



Clustering Methods

Clustering Methods

- Clustering is the process of grouping the data into classes or clusters, so that objects within a cluster have high similarity in comparison to one another but are very dissimilar to objects in other clusters.
- Some major clustering approaches are:
 - Partitioning Methods
 - Hierarchical Methods
 - Density based methods

Requirements and Challenges

- **Scalability**
 - Clustering all the data instead of only on samples
- **Ability to deal with different types of attributes**
 - Numerical, binary, categorical, ordinal, and mixture of these
- **Constraint-based clustering**
 - User may give inputs on constraints
 - Use domain knowledge to determine input parameters
- **Interpretability and usability**
- **Others**
 - Discovery of clusters with arbitrary shape
 - Ability to deal with noisy data
 - Incremental clustering and insensitivity to input order
 - High dimensionality

Partitioning Methods

- This method generally results in a set of k clusters. Each cluster maybe represented by a centroid or a cluster representative
- A partitioning method creates an initial partitioning, given k , the number of partitions to construct. It uses an iterative relocation technique that attempts to improve partitioning by moving objects from one group to another
- Partitioning methods can be used for high dimensional data. However, one disadvantage is that a user needs to initialize the number of clusters.

K-Means Clustering Algorithm

Algorithm: *K-Means*

Input: k , D

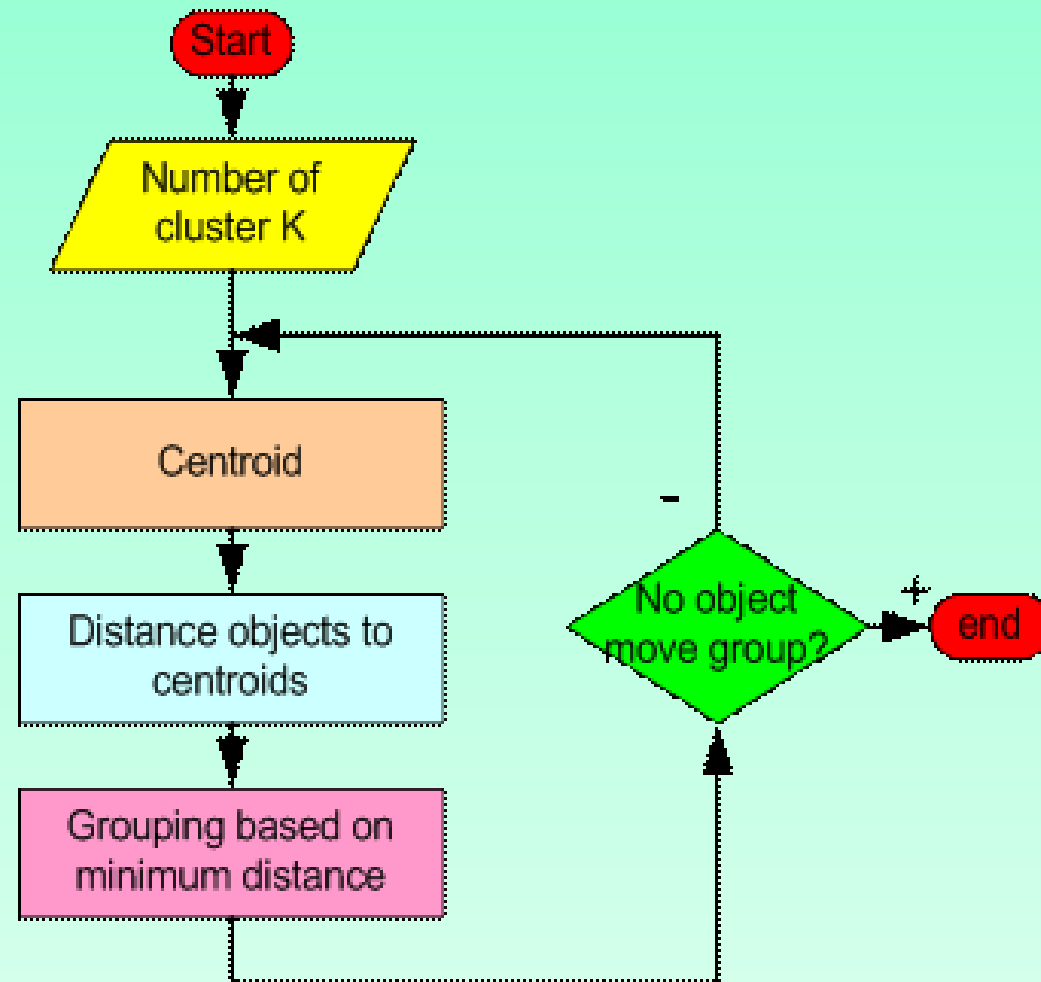
k , is the number of clusters, D a data set containing n objects.

Output: A set of k clusters

Method:

- 1: arbitrarily choose k objects from D as the initial cluster centers;
- 2: **repeat**
- 3: (re)assign each object to the cluster to which the object is the most similar based on the mean value of the objects in the cluster;
- 4: update the cluster means, i.e. calculate the mean value of the objects for each cluster;
- 5: **until** no change;

K-Means Clustering



Example 1: K-Means Clustering

Use k-Means and Euclidean Distance to cluster the following 8 samples into 3 clusters:

$A1=(2,10)$, $A2=(2,5)$, $A3=(8,4)$, $A4=(5,8)$, $A5=(7,5)$, $A6=(6,4)$, $A7=(1,2)$, $A8=(4,9)$

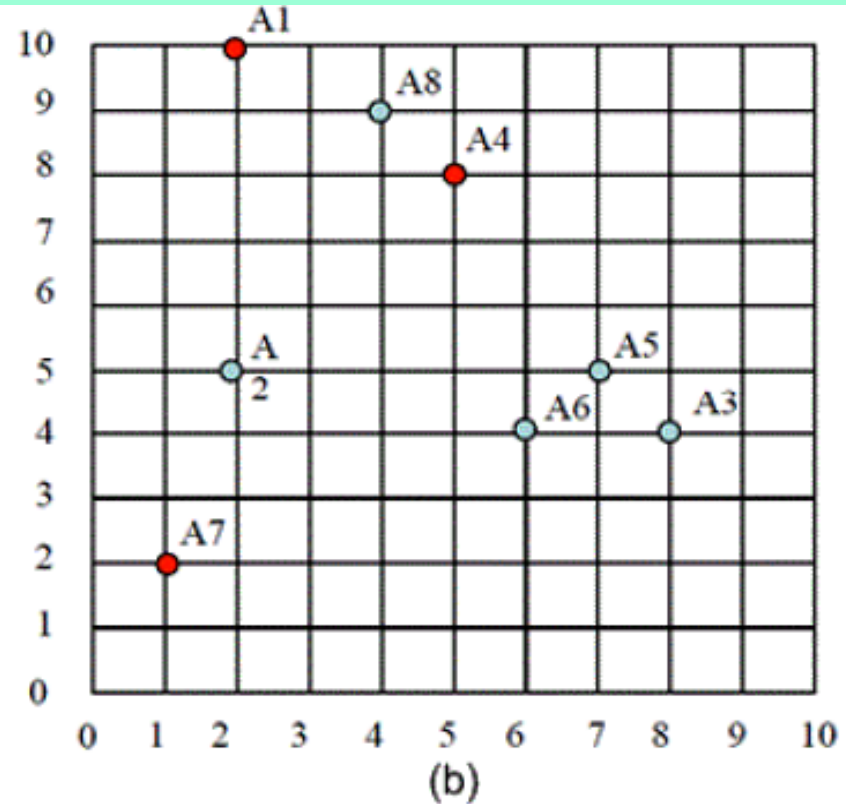
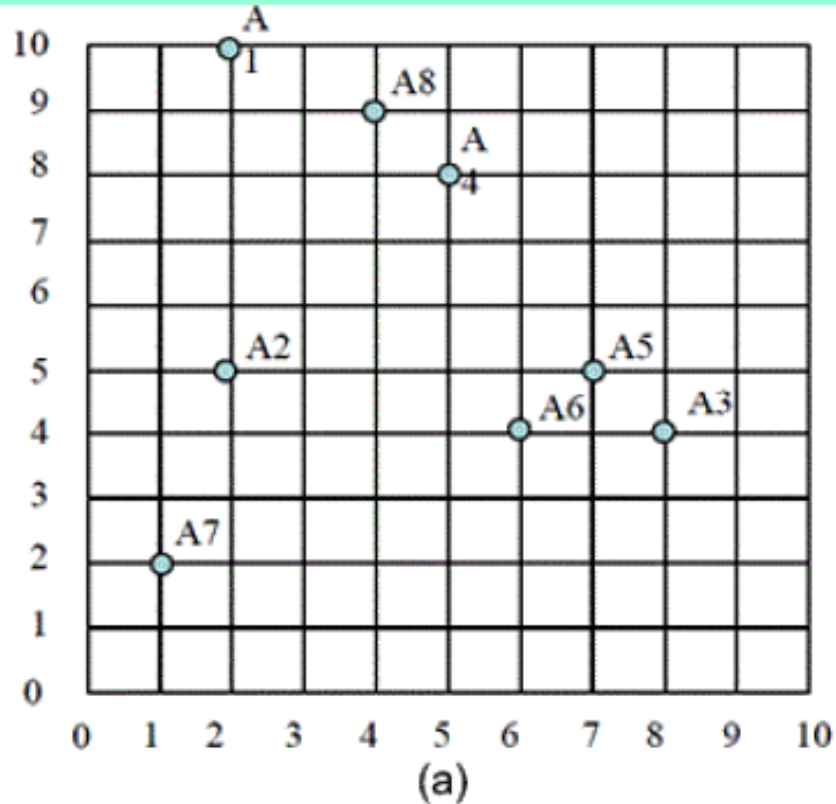


Figure 1: (a) 8 samples (b) 3 clusters (seed1, seed2 and seed3) in red

At the beginning, we initial the centre of the 3 clusters (seed1, seed2 and seed3) to A1, A4 and A7 (Figure 1 (b)).

K-Means Clustering

<p>A1:</p> <ul style="list-style-type: none">• $d(A1, \text{seed1})=0$ as A1 is seed1 (smallest)• $d(A1, \text{seed2})=\sqrt{13}$• $d(A1, \text{seed3})=\sqrt{65}$ <p>→ A1 \in cluster1</p>	<p>A2:</p> <ul style="list-style-type: none">• $d(A2, \text{seed1})=\sqrt{25}$• $d(A2, \text{seed2})=\sqrt{18}$• $d(A2, \text{seed3})=\sqrt{10}$ (smallest) <p>→ A2 \in cluster3</p>
<p>A3:</p> <ul style="list-style-type: none">• $d(A3, \text{seed1})=\sqrt{36}$• $d(A3, \text{seed2})=\sqrt{25}$ (smallest)• $d(A3, \text{seed3})=\sqrt{53}$ <p>→ A3 \in cluster2</p>	<p>A4:</p> <ul style="list-style-type: none">• $d(A4, \text{seed1})=\sqrt{13}$• $d(A4, \text{seed2})=0$ as A4 is seed2 (smallest)• $d(A4, \text{seed3})=\sqrt{52}$ <p>→ A4 \in cluster2</p>
<p>A5:</p> <ul style="list-style-type: none">• $d(A5, \text{seed1})=\sqrt{50}$• $d(A5, \text{seed2})=\sqrt{13}$ (smallest)• $d(A5, \text{seed3})=\sqrt{45}$ <p>→ A5 \in cluster2</p>	<p>A6:</p> <ul style="list-style-type: none">• $d(A6, \text{seed1})=\sqrt{52}$• $d(A6, \text{seed2})=\sqrt{17}$ (smallest)• $d(A6, \text{seed3})=\sqrt{29}$ <p>→ A6 \in cluster2</p>
<p>A7:</p> <ul style="list-style-type: none">• $d(A7, \text{seed1})=\sqrt{65}$• $d(A7, \text{seed2})=\sqrt{52}$• $d(A7, \text{seed3})=0$ as A7 is seed3 (smallest) <p>→ A7 \in cluster2</p>	<p>A8:</p> <ul style="list-style-type: none">• $d(A8, \text{seed1})=\sqrt{5}$• $d(A8, \text{seed2})=\sqrt{2}$ (smallest)• $d(A8, \text{seed3})=\sqrt{58}$ <p>→ A8 \in cluster2</p>

K-Means Clustering

New cluster 1: {A1}, 2: {A3, A4, A5, A6, A8}, 3: {A2, A7}, see Figure 2 (a)

Then, the centres of the new clusters (see Figure 2 (b)):

$$C1=(2,10), C2=(\frac{8+5+7+6+4}{5}, \frac{4+8+5+4+9}{5})=(6,6), C3=(\frac{2+1}{2}, \frac{5+2}{2})=(1.5, 3.5)$$

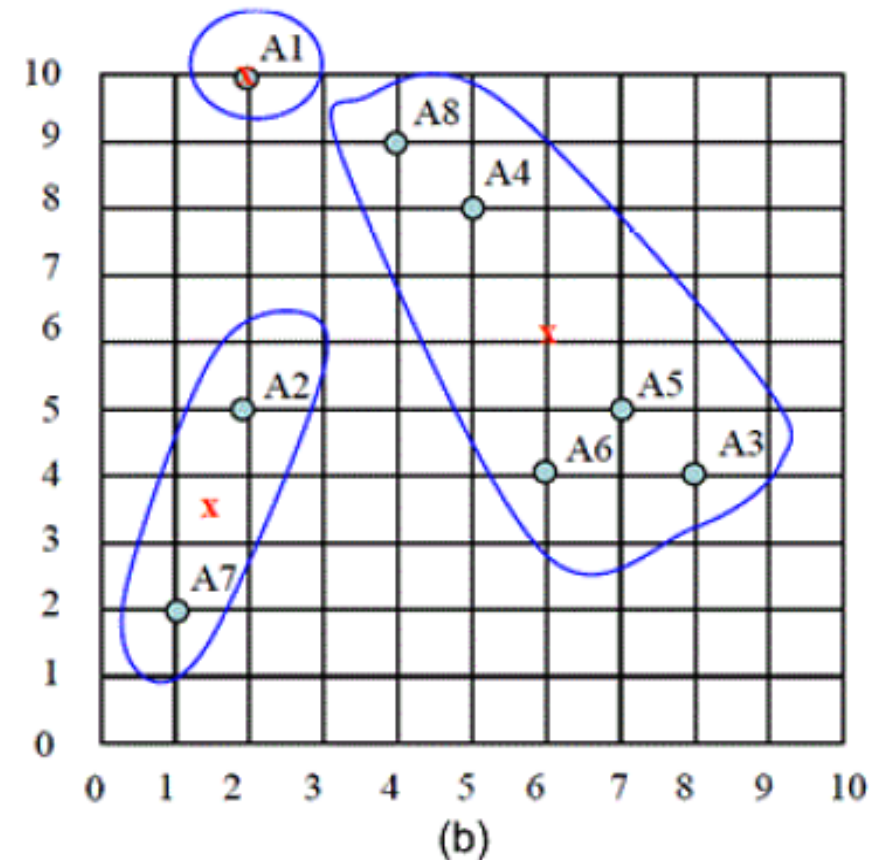
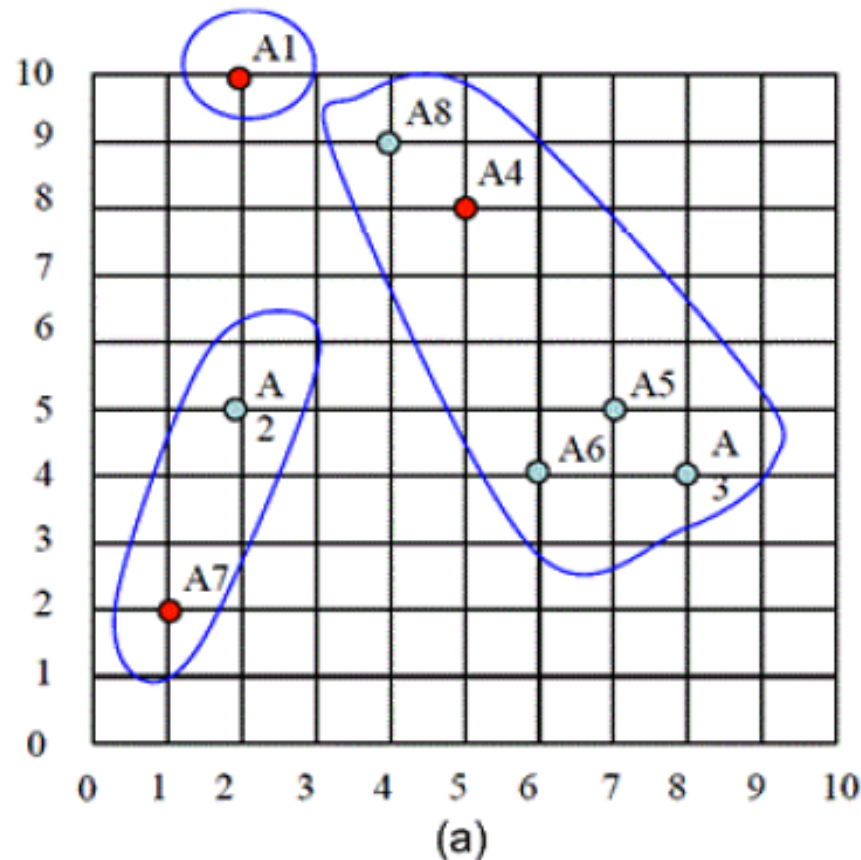


Figure 2: (a) before Epoch 1; (b) centroid has been changed after **Epoch 1**

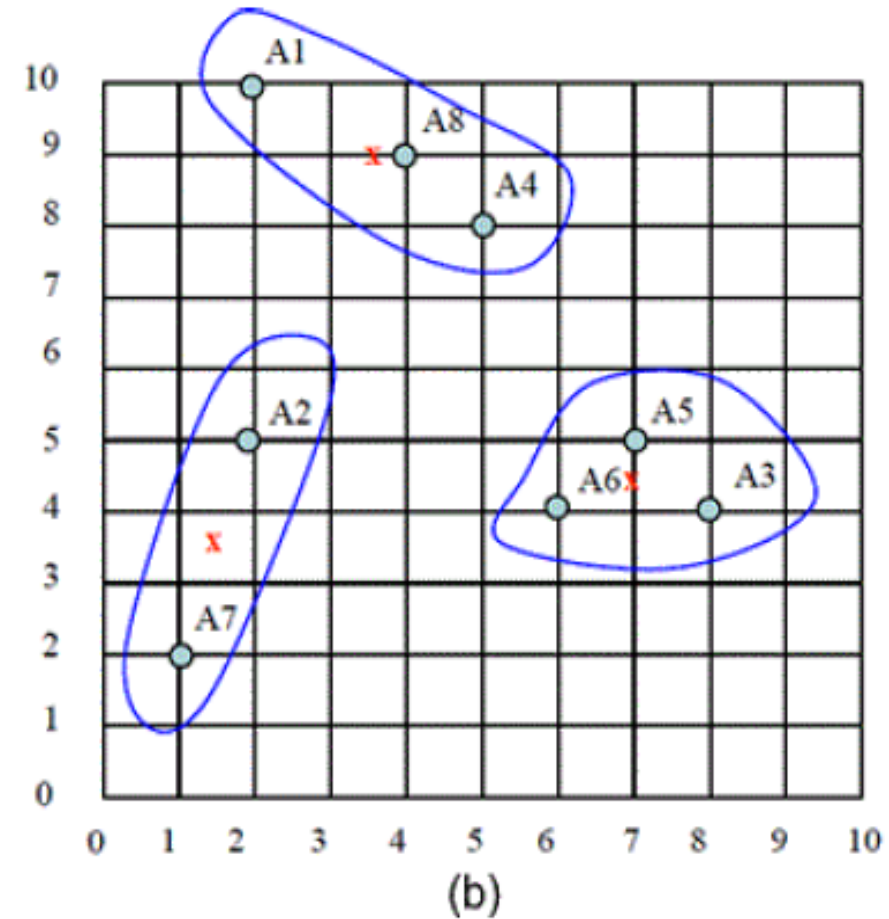
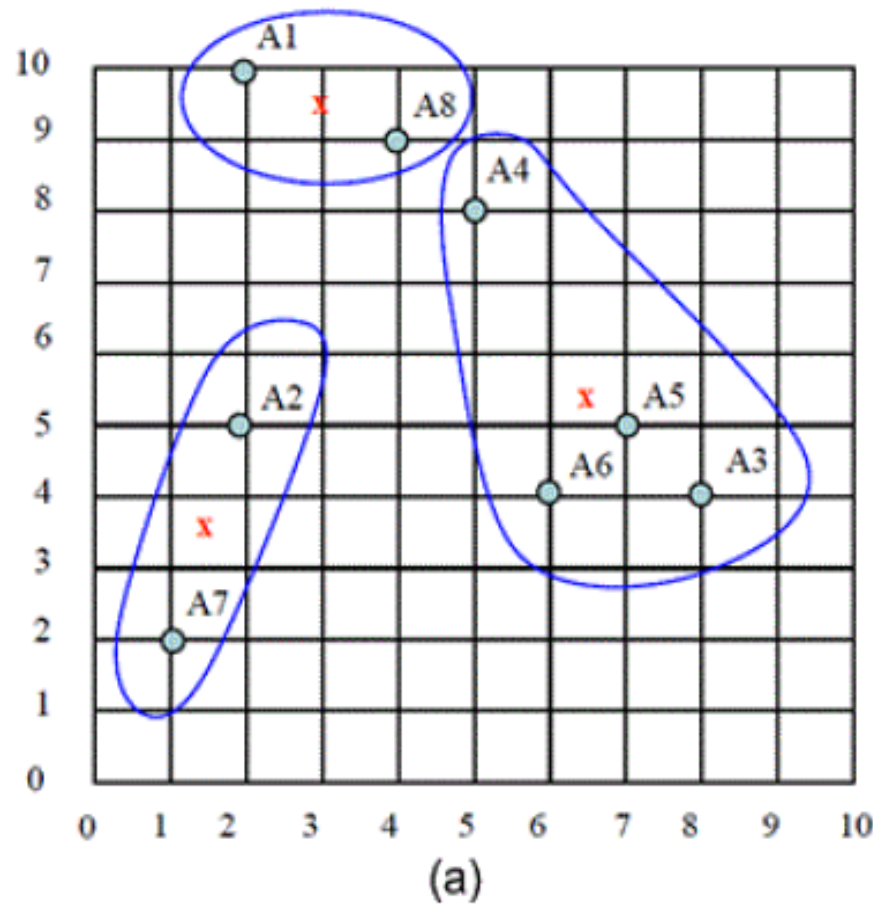
K-Means Clustering

Epoch 2 (Omit), after the 2nd epoch the results would be:

1: {A1, A8}, 2: {A3, A4, A5, A6}, 3: {A2, A7} with centres $C1 = (3, 9.5)$, $C2 = (6.5, 5.25)$ and $C3 = (1.5, 3.5)$

Epoch 3 (Omit), after the 3rd epoch the results would be:

1: {A1, A4, A8}, 2: {A3, A5, A6}, 3: {A2, A7} with centre $C1 = (3.66, 9)$, $C2 = (7, 4.33)$ and $C3 = (1.5, 3.5)$



Strengths and weakness of K-Means Method

- **Strength:** *Efficient*: $O(tkn)$, where n is # objects, k is # clusters, and t is # iterations. Normally, $k, t \ll n$.
- **Weakness**
 - Need to specify k , the *number* of clusters, in advance (there are ways to automatically determine the best k).
 - Sensitive to noisy data and *outliers*
 - Not suitable to discover clusters with *non-convex shapes*