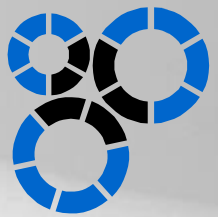
Machine Learning
CSE - 465

Lecture - 08

Lecture 08 (Unsupervised Learning) Clustering

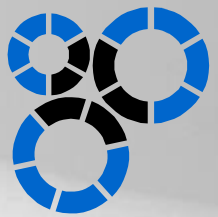
Outline

- Introduction to clustering
- K-means Clustering
- Weaknesses of K-means clustering
- Hierarchical Clustering

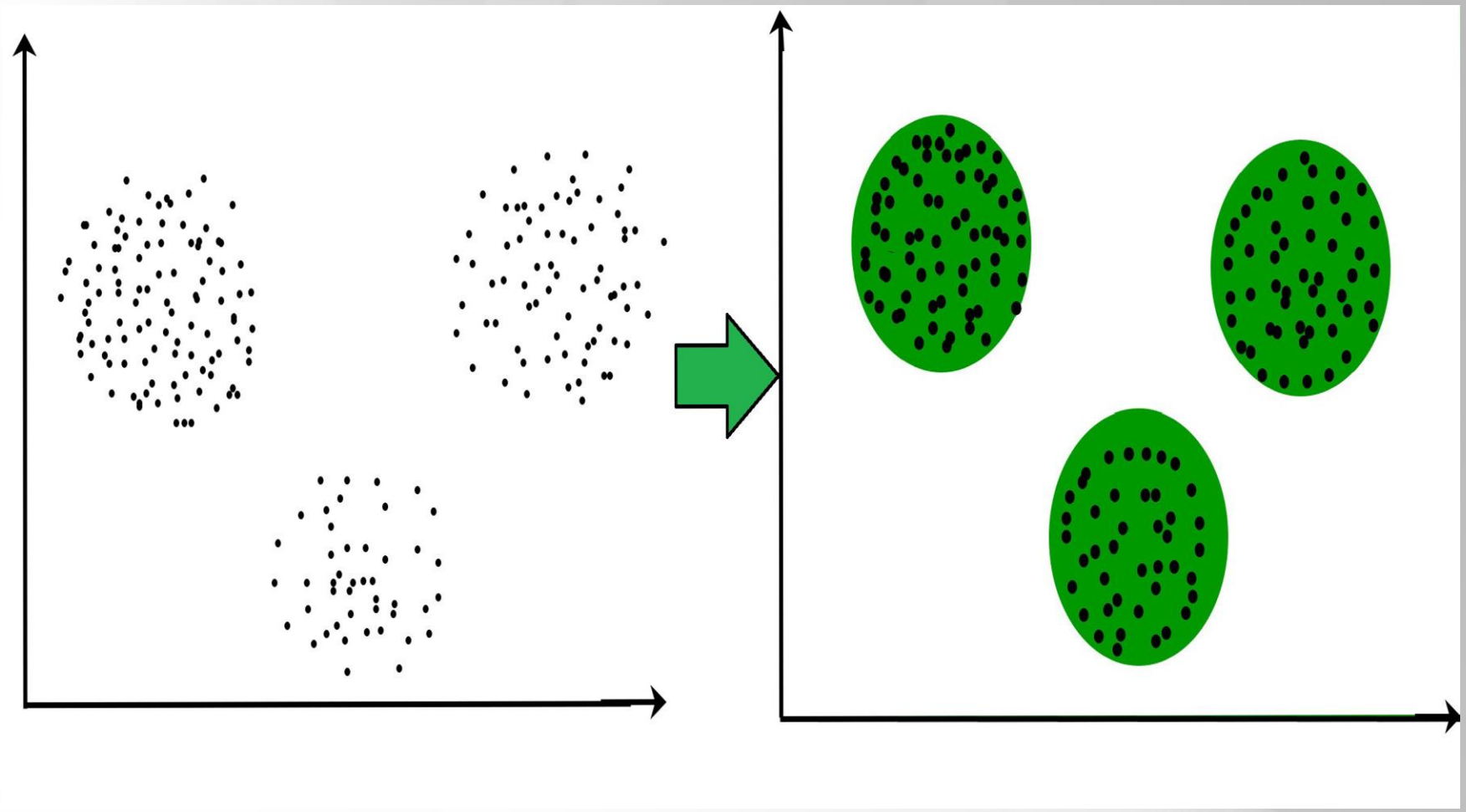


Introduction to Clustering

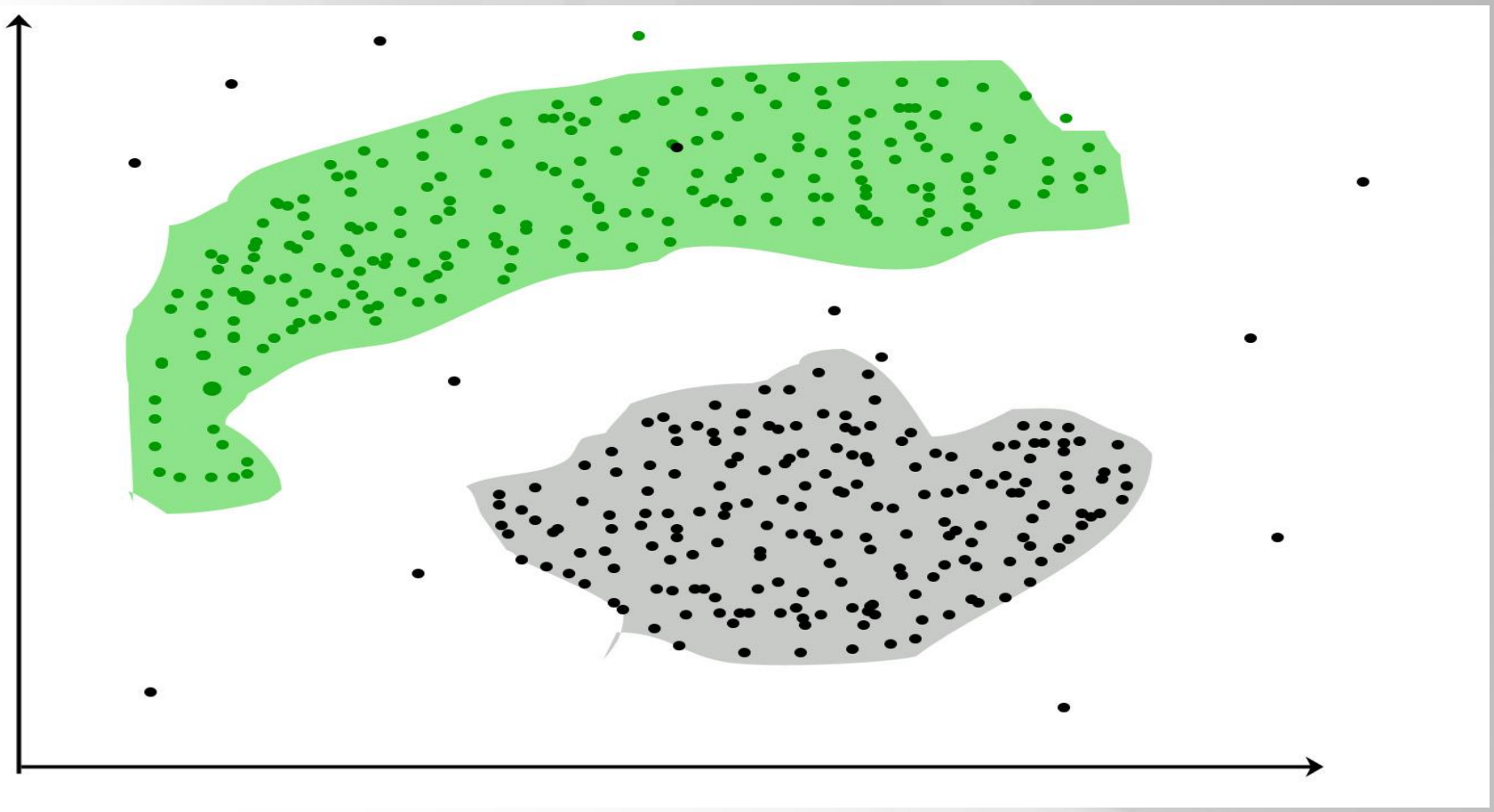
- **Clustering** is the task of dividing the population or data points into a number of groups such that data points in the same groups are more similar to other data points in the same group and dissimilar to the data points in other groups.
- It is basically a collection of objects on the basis of similarity and dissimilarity between them.
- It is basically a type of *unsupervised learning method*



Introduction to Clustering

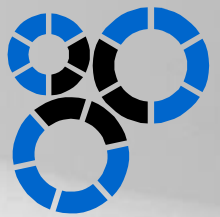


Introduction to Clustering



K-means Clustering

- Input: n objects (or points) and a number k
- Algorithm
  1. Randomly place K points into the space represented by the objects that are being clustered. These points represent initial group centroids.
  2. Assign each object to the group that has the closest centroid.
  3. When all objects have been assigned, recalculate the positions of the K centroids.
  4. Repeat Steps 2 and 3 until the stopping criteria is met.



Euclidean distance and Manhattan distance

- If $h = 2$, it is the Euclidean distance

$$dist(\mathbf{x}_i, \mathbf{x}_j) = \sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2 + \ldots + (x_{ir} - x_{jr})^2}$$

- If $h = 1$, it is the Manhattan distance

$$dist(\mathbf{x}_i, \mathbf{x}_j) = |x_{i1} - x_{j1}| + |x_{i2} - x_{j2}| + \ldots + |x_{ir} - x_{jr}|$$

- ***Input:*** Number of Objects = 6, Number of Clusters = 2

No	X	Y
1	1	1
2	2	3
3	1	2
4	3	3
5	2	2
6	3	1

- Step-1: Choose random K points and set as cluster centers

C1 = (2, 2) C2 = (3 , 3)

- Step-2: Calculating the distance between objects into cluster centroids by using Euclidean Distance

$$D_i = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

1. D1 $= \{(1, 1), (2, 2)\}$

$= \sqrt{(2-1)^2 + (2-1)^2}$

$= 1.41$

2. D1 $= \{(2, 3), (2, 2)\}$

$= \sqrt{(2-2)^2 + (2-3)^2}$

$= 1$

3. D1 $= \{(1, 2), (2, 2)\}$

$= \sqrt{(2-1)^2 + (2-2)^2}$

1. D2 $= \{(1, 1), (3, 3)\}$

$= \sqrt{(3-1)^2 + (3-1)^2}$

$= 2.82$

2. D2 $= \{(2, 3), (3, 3)\}$

$= \sqrt{(3-2)^2 + (3-3)^2}$

$= 1$

3. D2 $= \{(1, 2), (3, 3)\}$

$= \sqrt{(3-1)^2 + (3-2)^2}$

4. D1 $= \{(3, 3), (2, 2)\}$

$= \sqrt{(2-3)^2 + (2-3)^2}$

$= 1.41$

4. D2 $= \{(3, 3), (3, 3)\}$

$= \sqrt{(3-3)^2 + (3-3)^2}$

$= 0$

5. D1 $= \{(2, 2), (2, 2)\}$

$= \sqrt{(2-2)^2 + (2-2)^2}$

$= 0$

5. D2 $= \{(2, 2), (3, 3)\}$

$= \sqrt{(3-2)^2 + (3-2)^2}$

$= 1.41$

6. D1 $= \{(3, 1), (2, 2)\}$

$= \sqrt{(2-3)^2 + (2-1)^2}$

$= 1.41$

6. D2 $= \{(3, 1), (3, 3)\}$

$= \sqrt{(3-3)^2 + (3-1)^2}$

$= 2$

C1 = {(1 , 1), (1 , 2), (2 , 2), (3 , 1)}

C2 = {(2 , 3), (3 , 3)}

- Step-3: Recalculating the position of the centroid

$$\text{Mean} = \left(\frac{x_1+x_2+ \ldots .x_n}{n}, \frac{y_1+y_2+ \ldots .y_n}{n}\right)$$

$$C1 = \left(\frac{1+1+2+3}{4}, \frac{1+2+2+1}{4}\right)$$

New $C1 = (1.75\ ,\ 1.5)$

$$C2 = \left(\frac{2+3}{2}, \frac{3+3}{2}\right)$$

New $C2 = (2.5,\ 3)$

No	X	Y
1	1	1
2	2	3
3	1	2
4	3	3
5	2	2
6	3	1

- Step-4: Go back to Step 2, unless the centroids are not changing.

k-means (Cont.)

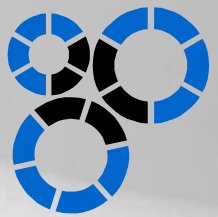
- D1={(1.75,1.5), (1,1)}=.90
- D1={(1.75,1.5), (2,3)}=1.52
- D1={(1.75,1.5), (1,2)}= .90
- D1={(1.75,1.5), (3,3)}=1.95
- D1={(1.75,1.5), (2,2)}=.55
- D1={(1.75,1.5), (3,1)}=1.34

- D2={(2.5,3), (1,1)}=2.5
- D2={(2.5,3), (2,3)}=.5
- D2={(2.5,3), (1,2)}= 1.80
- D2={(2.5,3), (3,3)}=.5
- D2={(2.5,3), (2,2)}=1.11
- D2={(2.5,3), (3,1)}=2.06

C1={(1,1),(1,2),(2,2),(3,1)}

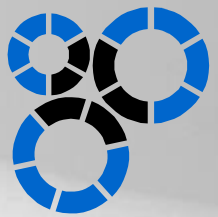
C2={(2,3),(3,3)}

No	X	Y
1	1	1
2	2	3
3	1	2
4	3	3
5	2	2
6	3	1

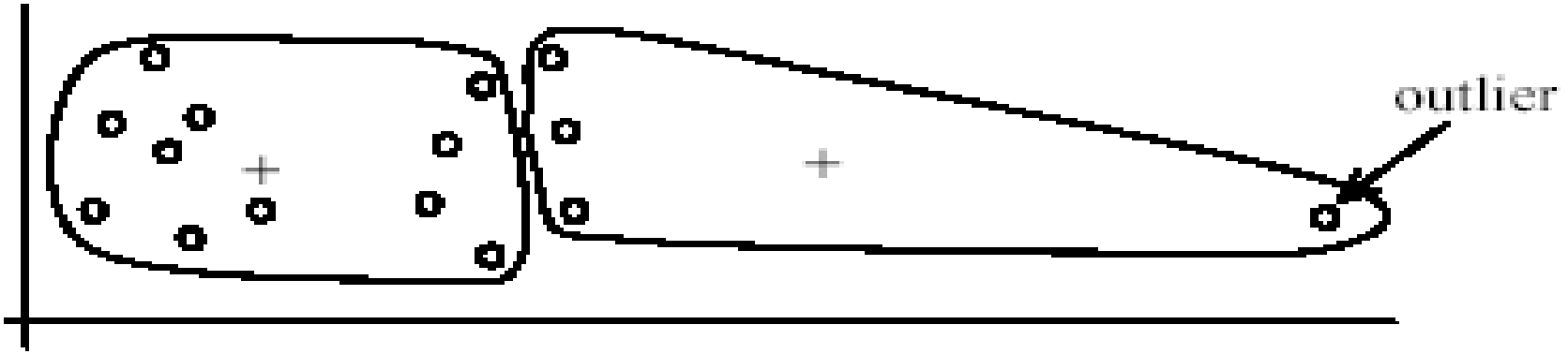


Weaknesses of k-means

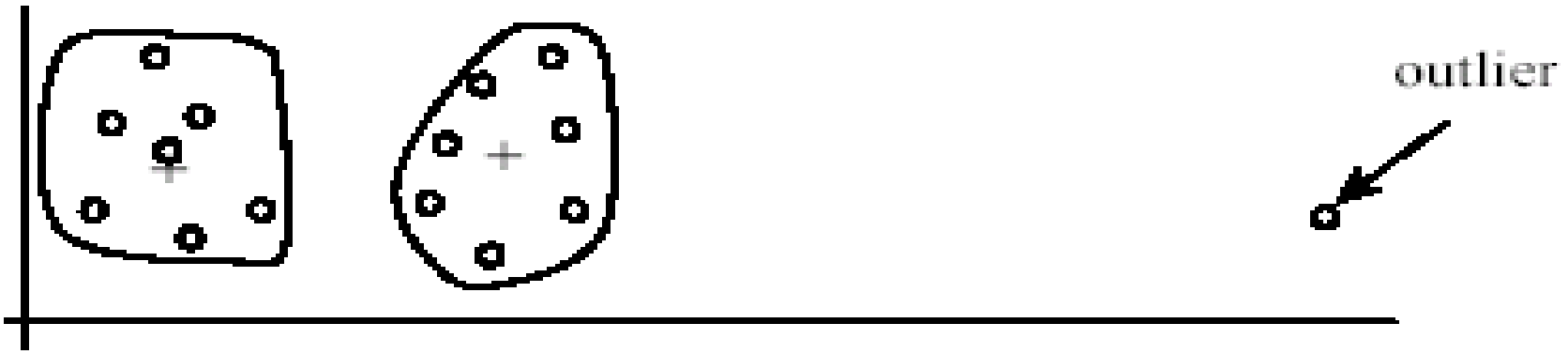
- The algorithm is only applicable if the mean is defined.
  - For categorical data, *k*-mode - the centroid is represented by most frequent values.
- The user needs to specify *k*.
- The algorithm is sensitive to **outliers**
  - Outliers are data points that are very far away from other data points.
  - Outliers could be errors in the data recording or some special data points with very different values.



Weaknesses of k-means: Problems with outliers



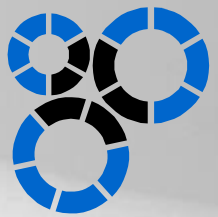
(A): Undesirable clusters



(B): Ideal clusters

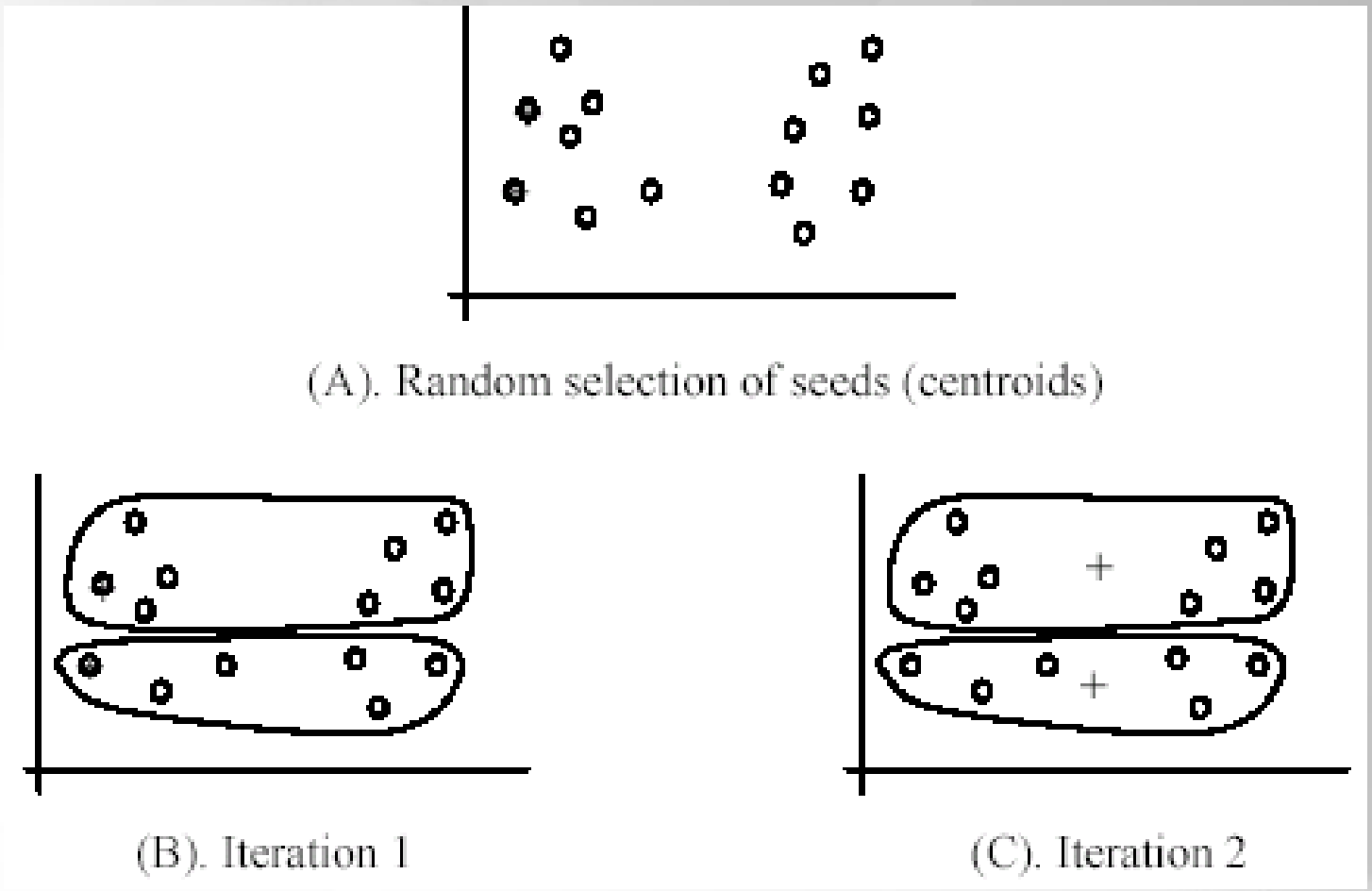
Weaknesses of k-means: To deal with outliers

- One method is to remove some data points in the clustering process that are much further away from the centroids than other data points.
  - To be safe, we may want to monitor these possible outliers over a few iterations and then decide to remove them.
- Another method is to perform random sampling. Since in sampling we only choose a small subset of the data points, the chance of selecting an outlier is very small.
  - Assign the rest of the data points to the clusters by distance or similarity comparison, or classification



Weaknesses of k-means

- The algorithm is sensitive to initial seeds.



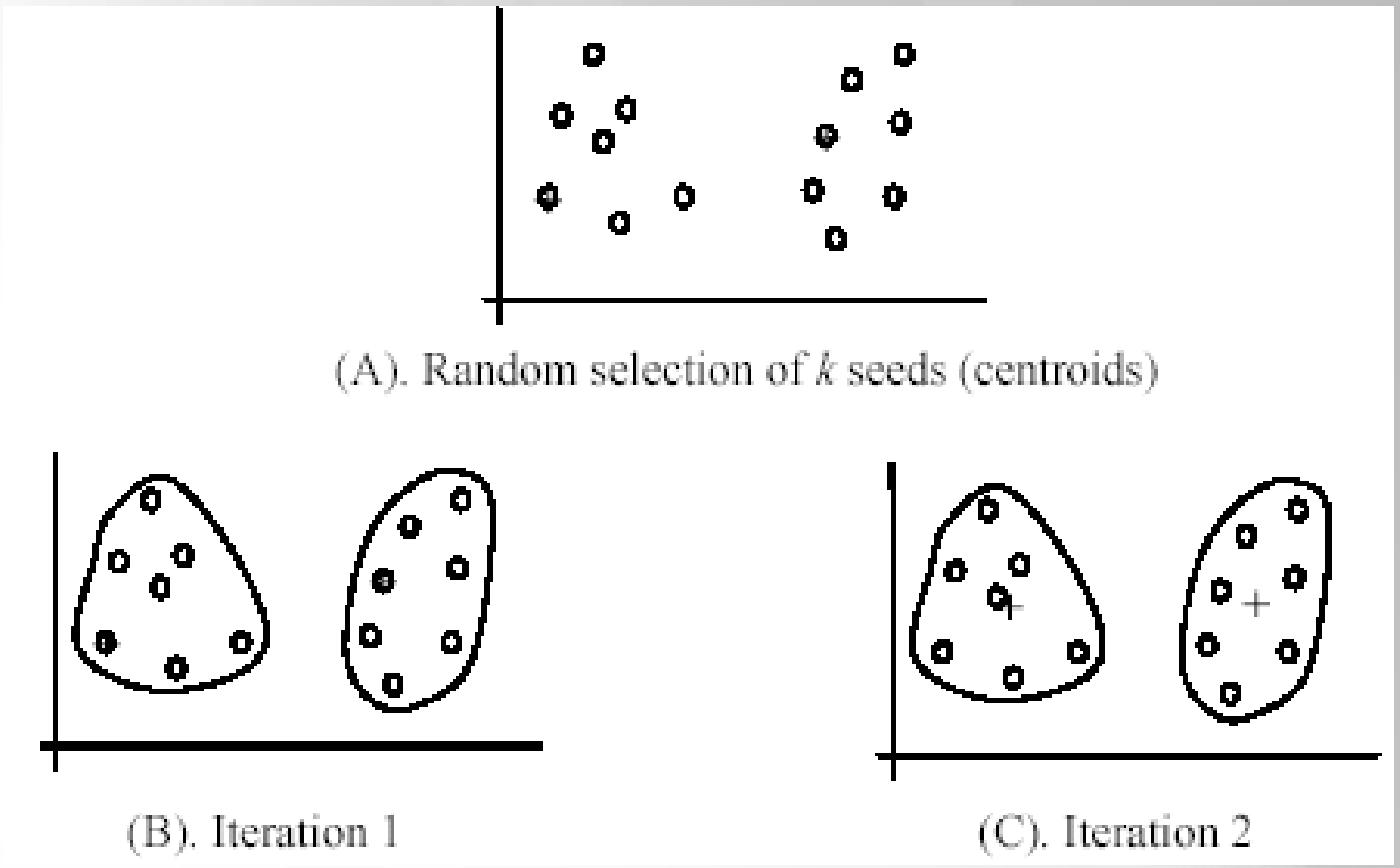
(A). Random selection of seeds (centroids)

(B). Iteration 1

(C). Iteration 2

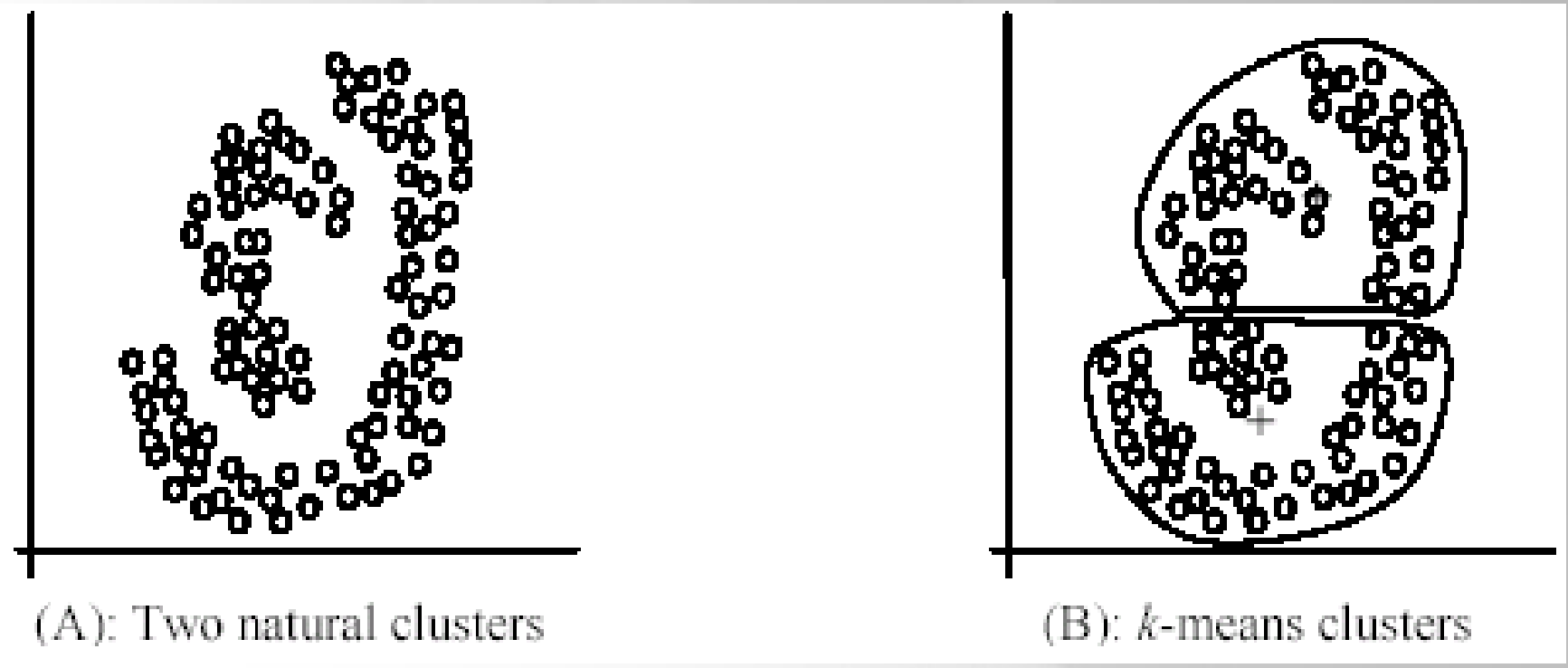
Weaknesses of k-means

If we use different seeds: good results



Weaknesses of k-means

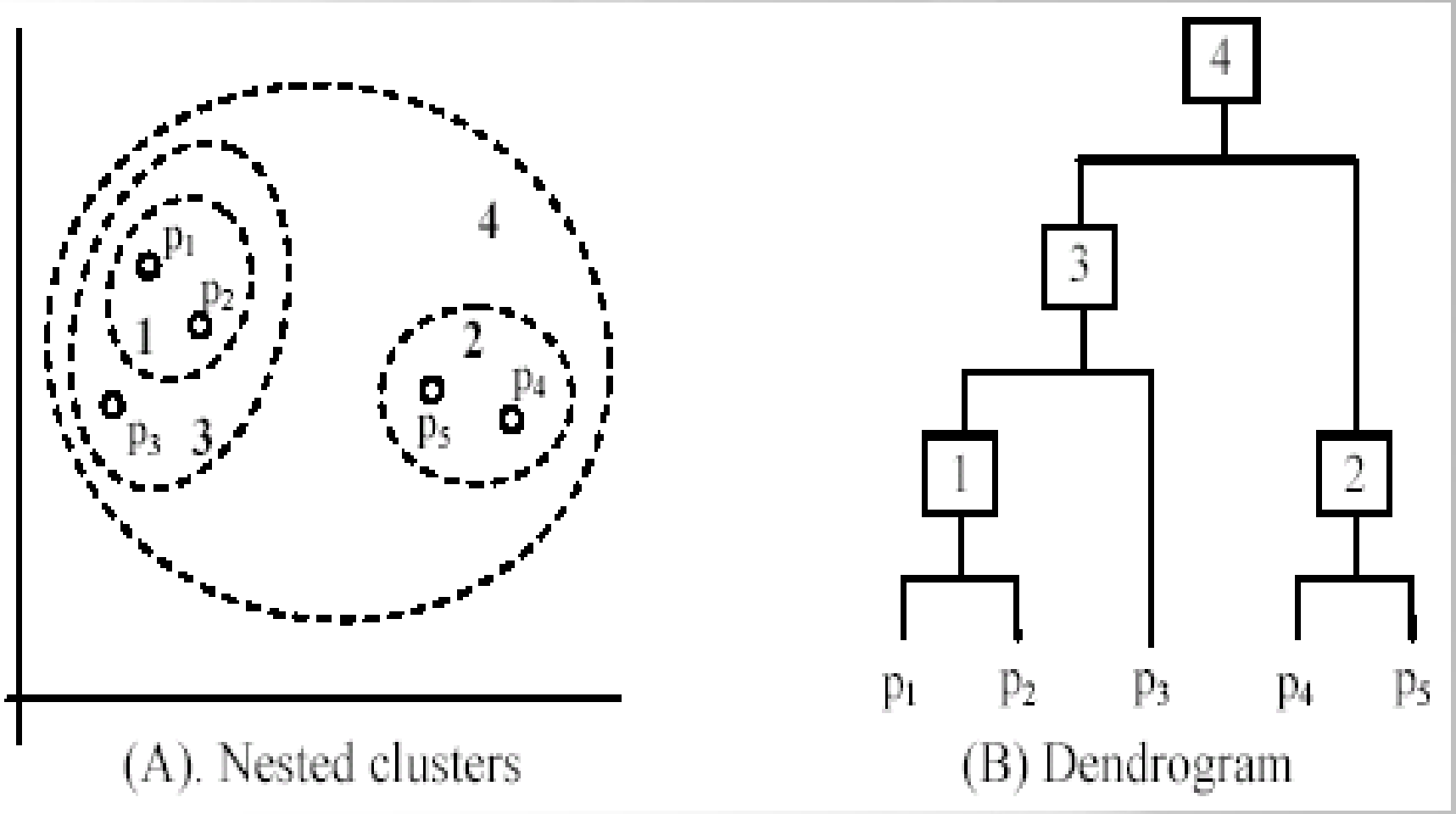
- The *k*-means algorithm is not suitable for discovering clusters that are not hyper-ellipsoids (or hyper-spheres).



(A): Two natural clusters

(B): *k*-means clusters

Hierarchical Clustering



Thank You