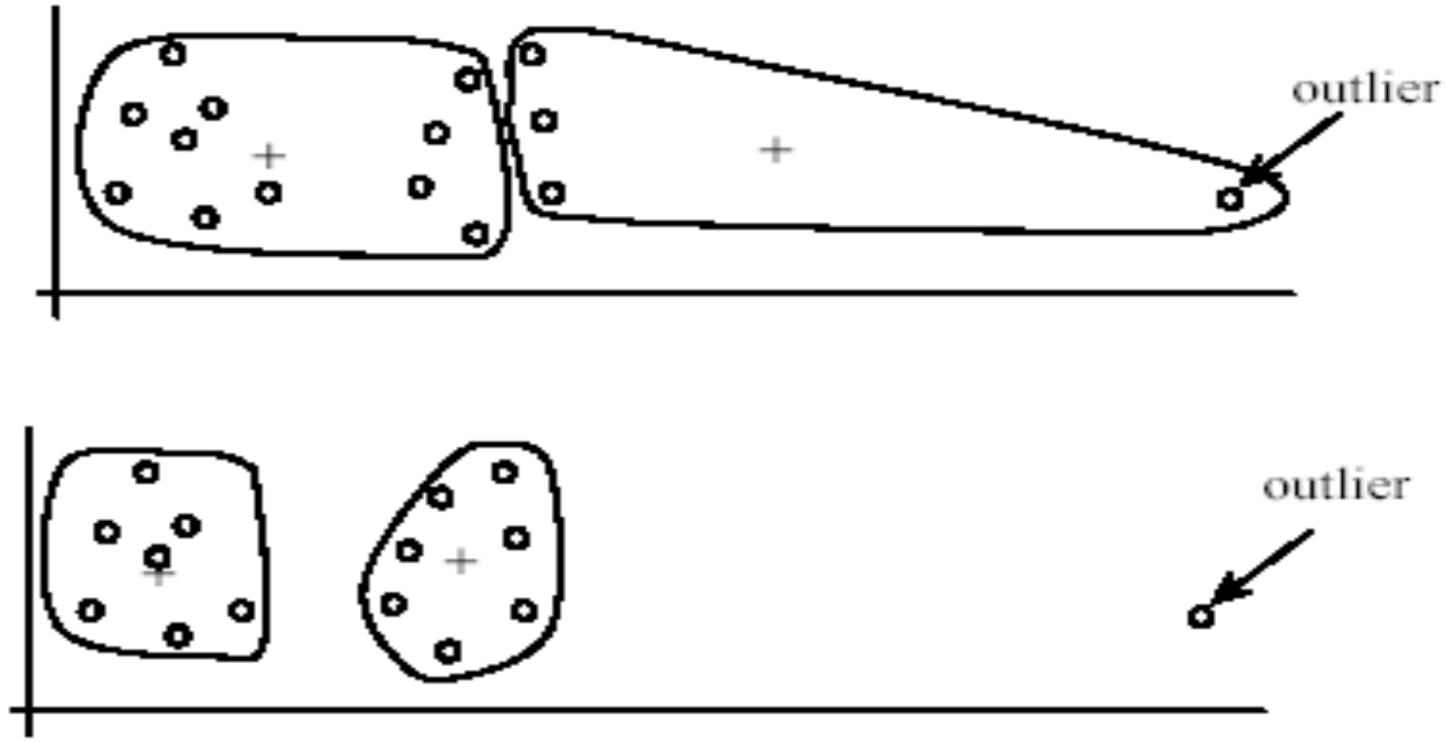
$$\mu_2 = \frac{1}{|C_2|} \sum_{\mathbf{x}_i \in C_2} \mathbf{x}_i$$

K-Means

- Strength:
  - Simple: easy to understand and implement
  - Efficient: O(KNT)
    - K is the number of clusters
    - N is the number of samples
    - T is the number of iterations
- Weakness:
  - Only applicable when the mean is defined
  - Sensitive to outliers
  - Sensitive to initialization

K-Means

- Outliers

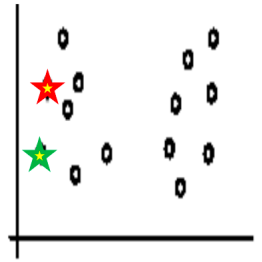


Remove some data points that are much further away from the centroids

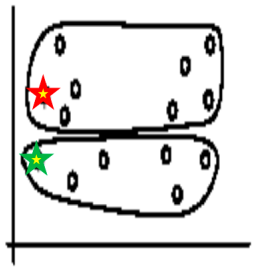
Fig: http://www.mit.edu/~9.54/fall14/slides/Class13.pdf

K-Means

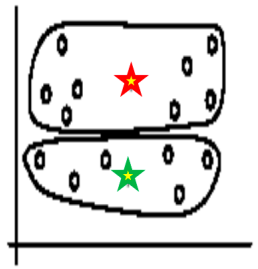
- Sensitive to initialization



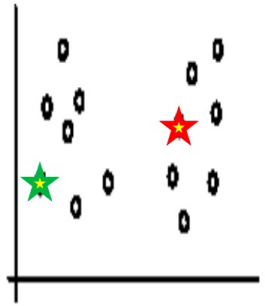
Random selection of seeds (centroids)



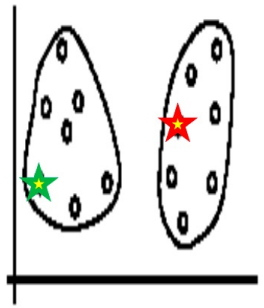
Iteration 1



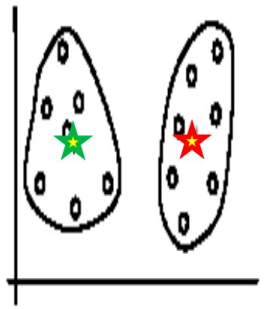
Iteration 2



Random selection of seeds (centroids)



Iteration 1

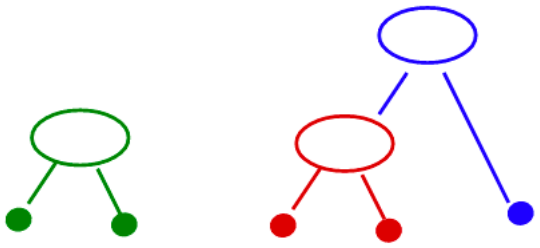


Iteration 2

Fig: http://www.mit.edu/~9.54/fall14/slides/Class13.pdf

Agglomerative Clustering

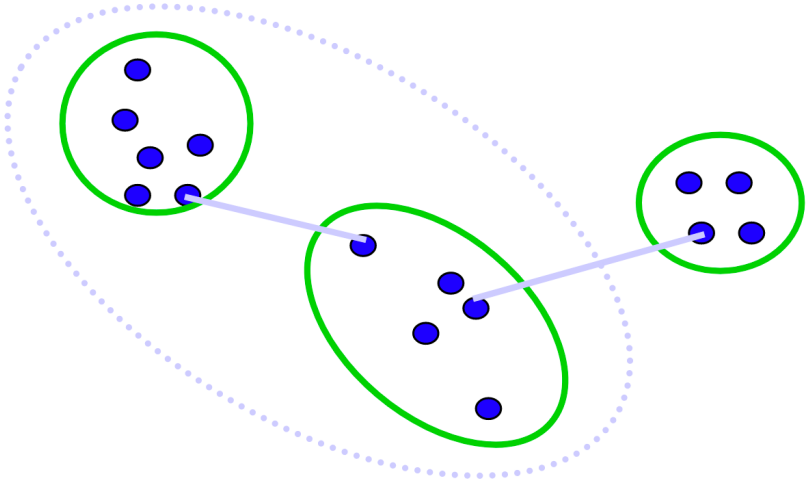
- Agglomerative clustering:
  - Each sample is a cluster
  - Repeat:
    - Pick the two closest clusters
    - Merge them into a new cluster
    - Stop when there's only one cluster left



Agglomerative Clustering

- How to measure the similarity between two clusters?
  - 1. Single link:
    - Distance of two closest samples in each cluster
    - Potentially long and skinny clusters

$$d_{\min}(D_i, D_j) = \min_{x \in D_i, y \in D_j} \| x - y \|$$

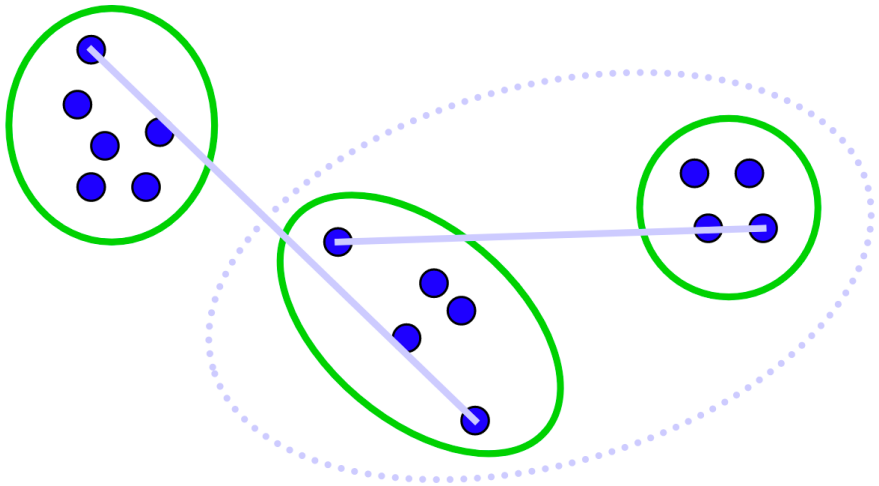


	y1	y2	y3
x1	2	5	7
x2	9	4	6

Agglomerative Clustering

- How to measure the similarity between two clusters?
  - 2. Complete link:
    - Distance of two farthest samples in each cluster
    - Tighter clusters

$$d_{\max}(D_i, D_j) = \max_{x \in D_i, y \in D_j} \| x - y \|$$

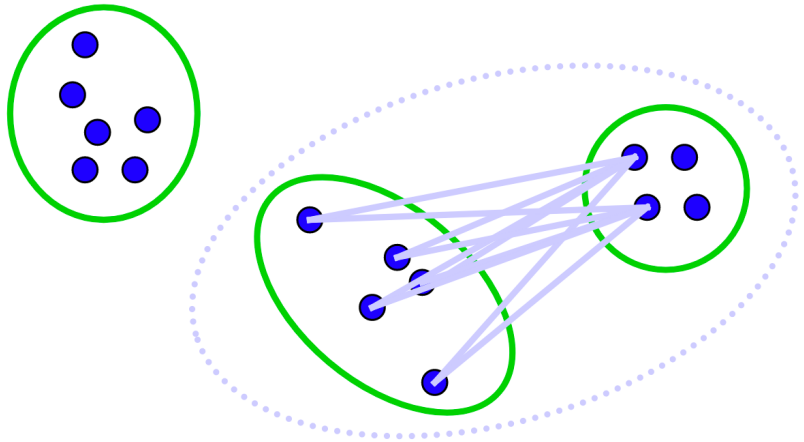


	y1	y2	y3
x1	2	5	7
x2	9	4	6

Agglomerative Clustering

- How to measure the similarity between two clusters?
  - 3. Average link:
    - Average distance of all pairs
    - Robust against noise
    - Most widely used method

$$d_{avg}(D_i, D_j) = \frac{1}{n_i n_j} \sum_{x \in D_i} \sum_{y \in D_j} \| x - y \|$$

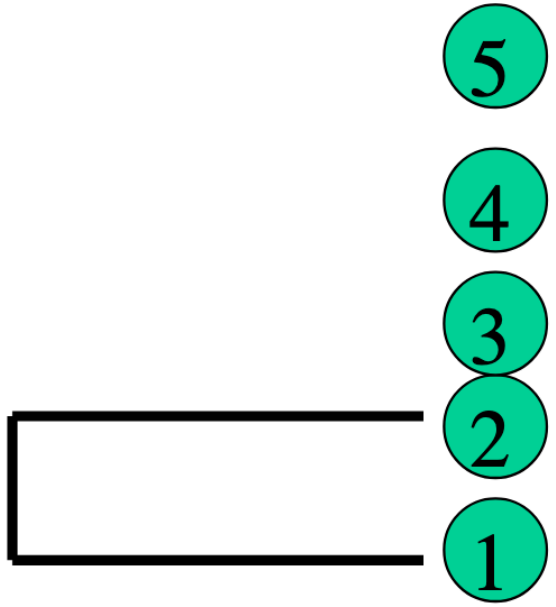


	y1	y2	y3
x1	2	5	7
x2	9	4	6

Agglomerative Clustering

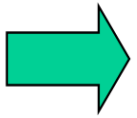
- Single link example

	1	2	3	4	5
1	0				
2	2	0			
3	6	3	0		
4	10	9	7	0	
5	9	8	5	4	0



Agglomerative Clustering

	1	2	3	4	5
1	0				
2	2	0			
3	6	3	0		
4	10	9	7	0	
5	9	8	5	4	0

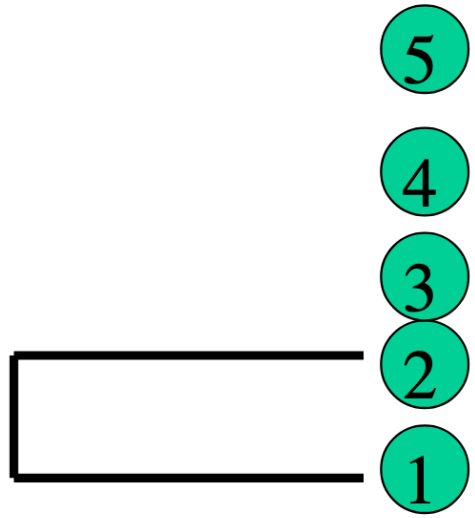


	(1,2)	3	4	5
(1,2)				
3				
4				
5				

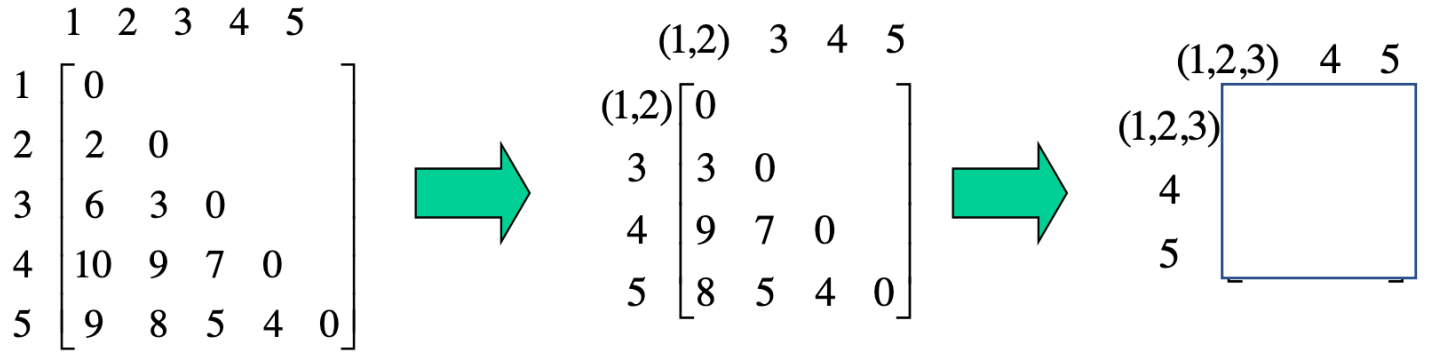
$d_{(1,2),3} =$

$d_{(1,2),4} =$

$d_{(1,2),5} =$

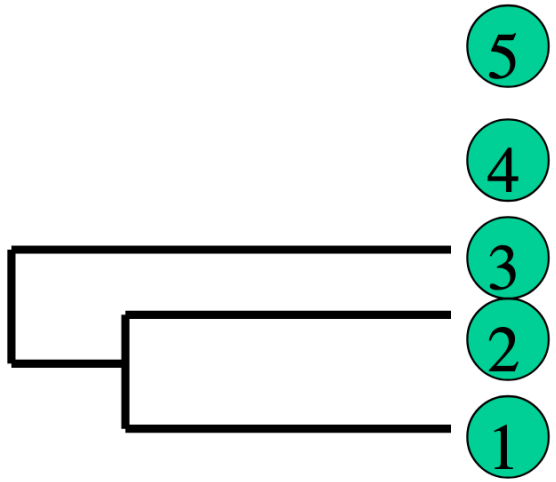


Agglomerative Clustering



$d_{(1,2,3),4} =$

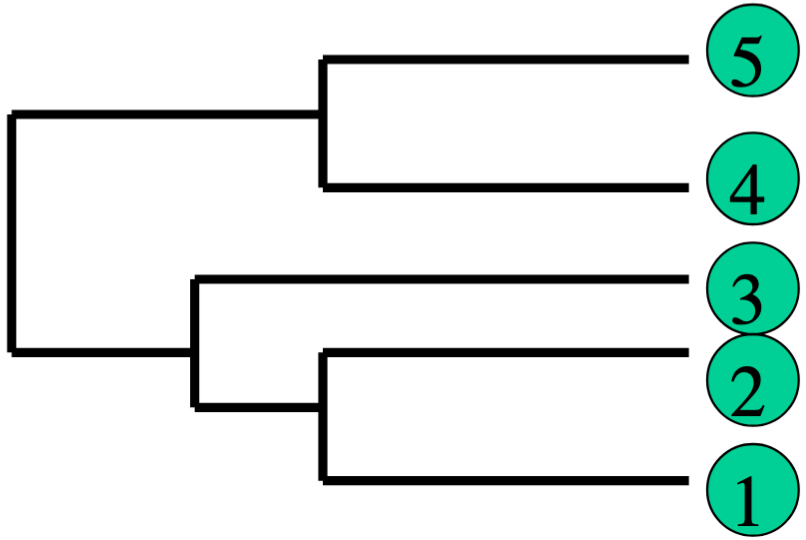
$d_{(1,2,3),5} =$



Agglomerative Clustering

$$\begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array}\begin{array}{c} \begin{array}{ccccc} 1 & 2 & 3 & 4 & 5 \end{array} \\ \begin{bmatrix} 0 & & & & \\ 2 & 0 & & & \\ 6 & 3 & 0 & & \\ 10 & 9 & 7 & 0 & \\ 9 & 8 & 5 & 4 & 0 \end{bmatrix} \end{array} \Rightarrow \begin{array}{c} \\ (1,2) \\ 3 \\ 4 \\ 5 \end{array}\begin{array}{c} \begin{array}{cccc} (1,2) & 3 & 4 & 5 \end{array} \\ \begin{bmatrix} 0 & & & \\ 3 & 0 & & \\ 9 & 7 & 0 & \\ 8 & 5 & 4 & 0 \end{bmatrix} \end{array} \Rightarrow \begin{array}{c} \\ (1,2,3) \\ 4 \\ 5 \end{array}\begin{array}{c} \begin{array}{ccc} (1,2,3) & 4 & 5 \end{array} \\ \begin{bmatrix} 0 & & \\ 7 & 0 & \\ 5 & 4 & 0 \end{bmatrix} \end{array}$$

$$d_{(1,2,3),(4,5)} = \min\{d_{(1,2,3),4}, d_{(1,2,3),5}\} = 5$$



Complete Link

Step 0: {x_1}, {x_2}, {x_3}, {x_4}, {x_5}

	X_1	X_2	X_3	X_4	X_5
X_1	0				
X_2	1	0			
X_3	5	2	0		
X_4	10	8	6	0	
X_5	9	7	4	3	0

Step 1: {x_1, x_2}, {x_3}, {x_4}, {x_5}

	{X_1, x_2}	X_3	X_4	X_5
{X_1,x_2}	0			
X_3	5	0		
X_4	10	6	0	
X_5	9	4	3	0

d({x_1, x_2}, x_3)=max{d(x_1, x_3), d(x_2, x_3)}=max{5, 2}=5

d({x_1, x_2}, x_4)=max{d(x_1, x_4), d(x_2, x_4)}=max{10, 8}=10

d({x_1, x_2}, x_5)=max{d(x_1, x_5), d(x_2, x_5)}=max{9, 7}=9

Complete Link

Step 2: {x_1, x_2}, {x_3}, {x_4, x_5}

	{X_1, x_2}	X_3	X_4	X_5
{X_1,x_2}	0			
X_3	5	0		
X_4	10	6	0	
X_5	9	4	3	0

	{X_1, x_2}	X_3	{X_4, x_5}
{X_1,x_2}	0		
X_3	5	0	
{X_4, x_5}	10	6	0

d({x_4, x_5}, {x_1, x_2})=max{d(x_4, {x_1, x_2}), d(x_5, {x_1, x_2})}=max{10, 9}=10

d({x_4, x_5}, x_3)=max{d(x_4, x_3), d(x_5, x_3)}=max{6, 4}=6

Complete Link

Step 3: {x_1, x_2, x_3}, {x_4, x_5}

	{X_1, x_2}	X_3	{X_4, x_5}
{X_1,x_2}	0		
X_3	5	0	
{X_4, x_5}	10	6	0

	{X_1, x_2, x_3}	{X_4, x_5}
{X_1,x_2, x_3}	0	
{X_4, x_5}	10	0

d({x_4, x_5}, {x_1, x_2, x_3})=max{d({x_4, x_5} , {x_1, x_2}), d({x_4, x_5}, x_3)}=max{10, 6}=10

Step 4: {x_1, x_2, x_3, x_4, x_5}

Agglomerative Clustering

- Properties:
  - No need to specify the number of clusters in advance
  - Not scale well
    - O(N^2)