CHAPTER 2

Unsupervised Learning

(Introduction and K-means)

Content

- **Introduction to Unsupervised Learning**
- **K-means clustering**
- Probabilistic clustering via EM algorithm
- Hierarchical clustering
- Unsupervised Learning with Python
- Determine Number of Clusters with Python

INTRODUCTION TO UNSUPERVISED LEARNING

Unsupervised Learning

- Definition of Unsupervised Learning:

  Learning useful structure *without* labeled classes, optimization criterion, feedback signal, or any other information beyond the raw data

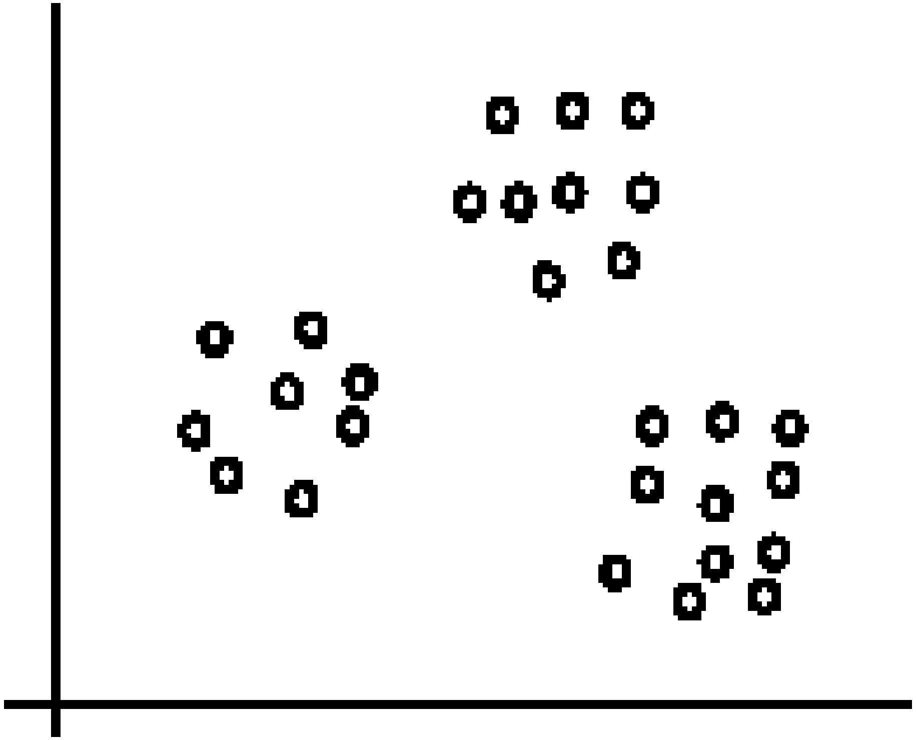
- Methods
  - Clustering (n-link, k-means, GAC,…)
  - Taxonomy creation (hierarchical clustering)
  - Novelty detection ("meaningful"outliers)
  - Trend detection (extrapolation from multivariate partial derivatives)

Clustering

- Clustering is a technique for finding similarity groups in data, called **clusters**. I.e.,
  - it groups data instances that are similar to (near) each other in one cluster and data instances that are very different (far away) from each other into different clusters.
- Clustering is often called an **unsupervised learning** task as no class values denoting an *a priori* grouping of the data instances are given, which is the case in supervised learning.
- Due to historical reasons, clustering is often considered synonymous with unsupervised learning.
  - In fact, association rule mining is also unsupervised
- This chapter focuses on clustering.

An illustration

- The data set has three natural groups of data points, i.e., 3 natural clusters.



What is clustering for?

- Let us see some real-life examples
- Example 1: groups people of similar sizes together to make “small”, “medium” and “large” T-Shirts.
  - Tailor-made for each person: too expensive
  - One-size-fits-all: does not fit all.
- Example 2: In marketing, segment customers according to their similarities
  - To do targeted marketing.

What is clustering for? (cont…)

- Example 3: Given a collection of text documents, we want to organize them according to their content similarities,
  - To produce a topic hierarchy
- In fact, clustering is one of the most utilized data mining techniques.
  - It has a long history, and used in almost every field, e.g., medicine, psychology, botany, sociology, biology, archeology, marketing, insurance, libraries, etc.
  - In recent years, due to the rapid increase of online documents, text clustering becomes important.

Aspects of clustering

- A clustering algorithm
  - Partitional clustering
  - Hierarchical clustering
  - …
- A distance (similarity, or dissimilarity) function
- Clustering quality
  - Inter-clusters distance $\Rightarrow$ maximized
  - Intra-clusters distance $\Rightarrow$ minimized
- The quality of a clustering result depends on the algorithm, the distance function, and the application.

K-MEANS CLUSTERING

K-means clustering

- K-means is a partitional clustering algorithm
- Let the set of data points (or instances) *D* be

  $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$,

  where $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ir})$ is a vector in a real-valued space $X \subseteq R^r$, and r is the number of attributes (dimensions) in the data.
- The *k*-means algorithm partitions the given data into *k* clusters.
  - Each cluster has a cluster **center** c, called **centroid**.
  - *k* is specified by the user

K-means clustering

- The goal of K-means clustering is to minimize the following objective function

$$\min_{I_{ik},\boldsymbol{c}_k} J(I_{ik},\boldsymbol{c}_k), where\ J(I_{ik},\boldsymbol{c}_k) = \sum_{i=1}^{n}\sum_{k=1}^{c} I_{ik}\left|\left|\boldsymbol{x}_i - \boldsymbol{c}_k\right|\right|^2$$

where $I_{ik} = \{0,1\}$ are binary variables and $\boldsymbol{c}_k$ are the cluster centers.

K-means clustering

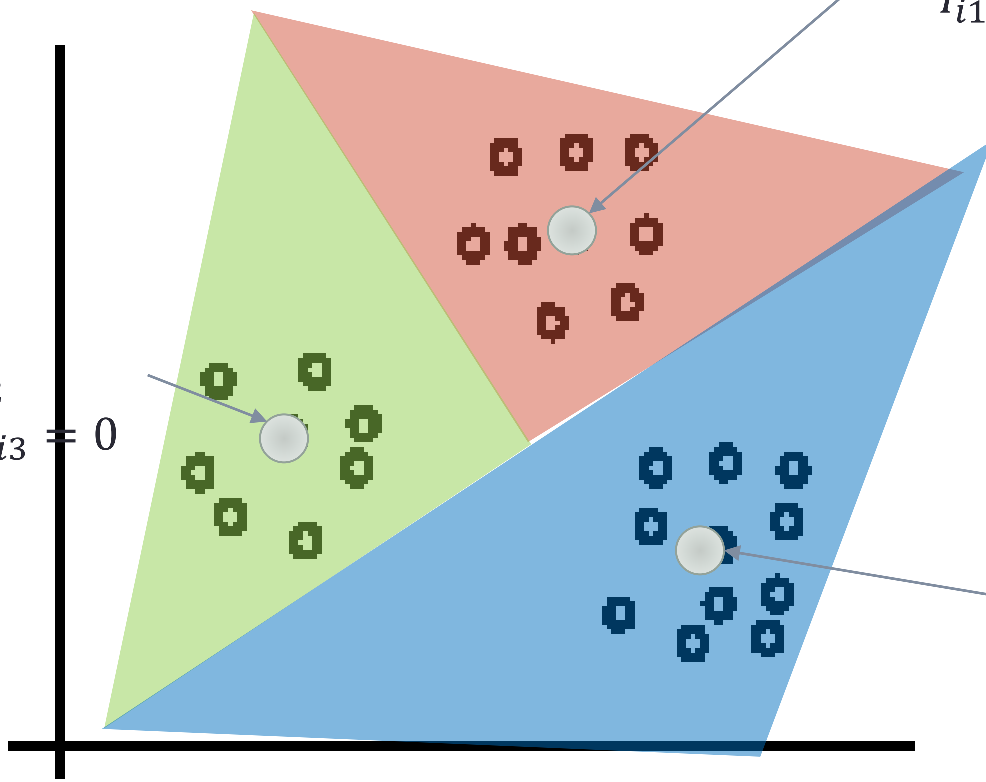
- If there are three clusters, the above problem becomes

$$\min_{I_{ik}, \boldsymbol{c}_k} \sum_{i=1}^{n} \left[I_{i1} ||\boldsymbol{x}_i - \boldsymbol{c}_1||^2 + I_{i2} ||\boldsymbol{x}_i - \boldsymbol{c}_2||^2 + I_{i3} ||\boldsymbol{x}_i - \boldsymbol{c}_3||^2 \right]$$

- That is, we want to find centers $\boldsymbol{c}_1, \boldsymbol{c}_2$ and $\boldsymbol{c}_3$ that are closest to the data samples $\boldsymbol{x}_i$ in three different regions.

K-means clustering

- Graphically illustration of K-means clustering



K-means clustering

How to find a local optimal solution for the k-means clustering problem?

- By the following alternating updating scheme

Step 1: Updating Assignment

Step 2: Updating Centroid

K-means clustering

Step 1: Updating Assignment

- Assign each sample to the closest centroid
- That is,

$$I_{ik} = 1 \text{ if } \left\| \boldsymbol{x}_i - \boldsymbol{c}_k \right\|^2 \leq \left\| \boldsymbol{x}_i - \boldsymbol{c}_j \right\|^2, \text{ for j=1,…c}$$

$$I_{ik} = 0 \quad \text{otherwise} \qquad .$$

Step 2: Updating Centroid

- Compute the centroids by the following formula

$$\boldsymbol{c}_k = \frac{\sum_{i=1}^{n} I_{ik} \boldsymbol{x}_i}{\sum_{i=1}^{n} I_{ik}}$$

K-means clustering

- Why the above two steps?
- Step 1: The update can further minimize the objective function $J(I_{ik}, \boldsymbol{c}_k)$

$$J(I_{ik}, \boldsymbol{c}_k) = \sum_{i=1}^{n} \sum_{k=1}^{c} I_{ik} \left|\left|\boldsymbol{x}_i - \boldsymbol{c}_k\right|\right|^2$$

- Step 2: The formula is obtained by taking first derivative of the objective function $J(I_{ik}, \boldsymbol{c}_k)$ with respect to $\boldsymbol{c}_k$.

K-means clustering

$$\frac{\partial J(I_{ik}, \boldsymbol{c}_k)}{\partial \boldsymbol{c}_k} = 2\sum_{i=1}^{n}\sum_{k=1}^{c} I_{ik}(\boldsymbol{c}_k - \boldsymbol{x}_i)$$

By taking $\frac{\partial J(I_{ik},\boldsymbol{c}_k)}{\partial \boldsymbol{c}_k} = 0$, we have

$$\boldsymbol{c}_k = \frac{\sum_{i=1}^{n} I_{ik}\boldsymbol{x}_i}{\sum_{i=1}^{n} I_{ik}}$$

- This implies that the updated centroids can further minimize the objective function $J(I_{ik}, \boldsymbol{c}_k)$.
- So, by alternating updating the two steps, the objective function can be minimized.
- Because of this, we have the following pseudo code.

K-means algorithm

- Given *k*, the *k-means* algorithm works as follows:
- Pseudo code:
  1) Randomly choose *k* data points (seeds) to be the initial centroids, cluster centers
  2) Assign each data point to the closest centroid
  3) Re-compute the centroids using the current cluster memberships.
  4) If a convergence criterion is not met, go to 2).

Stopping/convergence criterion

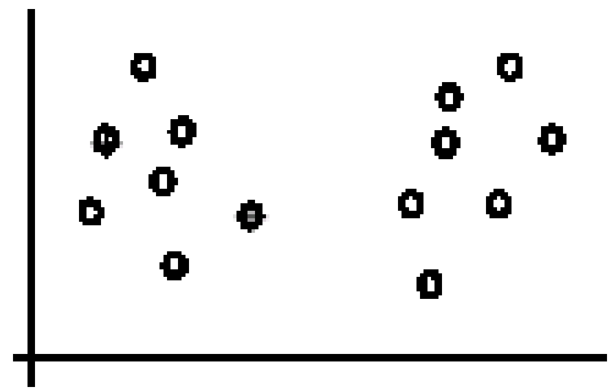
1. No re-assignments of data points to different clusters
    - That is, there is no change of the variables I_{ik} between two iterations
2. No change of centroids
    - That is, there is no change of the variables c_k between two iterations.

Animation of K-means Clustering

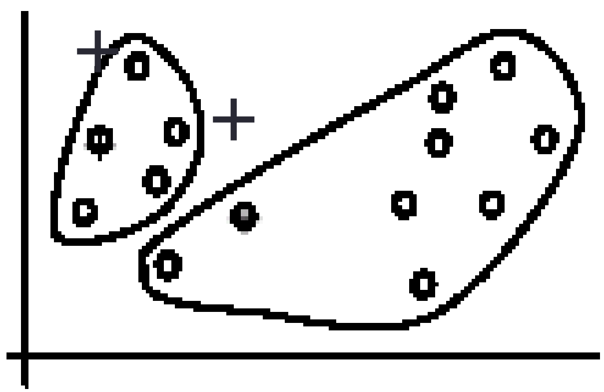
- You may visit the following link for an illustrative animation of K-means clustering:

http://shabal.in/visuals/kmeans/1.html

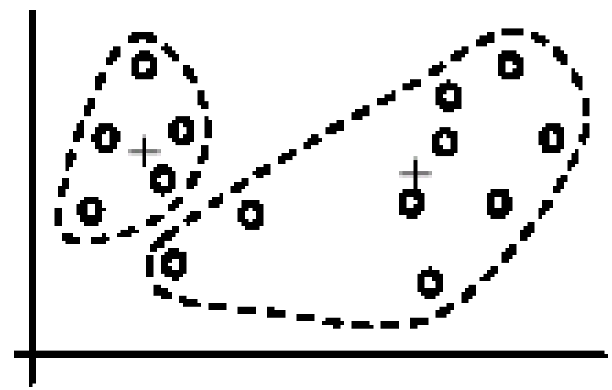
An example



(A). Random selection of *k* centers

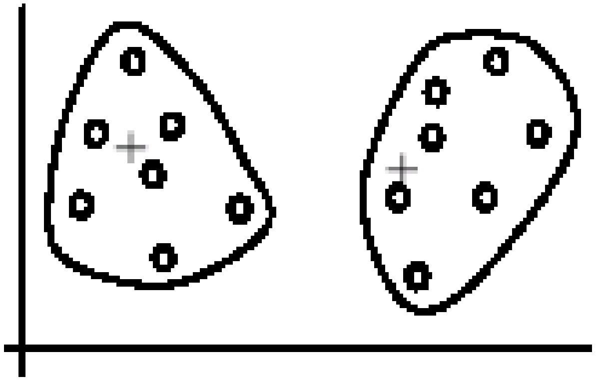


Iteration 1: (B). Cluster assignment

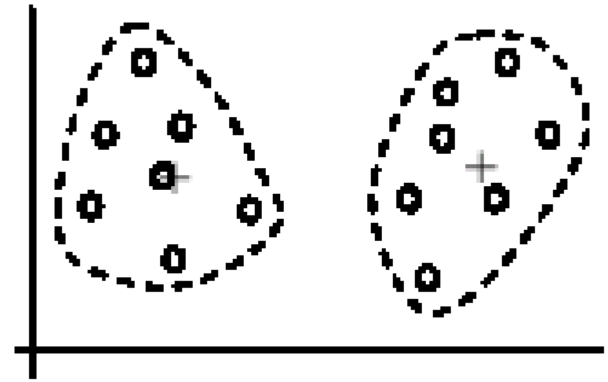


(C). Re-compute centroids

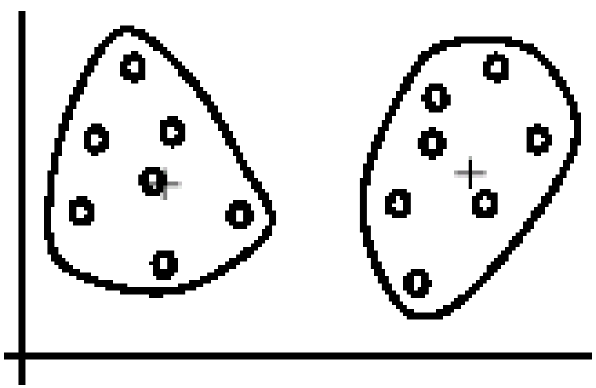
An example (cont …)



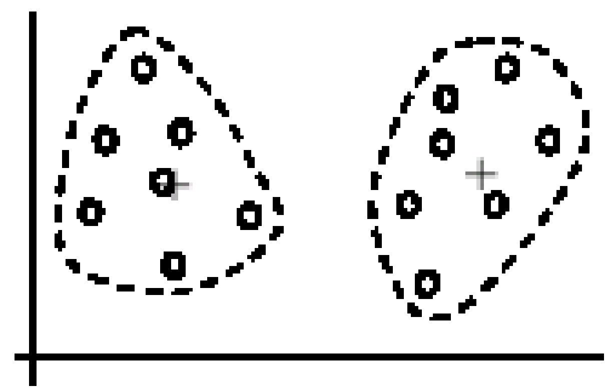
Iteration 2: (D). Cluster assignment



(E). Re-compute centroids



Iteration 3: (F). Cluster assignment



(G). Re-compute centroids

Manual Example 1

- Consider the following data set consisting of the scores of two variables on each of seven individuals:

Subject	A	B
1	1.0	1.0
2	1.5	2.0
3	3.0	4.0
4	5.0	7.0
5	3.5	5.0
6	4.5	5.0
7	3.5	4.5

- Objective: Divide the data into two groups with initial centroids, subject 1 and subject 4.
- [From http://mnemstudio.org/clustering-k-means-example-1.htm]

Manual Example 1

- Answer: The initial centroids are

	Individual	centroid
Group 1	1	(1.0, 1.0)
Group 2	4	(5.0, 7.0)

- Compute the distances between each of the samples and the above two centroids by the following formula:

$$d(x, y) = (x_1 - y_1)^2 + \cdots (x_{\mathrm{D}} - y_{\mathrm{D}})^2$$

where D is the dimension of the samples. In our case, $D = 2$.

Manual Example 1

- Step 1:Updating Assignment
- Distance table

$(1-1.5)^2+(2-1)^2 \;=1.25$

$(5-1.5)^2+(7-2)^2 \;=37.25$

Subject	Distance to Centroid 1	Distance to Centroid 2
1	0	52
2	1.25	37.25
3	13	13
4	52	0
5	22.25	6.25
6	28.25	4.25
7	18.5	8.5

Manual Example 1

- The first three subjects are assigned to centroid 1.
- The last four subjects are assigned to centroid 2.

Closer to Centroid 1

Subject	Distance to Centroid 1	Distance to Centroid 2
1	0	52
2	1.25	37.25
3	13	13
4	52	0
5	22.25	6.25
6	28.25	4.25
7	18.5	8.5

It can
group to centroid 2

Closer to Centroid 2

Manual Example 1

- Step 2: Updating Centroid
- Go back to the original data table
- Select the closest elements of each group and compute the mean

Subject	A	B
1	1.0	1.0
2	1.5	2.0
3	3.0	4.0
4	5.0	7.0
5	3.5	5.0
6	4.5	5.0
7	3.5	4.5

Centroid 1:

A = $\frac{1+1.5+3}{3} = 1.833;$

B= $\frac{1+2+4}{3} = 2.333$

Centroid 2:

A= $\frac{5+3.5+4.5+3.5}{4} = 4.125$

B = $\frac{7+5+5+4.5}{4} = 5.3750$

Manual Example 1

- The above steps finish a single iteration.
- We have to repeat the above steps until there is no change about the centroids or the assignments.

Manual Example 1

- Step 1:Updating Assignment
- Distance table

$$(4.125 - 1.5)^2 + (5.375 - 2)^2 = 18.2813$$

$$(1.833 - 1.5)^2 + (2.33 - 2)^2 = 0.2222$$

Subject	Distance to Centroid 1	Distance to Centroid 2
1	2.4722	28.9063
2	0.2222	18.2813
3	4.1389	3.1563
4	31.8056	3.4063
5	9.8889	0.5313
6	14.2222	0.2813
7	7.4722	1.1563

Manual Example 1

- The first two subjects are assigned to centroid 1.
- The last five subjects are assigned to centroid 2.

Closer to Centroid 1

Subject	Distance to Centroid 1	Distance to Centroid 2
1	2.4722	28.9063
2	0.2222	18.2813
3	4.1389	3.1563
4	31.8056	3.4063
5	9.8889	0.5313
6	14.2222	0.2813
7	7.4722	1.1563

Closer to Centroid 2

Manual Example 1

- Step 2: Updating Centroid
- Go back to the original data table
- Select the closest elements of each group and compute the mean

Subject	A	B
1	1.0	1.0
2	1.5	2.0
3	3.0	4.0
4	5.0	7.0
5	3.5	5.0
6	4.5	5.0
7	3.5	4.5

Centroid 1:

A = $\frac{1+1.5}{2} = 1.25;$

B= $\frac{1+2}{2} = 1.5$

Centroid 2:

A= $\frac{3+5+3.5+4.5+3.5}{5} = 3.9$

B = $\frac{4+7+5+5+4.5}{5} = 5.1$

Manual Example 1

- The above steps finish the second iteration.
- However, the current centroids are different from the previous centroids.
- We have to repeat the above steps until there is no change about the centroids or the assignments.

Current Centroid 1:

A = $\frac{1+1.5}{2} = 1.25;$

B= $\frac{1+2}{2} = 1.5$

Current Centroid 2:

A= $\frac{3+5+3.5+4.5+3.5}{5} = 3.9$

B = $\frac{4+7+5+5+4.5}{5} = 5.1$

Previous Centroid 1:

A = $\frac{1+1.5+3}{3} = 1.833;$

B= $\frac{1+2+4}{3} = 2.333$

Previous Centroid 2:

A= $\frac{5+3.5+4.5+3.5}{4} = 4.5$

B = $\frac{7+5+5+4.5}{4} = 5.3750$

Manual Example 1

- Step 1:Updating Assignment
- Distance table

$$(3.9 - 1.5)^2 + (5.1 - 2)^2 \ = 15.37$$

$$(1.25 - 1.5)^2 + (1.5 - 2)^2 \ = 0.3125$$

Subject	Distance to Centroid 1	Distance to Centroid 2
1	0.3125	25.2200
2	0.3125	15.3700
3	9.3125	2.0200
4	44.3125	4.8200
5	17.3125	0.1700
6	22.8125	0.3700
7	14.0625	0.5200

Manual Example 1

- The first two subjects are assigned to centroid 1.
- The last five subjects are assigned to centroid 2.

Closer to Centroid 1

Subject	Distance to Centroid 1	Distance to Centroid 2
1	0.3125	25.2200
2	0.3125	15.3700
3	9.3125	2.0200
4	44.3125	4.8200
5	17.3125	0.1700
6	22.8125	0.3700
7	14.0625	0.5200

Closer to Centroid 2

Manual Example 1

- Step 2: Updating Centroid
- Go back to the original data table
- Select the closest elements of each group and compute the mean

Subject	A	B
1	1.0	1.0
2	1.5	2.0
3	3.0	4.0
4	5.0	7.0
5	3.5	5.0
6	4.5	5.0
7	3.5	4.5

Centroid 1:

A = $\frac{1+1.5}{2} = 1.25;$

B= $\frac{1+2}{2} = 1.5$

Centroid 2:

A= $\frac{3+5+3.5+4.5+3.5}{5} = 3.9$

B = $\frac{4+7+5+5+4.5}{5} = 5.1$

Manual Example 1

- The above steps finish the second iteration.
- However, the current centroids are the same as previous centroids.
- We stop here and output the solutions.

Current	Previous
Current Centroid 1: $A = \frac{1+1.5}{2} = 1.25$; $B= \frac{1+2}{2} = 1.5$ Current Centroid 2: $A= \frac{3+5+3.5+4.5+3.5}{5} = 3.9$ $B = \frac{4+7+5+5+4.5}{5} = 5.1$	Previous Centroid 1: $A = \frac{1+1.5}{2} = 1.25$; $B= \frac{1+2}{2} = 1.5$ Current Centroid 2: $A= \frac{3+5+3.5+4.5+3.5}{5} = 3.9$ $B = \frac{4+7+5+5+4.5}{5} = 5.1$

Manual Example 2

- Consider the following data set consisting of the scores of two variables on each of five individuals:

Subject	A	B
1	7.0	10.0
2	1.0	10.0
3	6.0	9.0
4	3.0	5.0
5	3.0	2.0

- Objective: Divide the data into two groups with initial centroids, subject 1 and subject 5.

Manual Example 2

- Answer: The initial centroids are

	Individual	centroid
Group 1	1	(7.0, 10.0)
Group 2	5	(3.0, 2.0)

- Compute the distances between each of the samples and the above two centroids by the following formula:

$$d(x, y) = (x_1 - y_1)^2 + \cdots (x_{\mathrm{D}} - y_{\mathrm{D}})^2$$

where D is the dimension of the samples. In our case, $D = 2$.

Manual Example 2

- Step 1:Updating Assignment
- Distance table

$$(7-1)^2+(10-10)^2 \;=36$$

$$(3-1)^2+(2-10)^2 \;=68$$

Subject	Distance to Centroid 1	Distance to Centroid 2
1	0	80
2	36	68
3	2	58
4	41	9
5	80	0

Manual Example 2

- The subjects 1, 2 & 4 are assigned to centroid 1.
- The subjects 3 & 5 are assigned to centroid 2.

Closer to Centroid 1

Subject	Distance to Centroid 1	Distance to Centroid 2
1	0	80
2	36	68
3	2	58
4	41	9
5	80	0

Closer to Centroid 2

Manual Example 2

- Step 2: Updating Centroid
- Go back to the original data table
- Select the closest elements of each group and compute the mean

Subject	A	B
1	7.0	10.0
2	1.0	10.0
3	6.0	9.0
4	3.0	5.0
5	3.0	2.0

Centroid 1:

A = $\frac{7+1+6}{3} = 4.667;$

B= $\frac{10+10+9}{3} = 9.6667$

Centroid 2:

A= $\frac{3+3}{2} = 3$

B = $\frac{5+2}{2} = 3.5$

Manual Example 2

- The above steps finish a single iteration.
- We have to repeat the above steps until there is no change about the centroids or the assignments.

Manual Example 2

- Step 1:Updating Assignment
- Distance table
- The subjects 1,2 & 4 are assigned to centroid 1.
- The subjects 3 & 5 are assigned to centroid 2.

Subject	Distance to Centroid 1	Distance to Centroid 2
1	5.55378	58.25
2	13.55578	46.25
3	2.222178	39.25
4	24.55598	2.25
5	61.55618	2.25

Manual Example 2

- Step 2: Updating Centroid
- Go back to the original data table
- Select the closest elements of each group and compute the mean

Subject	A	B
1	7.0	10.0
2	1.0	10.0
3	6.0	9.0
4	3.0	5.0
5	3.0	2.0

Centroid 1:

A = $\frac{7+1+6}{3} = 4.667;$

B= $\frac{10+10+9}{3} = 9.6667$

Centroid 2:

A= $\frac{3+3}{2} = 3$

B = $\frac{5+2}{2} = 3.5$

Manual Example 2

- The above steps finish the second iteration.
- However, the current centroids are the same as previous centroids.
- We stop here and output the solutions.

Current
Centroid 1:
A = $\frac{7+1+6}{3} = 4.667;$
B= $\frac{10+10+9}{3} = 9.6667$
Centroid 2:
A= $\frac{3+3}{2} = 3$
B = $\frac{5+2}{2} = 3.5$

Previous
Centroid 1:
A = $\frac{7+1+6}{3} = 4.667;$
B= $\frac{10+10+9}{3} = 9.6667$
Centroid 2:
A= $\frac{3+3}{2} = 3$
B = $\frac{5+2}{2} = 3.5$

Strengths of k-means

- Strengths:
  - Simple: easy to understand and to implement
  - Efficient: Time complexity: $O(tkn)$,

    where n is the number of data points,

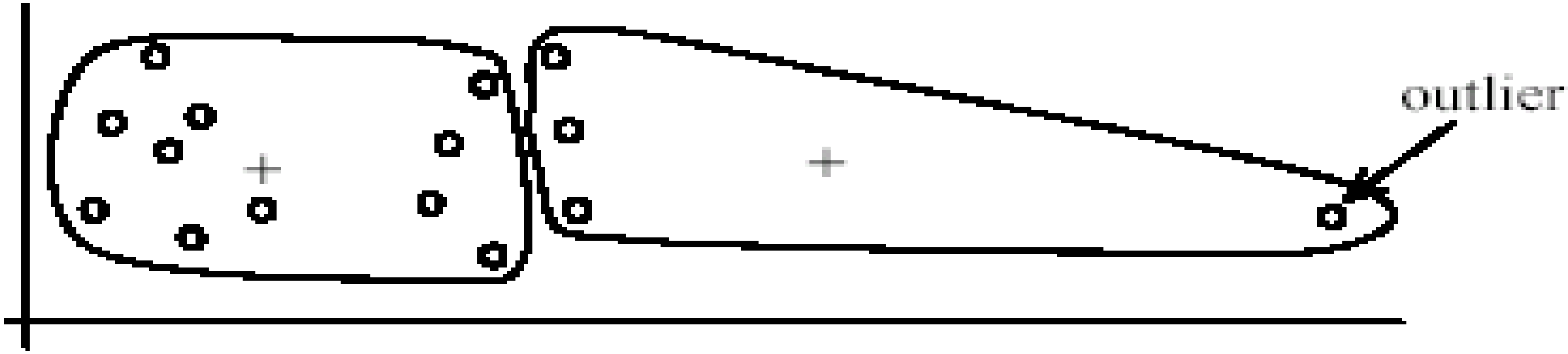
    k is the number of clusters, and

    t is the number of iterations.
  - Since both k and t are small. k-means is considered a linear algorithm.
- K-means is the most popular clustering algorithm.
- Note that: it terminates at a local optimum if the sum of square error is used. The global optimum is hard to find due to complexity.

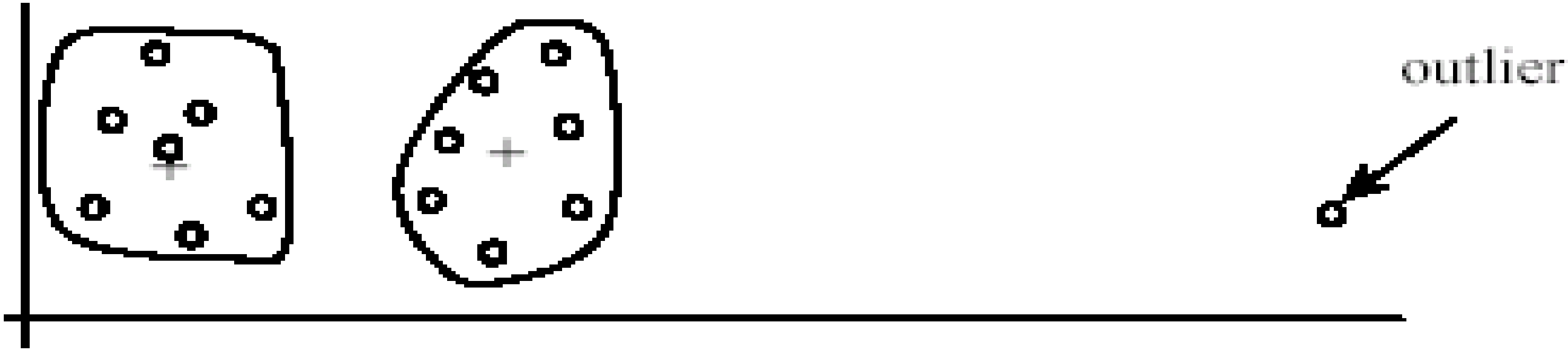
Weaknesses of k-means

- 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



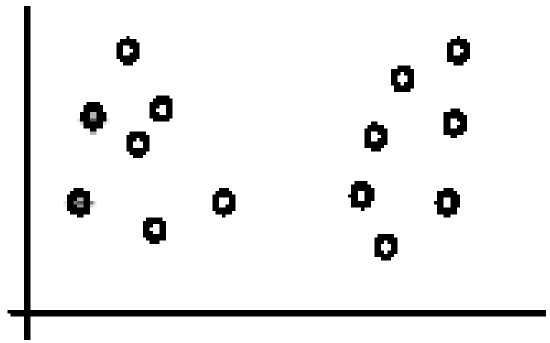
(B): Ideal clusters

Weaknesses of k-means: To deal with outliers

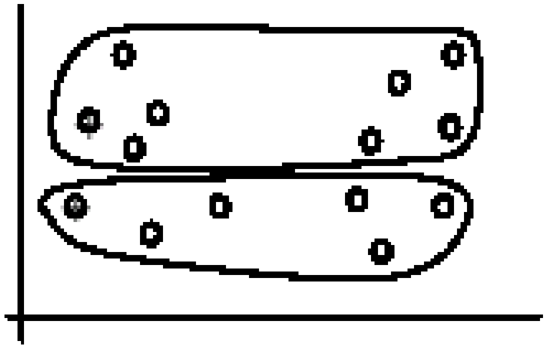
- One method is to remove some data points in the clustering process that are much further away from the centroids than other data points.
  - To be safe, we may want to monitor these possible outliers over a few iterations and then decide to remove them.
- Another method is to perform random sampling. 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 (cont …)

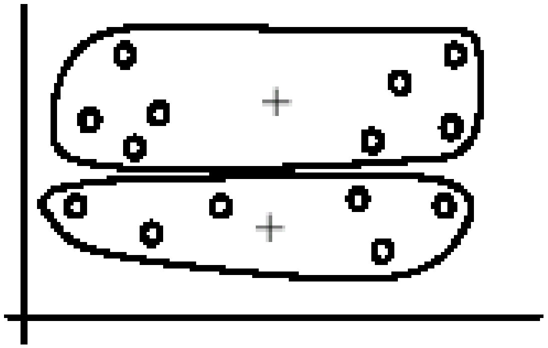
- The algorithm is sensitive to initial seeds.



(A). Random selection of seeds (centroids)



(B). Iteration 1

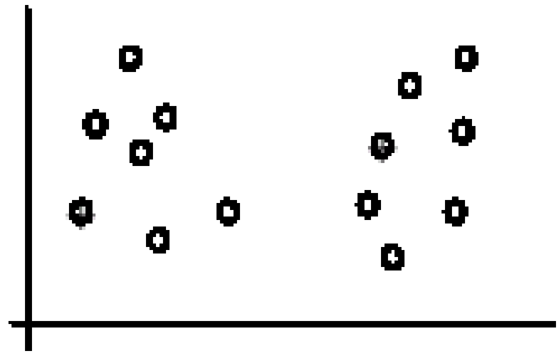


(C). Iteration 2

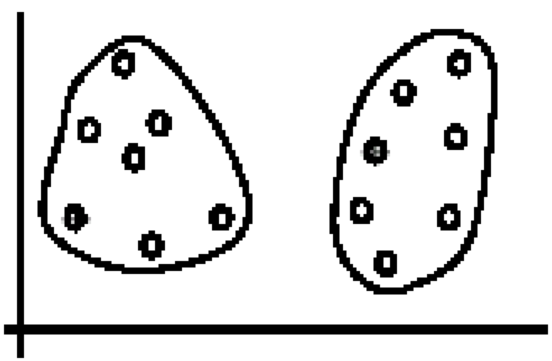
Weaknesses of k-means (cont …)

- If we use different seeds: good results

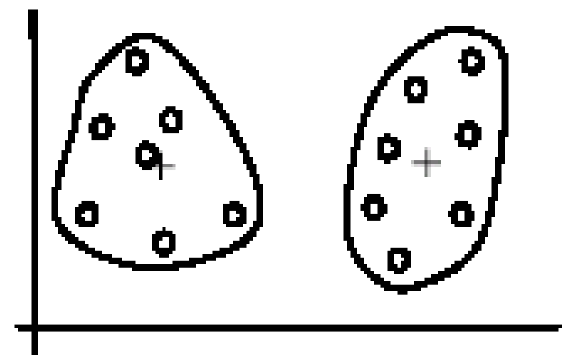
There are some methods to help choose good seeds



(A). Random selection of *k* seeds (centroids)



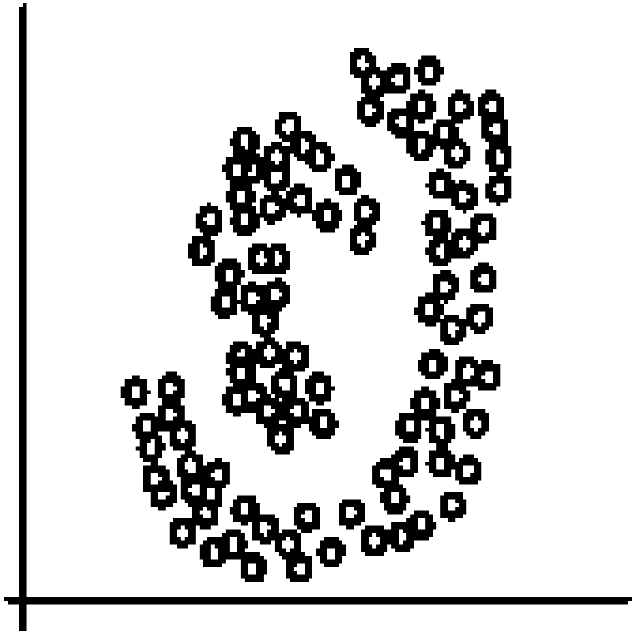
(B). Iteration 1



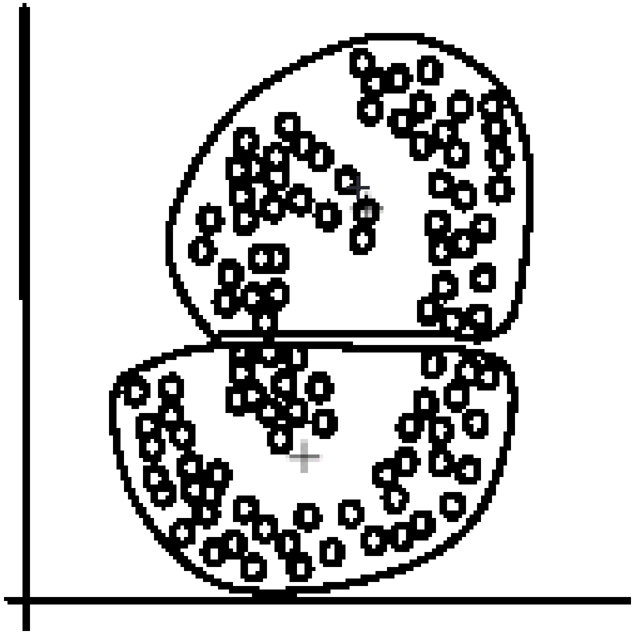
(C). Iteration 2

Weaknesses of k-means (cont …)

- 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

Variants of K-means clustering

- K-means clustering employs the following objective function

$$J(I_{ik}, \boldsymbol{c}_k) = \sum_{i=1}^{n} \sum_{k=1}^{c} I_{ik} \left|\left|\boldsymbol{x}_i - \boldsymbol{c}_k\right|\right|^2$$

- The term $\left|\left|x\right|\right|^2$ is also known as squared l_2 distance.
- Variants of K-means clustering were developed based on the use of different types of distance.

Distance functions

- Most commonly used functions are
  - Euclidean distance and
  - Manhattan (city block) distance
- We denote distance with: $dist(\mathbf{x}_i, \mathbf{x}_j)$, where $\mathbf{x}_i$ and $\mathbf{x}_j$ are data points (vectors)
- They are special cases of Minkowski distance. h is positive integer.

$$dist(\mathbf{x}_i, \mathbf{x}_j) = ((x_{i1} - x_{j1})^h + (x_{i2} - x_{j2})^h + ... + (x_{ir} - x_{jr})^h)^{\frac{1}{h}}$$

Euclidean distance and Manhattan distance

- If h = 2, it is the Euclidean distance

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

- If h = 1, it is the Manhattan distance

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

- Weighted Euclidean distance

$$dist(\mathbf{x}_i, \mathbf{x}_j) = \sqrt{w_1(x_{i1} - x_{j1})^2 + w_2(x_{i2} - x_{j2})^2 + ... + w_r(x_{ir} - x_{jr})^2}$$

Squared distance and Chebychev distance

- **Squared Euclidean distance:** to place progressively greater weight on data points that are further apart.

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

- **Chebychev distance:** one wants to define two data points as "different" if they are different on any one of the attributes.

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

$Manhattan\ distance\ based$
$K-means\ Clustering$

- Illustrate the use of Manhattan distance in K-means clustering.
- Replacing the squared l_2 distance in K-means objective function by Manhattan distance, we have

$$J(I_{ik}, \boldsymbol{c}_k) = \sum_{i=1}^{n} \sum_{k=1}^{c} I_{ik} ||\boldsymbol{x}_i - \boldsymbol{c}_k||_1$$

where $||x||_1$ is the Manhattan distance.

$Manhattan\ distance\ based$ $K - means\ Clustering$

- Strategy in solving this type of K-means clustering algorithm is similar to that of the classical K-means clustering.
- It employs the alternative updating scheme:

1. Updating the assignment
2. Updating the centroids

Manual Example 1

- Consider the following data set consisting of the scores of two variables on each of seven individuals:

Subject	A	B
1	1.0	1.0
2	1.5	2.0
3	3.0	4.0
4	5.0	7.0
5	3.5	5.0
6	4.5	5.0
7	3.5	4.5

- Objective: Divide the data into two groups with initial centroids, subject 1 and subject 4 using Manhattan distance based K-means Clustering.
- [From http://mnemstudio.org/clustering-k-means-example-1.htm]

Manual Example 1

- Answer: The initial centroids are

	Individual	centroid
Group 1	1	(1.0, 1.0)
Group 2	4	(5.0, 7.0)

- Compute the distances between each of the samples and the above two centroids by the following formula:

$$d(x, y) = |x_1 - y_1| + \cdots |x_{\mathrm{D}} - y_{\mathrm{D}}|$$

where D is the dimension of the samples. In our case, $D = 2$.

Manual Example 1

- Step 1:Updating Assignment
- Distance table

$$|5 - 1.5| + |7 - 2| \; = 8.5$$

$$|1 - 1.5| + |2 - 1| \; = 1.5$$

Subject	Distance to Centroid 1	Distance to Centroid 2
1	0	10
2	1.5	8.5
3	5	5
4	10	0
5	6.5	3.5
6	7.5	2.5
7	6	4

Manual Example 1

- The first three subjects are assigned to centroid 1.
- The last four subjects are assigned to centroid 2.

Closer to Centroid 1

Subject	Distance to Centroid 1	Distance to Centroid 2
1	0	10
2	1.5	8.5
3	5	5
4	10	0
5	6.5	3.5
6	7.5	2.5
7	6	4

It can
group to centroid 2

Closer to Centroid 2

Manual Example 1

- Step 2: Updating Centroid
- Go back to the original data table
- Select the closest elements of each group and compute the mean

Subject	A	B
1	1.0	1.0
2	1.5	2.0
3	3.0	4.0
4	5.0	7.0
5	3.5	5.0
6	4.5	5.0
7	3.5	4.5

Centroid 1:
A = median(1,1.5,3) $= 1.5;$
B= $median(1,2,4) = 2$
Centroid 2:
A= $median(3.5,3.5,4.5,5) = 4$
B = $\mathrm{median}(4.5,5,5,7) = 5$

Why median here?
Will explain later

Manual Example 1

- The above steps finish a single iteration.
- We have to repeat the above steps until there is no change about the centroids or the assignments.

Manual Example 1

- Step 1:Updating Assignment
- Distance table

$$|4 - 1.5| + |5 - 2| \ = 5.5$$

$$|1.5 - 1.5| + |2 - 2| \ = 0$$

Subject	Distance to Centroid 1	Distance to Centroid 2
1	1.5	7
2	0	5.5
3	3.5	2
4	8.5	3
5	5	0.5
6	6	0.5
7	4	1

Manual Example 1

- The first two subjects are assigned to centroid 1.
- The last five subjects are assigned to centroid 2.

Closer to Centroid 1

Subject	Distance to Centroid 1	Distance to Centroid 2
1	1.5	7
2	0	5.5
3	3.5	2
4	8.5	3
5	5	0.5
6	6	0.5
7	4	1

Closer to Centroid 2

Manual Example 1

- Step 2: Updating Centroid
- Go back to the original data table
- Select the closest elements of each group and compute the mean

Subject	A	B
1	1.0	1.0
2	1.5	2.0
3	3.0	4.0
4	5.0	7.0
5	3.5	5.0
6	4.5	5.0
7	3.5	4.5

Centroid 1:
A = median(1,1.5) $= 1.25;$
B$= median(1,2) = 1.5$
Centroid 2:
A$= median(3,3.5,3.5,4.5,5) = 3.5$
B = $\mathrm{median}(4.0,4.5,5,5,7) = 5$

Manual Example 1

- The above steps finish the second iteration.
- However, the current centroids are different from the previous centroids.
- We have to repeat the above steps until there is no change about the centroids or the assignments.

Current
A = median(1,1.5) $= 1.25$;
B= $median(1,2) = 1.5$
Centroid 2:
A= $median(3,3.5,3.5,4.5,5) = 3.5$
B = $\mathrm{median}(4.0,4.5,5,5,7) = 5$

Previous
A = median(1,1.5,3) $= 1.5$;
B= $median(1,2,4) = 2$
Centroid 2:
A= $median(3.5,3.5,4.5,5) = 4$
B = $\mathrm{median}(4.5,5,5,7) = 5$

Manual Example 1

- Step 1:Updating Assignment
- Distance table

$$|3.5 - 1.5| + |5.0 - 2| = 5$$

$$|1.25 - 1.5| + |1.5 - 2| = 0.75$$

Subject	Distance to Centroid 1	Distance to Centroid 2
1	0.75	6.5
2	0.75	5
3	4.25	1.5
4	9.25	3.5
5	5.75	0
6	6.75	1
7	5.25	0.5

Manual Example 1

- The first two subjects are assigned to centroid 1.
- The last five subjects are assigned to centroid 2.

Closer to Centroid 1

Subject	Distance to Centroid 1	Distance to Centroid 2
1	0.75	6.5
2	0.75	5
3	4.25	1.5
4	9.25	3.5
5	5.75	0
6	6.75	1
7	5.25	0.5

Closer to Centroid 2

Manual Example 1

- Step 2: Updating Centroid
- Go back to the original data table
- Select the closest elements of each group and compute the mean

Subject	A	B
1	1.0	1.0
2	1.5	2.0
3	3.0	4.0
4	5.0	7.0
5	3.5	5.0
6	4.5	5.0
7	3.5	4.5

Centroid 1:
A = median(1,1.5) $= 1.25;$
B$= median(1,2) = 1.5$
Centroid 2:
A$= median(3,3.5,3.5,4.5,5) = 3.5$
B = $\mathrm{median}(4.0,4.5,5,5,7) = 5$

Manual Example 1

- The above steps finish the third iteration.
- However, the current centroids are the same as the previous centroids.
- We have to repeat the above steps until there is no change about the centroids or the assignments.

Current
Centroid 1:
A = median(1,1.5) $= 1.25$;
B= $median(1,2) = 1.5$
Centroid 2:
A= $median(3,3.5,3.5,4.5,5) = 3.5$
B = $\mathrm{median}(4.0,4.5,5,5,7) = 5$

Previous
Centroid 1:
A = median(1,1.5) $= 1.25$;
B= $median(1,2) = 1.5$
Centroid 2:
A= $median(3,3.5,3.5,4.5,5) = 3.5$
B = $\mathrm{median}(4.0,4.5,5,5,7) = 5$

Updating the Centroids for the Manhattan distance $based\ K - means\ Clustering$ [Reference]

- Now, we explain why the updating equation for the centroids are the median of the data.
- In the classical K-means clustering algorithm, the updating formula is obtained by taking the first derivative of the objective function with respect to the cluster centers c_k.
- It is basically the same for the case of Mahnhattan distance clustering.

Updating the Centroids for the Manhattan distance $based\ K - means\ Clustering$ [Reference]

- Recall that the objective function is

$$J(I_{ik}, \boldsymbol{c}_k) = \sum_{i=1}^{n} \sum_{k=1}^{c} I_{ik} ||\boldsymbol{x}_i - \boldsymbol{c}_k||_1$$

It is noted that the minimum of

$$\sum_{i=1}^{\mathrm{n}} |x_i - c|$$

Is attained at

$$c = median\{x_i\}$$

Manual Example 2

- Consider the following data set consisting of the scores of two variables on each of five individuals:

Subject	A	B
1	7.0	10.0
2	1.0	10.0
3	6.0	9.0
4	3.0	5.0
5	3.0	2.0

- Objective: Divide the data into two groups with initial centroids, subject 1 and subject 5 using Manhattan distance based K-means Clustering.

Manual Example 2

- Answer: We can follow exactly the same procedure as in Example 1 to obtain the solution.
- At first, we have the following distance table

Subject	Distance to Centroid 1	Distance to Centroid 2
1	0	12
2	6	10
3	2	10
4	9	3
5	12	0

- The new cluster centers are (6,10) and (3,3.5).

Manual Example 2

- Next, the distance table is updated as below

Subject	Distance to Centroid 1	Distance to Centroid 2
1	1	10.5
2	5	8.5
3	1	8.5
4	8	1.5
5	11	1.5

- The new cluster centers are (6,10) and (3,3.5).
- There is no change in cluster centers. So, it converges. We can output the solutions.

K-means summary

- Despite weaknesses, *k*-means is still the most popular algorithm due to its simplicity, efficiency and
  - other clustering algorithms have their own lists of weaknesses.
- No clear evidence that any other clustering algorithm performs better in general
  - although they may be more suitable for some specific types of data or applications.
- Comparing different clustering algorithms is a difficult task. No one knows the correct clusters!