Machine Learning Week 8

outline

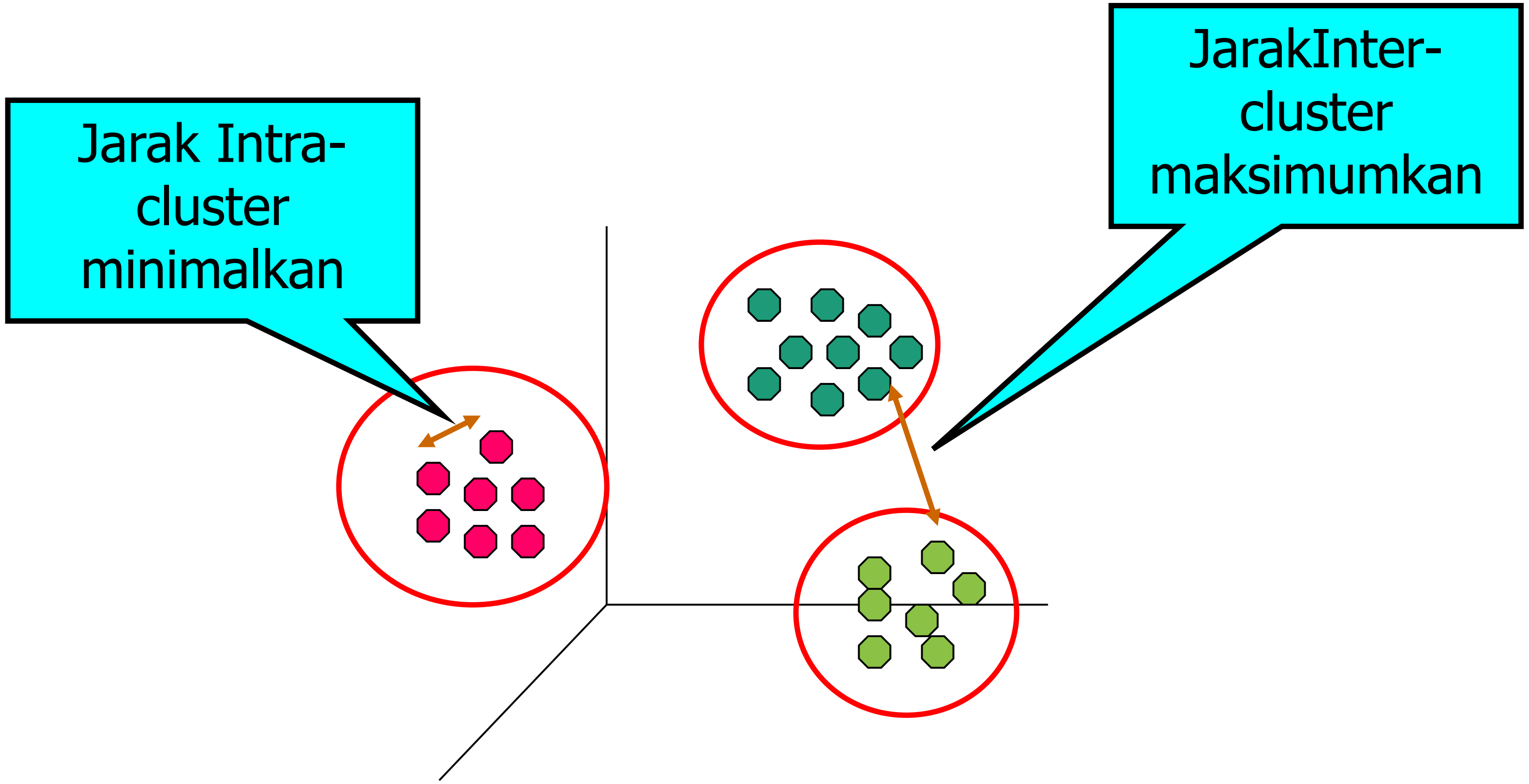
Week	Topics
8	Unsupervised Learning: Clustering 1- K Means Clustering
9	Unsupervised Learning: Clustering II – Hierarchical Clustering
10	Unsupervised Learning: EM Algorithm
11	Pengujian Unsupervised Learning
12	Reinforcement Learning
13	Feature reduction
14	Ensemble Learning

Clustering
- Unsupervised Learning -

Clustering - Basic Concept

- Cluster analysis groups data objects based only on information found in the data that describes the objects and their relationships.
- The goal is that the objects within a group be similar (or related) to one another and different from (or unrelated to) the objects in other groups.
- The greater the similarity (or homogeneity) within a group and the greater the difference between groups, the better or more distinct the clustering.

Clustering - Basic Concept (cont.)



Jarak Intra-cluster minimalkan

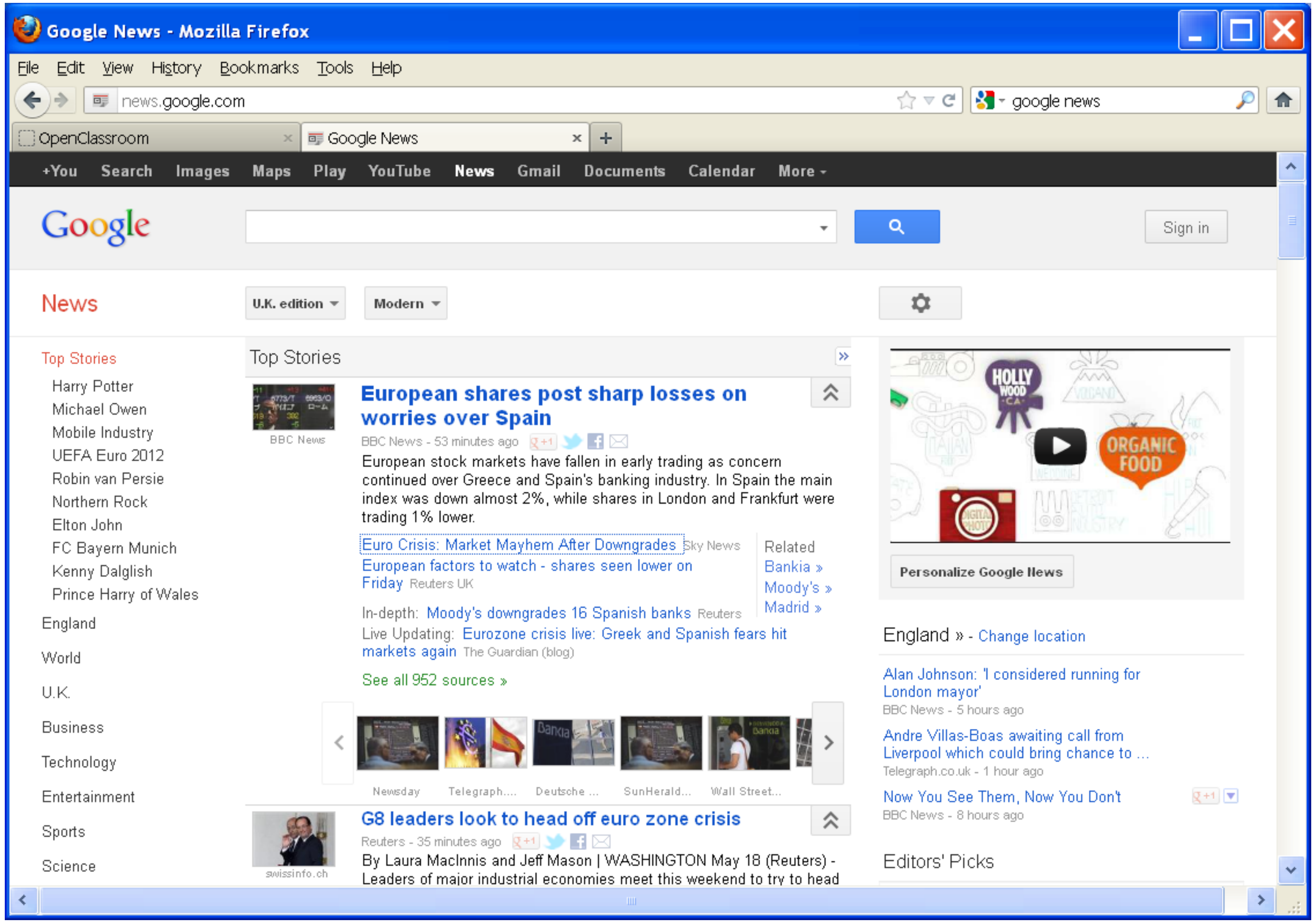
JarakInter-cluster maksimumkan

Clustering - Basic Concept (cont.)

- Clustering can be regarded as a form of classification in that **it creates a labeling of objects** with class (cluster) labels.
- However, it derives these labels only from the data. For this reason, cluster analysis is sometimes referred to as **unsupervised classification**.

Application

- Real Applications: Google News


Google News - Mozilla Firefox
File Edit View History Bookmarks Tools Help
news.google.com
google news
OpenClassroom
Google News
+You Search Images Maps Play YouTube News Gmail Documents Calendar More
Google
Sign in
News
U.K. edition
Modern
Personalize Google News

Top Stories
Harry Potter
Michael Owen
Mobile Industry
UEFA Euro 2012
Robin van Persie
Northern Rock
Elton John
FC Bayern Munich
Kenny Dalglish
Prince Harry of Wales
England
World
U.K.
Business
Technology
Entertainment
Sports
Science

Top Stories

European shares post sharp losses on worries over Spain
BBC News - 53 minutes ago
European stock markets have fallen in early trading as concern continued over Greece and Spain's banking industry. In Spain the main index was down almost 2%, while shares in London and Frankfurt were trading 1% lower.

Euro Crisis: Market Mayhem After Downgrades Sky News
European factors to watch - shares seen lower on Friday Reuters UK
In-depth: Moody's downgrades 16 Spanish banks Reuters
Live Updating: Eurozone crisis live: Greek and Spanish fears hit markets again The Guardian (blog)
See all 952 sources »

Related
Bankia »
Moody's »
Madrid »

Newsday Telegraph.... Deutsche ... SunHerald... Wall Street...

G8 leaders look to head off euro zone crisis
Reuters - 35 minutes ago
By Laura MacInnis and Jeff Mason | WASHINGTON May 18 (Reuters) - Leaders of major industrial economies meet this weekend to try to head

England » - Change location

Alan Johnson: 'I considered running for London mayor'
BBC News - 5 hours ago

Andre Villas-Boas awaiting call from Liverpool which could bring chance to ...
Telegraph.co.uk - 1 hour ago

Now You See Them, Now You Don't
BBC News - 8 hours ago

Editors' Picks

Application

- A technique demanded by many real world tasks
  - Bank/Internet Security: fraud/spam pattern discovery
  - Biology: taxonomy of living things such as kingdom, phylum, class, order, family, genus and species
  - City-planning: Identifying groups of houses according to their house type, value, and geographical location
  - Climate change: understanding earth climate, find patterns of atmospheric and ocean
  - Finance: stock clustering analysis to uncover correlation underlying shares
  - Image Compression/segmentation: coherent pixels grouped
  - Information retrieval/organisation: Google search, topic-based news
  - Land use: Identification of areas of similar land use in an earth observation database
  - Marketing: Help marketers discover distinct groups in their customer bases, and then use this knowledge to develop targeted marketing programs
  - Social network mining: special interest group automatic discovery

Clustering - Basic Concept (cont.)

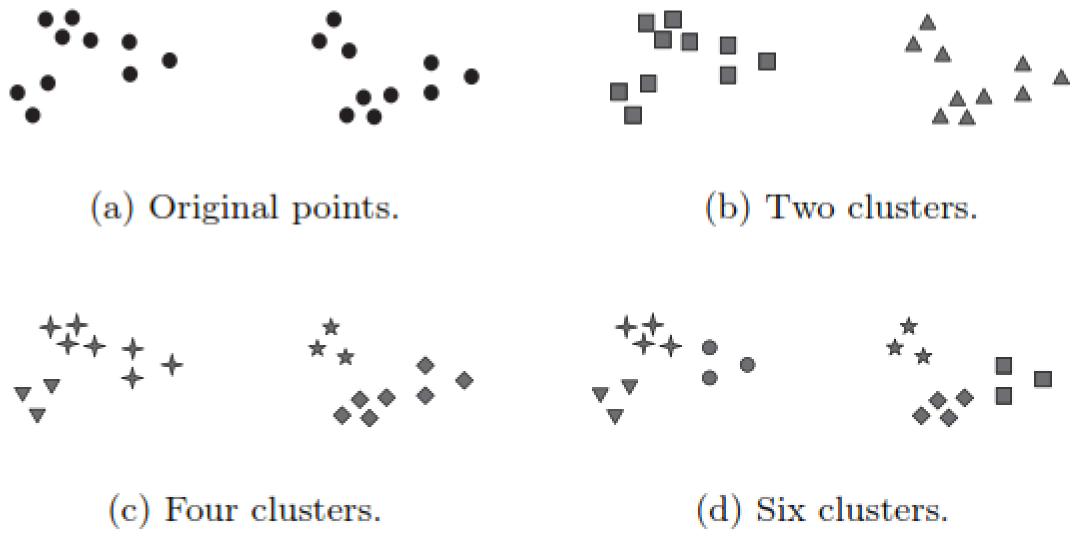


Figure 8.1. Different ways of clustering the same set of points.

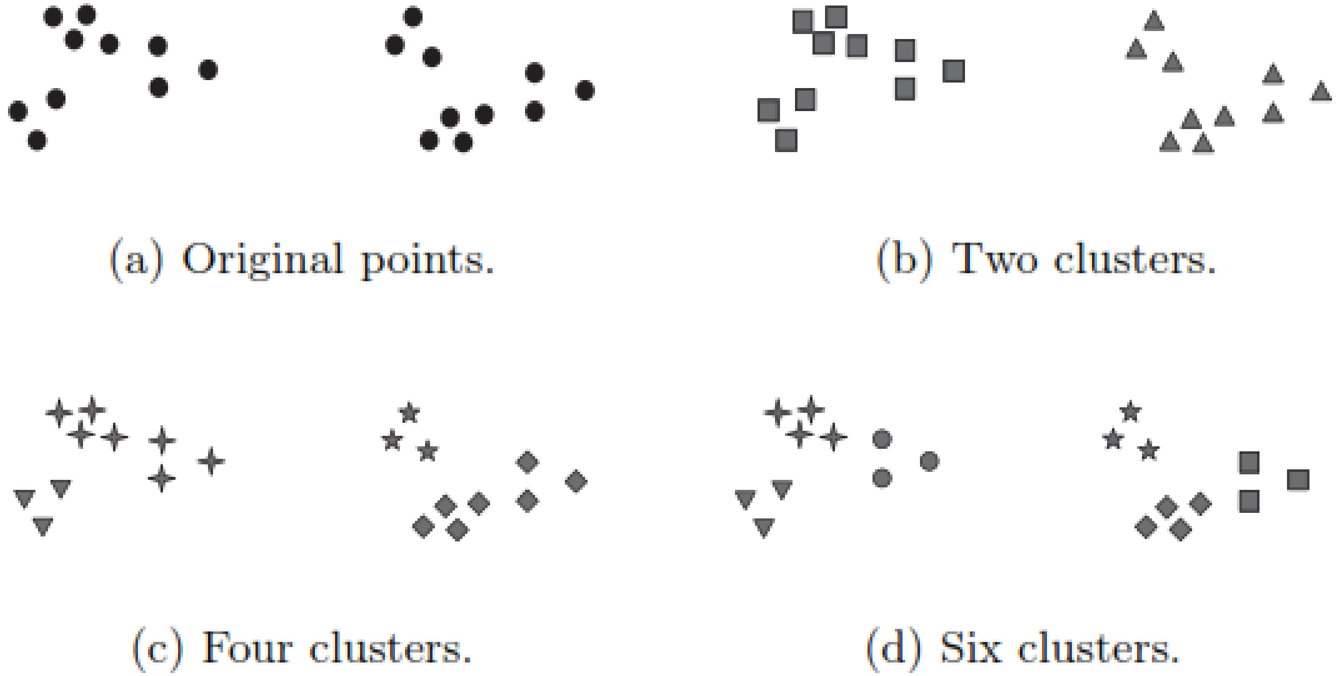
- This figure illustrates that the definition of a cluster is imprecise and that the best definition depends on the nature of data and the desired results.

Different Types of Clusterings

- Hierarchical versus Partitional
- Exclusive versus Overlapping versus Fuzzy
- Complete versus Partial

Hierarchical versus Partitional

- A **partitional clustering** is simply a division of the set of data objects into non-overlapping subsets (clusters) such that each data object is in exactly one subset.
- If we permit clusters to have subclusters, then we obtain a **hierarchical clustering**, which is a set of nested clusters that are organized as a tree.
- Each node (cluster) in the tree (except for the leaf nodes) is the union of its children (subclusters), and the root of the tree is the cluster containing all the objects.
- Often, but not always, the leaves of the tree are singleton clusters of individual data objects.



(a) Original points.

(b) Two clusters.

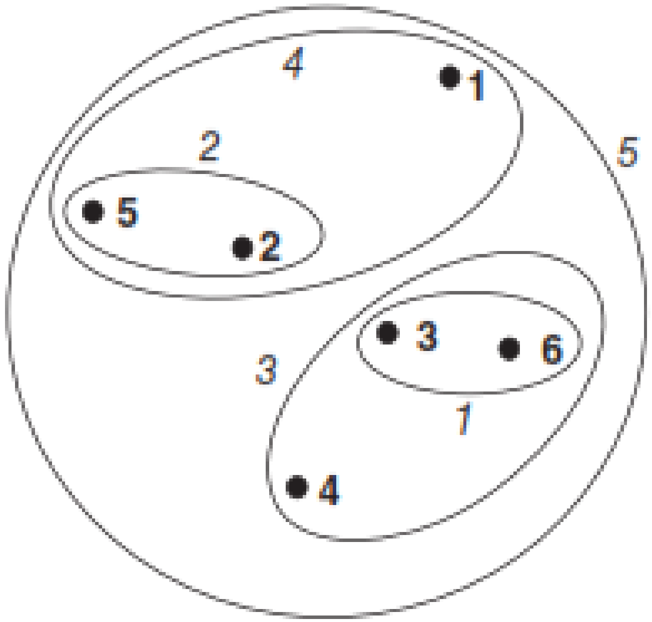
(c) Four clusters.

(d) Six clusters.

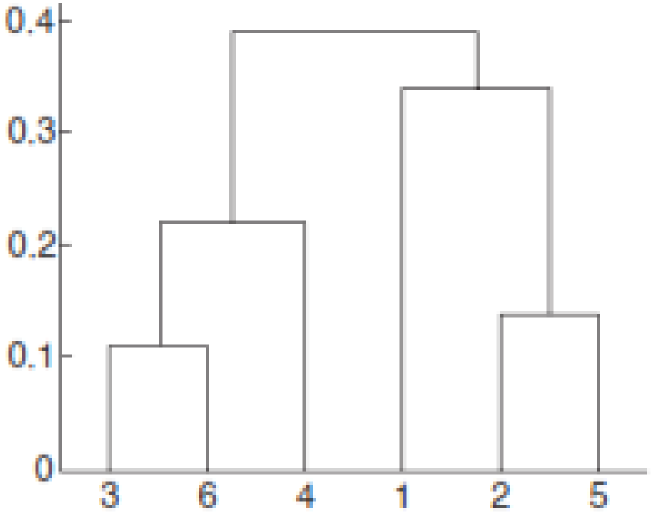
Figure 8.1. Different ways of clustering the same set of points.

- Taken individually, each collection of clusters in Figures 8.1 (b–d) is a partitional clustering.
- If we allow clusters to be nested, then one interpretation of Figure 8.1(a) is that it has two subclusters (Figure 8.1(b)), each of which, in turn, has three subclusters (Figure 8.1(d)).

Some representations of hierarchical clustering



Nested cluster diagram.



Dendrogram.

Hierarchical versus Partitional (cont.)

- A hierarchical clustering can be viewed as a sequence of partitional clusterings.
- A partitional clustering can be obtained by taking any member of that sequence; i.e., by cutting the hierarchical tree at a particular level.

Exclusive vs. Overlapping vs. Fuzzy

- Exclusive: each object is assigned into a single cluster.
- Overlapping or non-exclusive: an object can simultaneously belong to more than one group (class).
    - A non-exclusive clustering is also often used when, for example, an object is "between" two or more clusters and could reasonably be assigned to any of these clusters.
- In a fuzzy clustering, every object belongs to every cluster with a membership weight that is between 0 (absolutely doesn't belong) and 1 (absolutely belongs) → clusters are treated as fuzzy sets.

Exclusive vs. Overlapping vs. Fuzzy (cont.)

- In fuzzy clustering, we often impose the additional constraint that the sum of the weights for each object must equal 1.
- Similarly, probabilistic clustering techniques compute the probability with which each point belongs to each cluster, and these probabilities must also sum to 1.
- A fuzzy or probabilistic clustering does not address true multiclass situations, such as non-exclusive clustering.
- In practice, a fuzzy or probabilistic clustering is often converted to an exclusive clustering by assigning each object to the cluster in which its membership weight or probability is highest

Complete vs. Partial

- A complete clustering assigns every object to a cluster.
  - For example, an application that uses clustering to organize documents for browsing needs to guarantee that all documents can be browsed
- A partial clustering does not assign every object to a cluster.
  - Some objects in a data set may not belong to well-defined groups, even some objects may represent noise or outliers.
  - For example, some newspaper stories may share a common theme, such as global warming, while other stories are more generic or one-of-a-kind.

K-Means Clustering

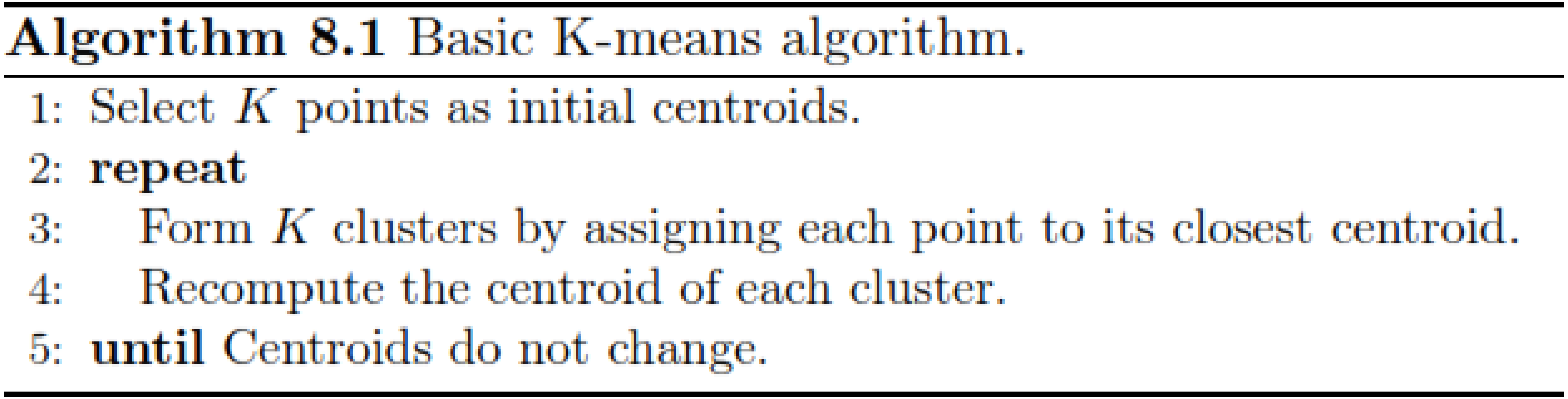
Prototype-based clustering

- K-means adalah prototype based clustering dan merupakan one-level partitioning dari objek yang ada pada data
- **K-means**
  - Mendefinisikan prototype dalam bentuk centroid(mean of a group of points)
  - A centroid almost never corresponds to an actual data point
- **K-medoid**
  - Prototype dalam bentuk medoid (the most representative point for a group of points)
  - A medoid, by its definition, must be an actual data point

Konsep K-Means Clustering

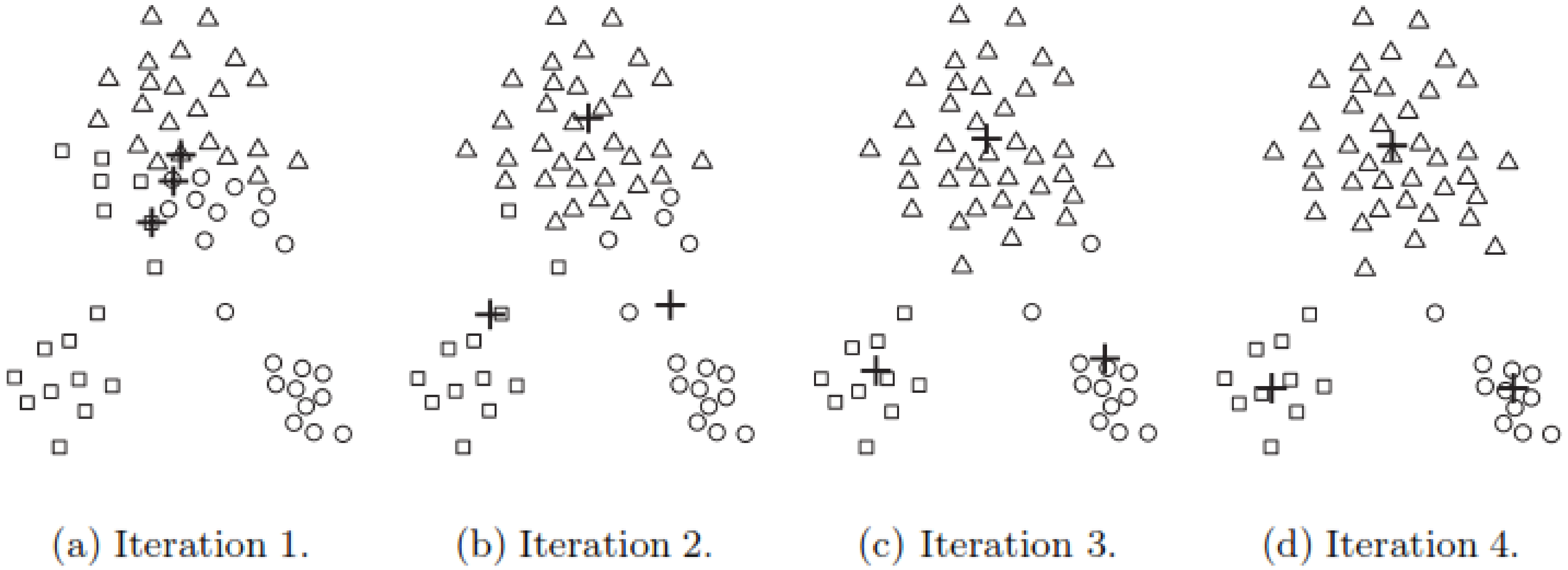
- Mengelompokkan n objek ke dalam K cluster berdasarkan nilai atribut dari objek tersebut.
- K adalah jumlah cluster, berupa bilangan integer positif.
- Merupakan jenis hard clustering → 1 objek hanya dapat menjadi anggota dari 1 cluster secara eksklusif.

The Basic K-means Algorithm



Algorithm 8.1 Basic K-means algorithm.

```
Select $K$ points as initial centroids.
repeat
    Form $K$ clusters by assigning each point to its closest centroid.
    Recompute the centroid of each cluster.
until Centroids do not change.
```



(a) Iteration 1.

(b) Iteration 2.

(c) Iteration 3.

(d) Iteration 4.

Figure 8.3. Using the K-means algorithm to find three clusters in sample data.

- Centroid dilambangkan dengan “+”
- All points belonging to the same cluster have the same marker shape.

Some Distance Measures

Distance/ Dissimilarity Measures

- Interval-scaled variables
  - Euclidean distance
  - Manhattan distance
  - Minkowski distance
- Categorical variables
- Ordinal variables

Interval-scaled variables

- Continous measurements of a roughly linear scale.
- Contoh: tinggi badan, berat badan, suhu ruangan, koordinat lintang dan bujur.
- Beberapa ukuran distance yang dapat digunakan:
  - Euclidean Distance
  - Minkowski Distance
  - Manhattan Distance

Mathematic requirement for distance measures

- $d(\boldsymbol{x}, \boldsymbol{y}) \geq 0$ Jarak tidak boleh negatif.
- $d(\boldsymbol{x}, \boldsymbol{x}) = 0$ Jarak objek terhadap dirinya sendiri adalah 0.
- $d(\boldsymbol{x}, \boldsymbol{y}) = d(\boldsymbol{y}, \boldsymbol{x})$ Jarak merupakan fungsi simetris.
- $d(\boldsymbol{x}, \boldsymbol{y}) \leq d(\boldsymbol{x}, \boldsymbol{z}) + d(\boldsymbol{z}, \boldsymbol{y})$ jarak objek x terhadap objek y secara langsung dalam sebuah ruang tidak melebihi jarak jika dilewatkan ke objek lain (*triangular inequality*).

Euclidean Distance

- Ukuran jarak yang paling sering digunakan untuk data numerik.
- Jarak Euclidean antara 2 titik atau tuple, $\boldsymbol{x} = (x_1, x_2, \dots, x_d)$ dan $\boldsymbol{y} = (y_1, y_2, \dots, y_n)$ pada ruang dimensi d adalah

$$d_{euc}(\boldsymbol{x}, \boldsymbol{y}) = \sqrt{\sum_{i=1}^{d} (x_i - y_i)^2}$$

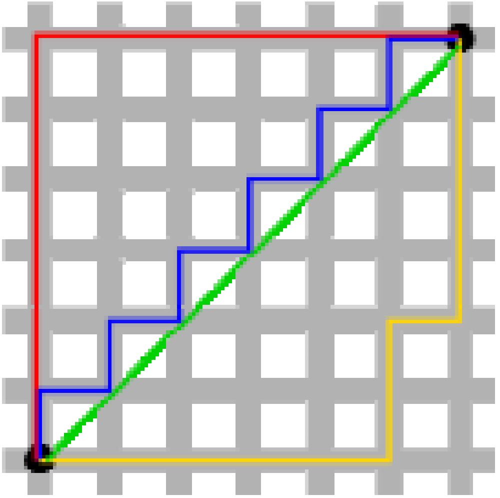
Manhattan (City Block Distance)

- Jumlah jarak dari semua atribut.
- Jarak Manhattan antara 2 titik atau tuple, $\boldsymbol{x} = (x_1, x_2, \dots, x_d)$ dan $\boldsymbol{y} = (y_1, y_2, \dots, y_n)$ pada ruang dimensi d adalah

$$d_{man}(\boldsymbol{x}, \boldsymbol{y}) = \sum_{i=1}^{d} |x_i - y_i|$$

Examples:

- Tentukan jarak titik A(3,10) dan titik B(6,8) menggunakan:
    - Jarak Euclidean
    - Jarak Manhattan



Manhattan versus Euclidean distance:

The red, blue, and yellow lines represent Manhattan distance. They all are of same length as **12**.

The green line represent Euclidian distance of length **$6 \times \sqrt{2} \approx 8.48$.**

Minkowski Distance

- Generalisasi dari jarak Euclidean dan Manhattan.
- Jarak Manhattan antara 2 titik atau tuple, $\boldsymbol{x} = (x_1, x_2, \dots, x_d)$ dan $\boldsymbol{y} = (y_1, y_2, \dots, y_n)$ pada ruang dimensi d adalah

$$d_{min}(\boldsymbol{x}, \boldsymbol{y}) = \left(\sum_{i=1}^{d} |x_i - y_i|^p \right)^{\frac{1}{p}} \qquad , p \geq 1$$

- Jika $p = 1$ → Manhattan distance
- Jika $p = 2$ → Euclidean distance

Normalisasi Data

- Data di tiap atribut umumnya dinormalisasi terlebih dahulu.
  - untuk menghindari *overweighing* atribut dengan range data besar (misalnya, income) terhadap atribut dengan range data kecil (misalnya, IPK).
- Beberapa cara untuk normalisasi antara lain:
  - Min-max normalization
  - Z-score normalization

Min-Max Normalization

Data dinormalisasi ke range $[x_{min}, x_{max}]$ dengan cara:

$$x' = \ x_{min} + \ [(x_{max} - x_{min}) \times \frac{x - x_{\mathrm{min}\_data})}{(x_{\mathrm{max}\_data} - x_{\mathrm{min}\_data})}]$$

Dimana:

x adalah nilai data semula

x' adalah data x yang telah dinormalisasi

x_{min} dan x_{max} adalah nilai minimum dan maksimum *range* data normalisasi

x_{min_data} dan x_{max_data} adalah nilai minimum dan maksimum pada data *input*

Z-Score Normalization

Data dinormalisasi berdasarkan nilai mean dan standar deviasi dari atribut dengan cara

$$v' = \frac{v - \bar{A}}{\sigma_A}$$

dimana:

v adalah nilai atribut A

$\bar{A}$ adalah nilai rata-rata atribut A

σ_A adalah standar deviasi dari atribut A

Categorical Variables

- Contoh categorical variables → warna : merah, kuning, hijau.
- Jarak Manhattan antara 2 objek $\boldsymbol{x}$ dan $\boldsymbol{y}$ yang dideskripsikan dengan categorical variables adalah:

$$d(\boldsymbol{x}, \boldsymbol{y}) = \frac{p - m}{p}$$

p adalah jumlah variabel

m adalah jumlah variabel yang bernilai sama pada objek $\boldsymbol{x}$ dan $\boldsymbol{y}$

Object identifier	Color (categorical)	Shape (categorical)	Material (categorical)
A	Yellow	Triangle	Wood
B	Red	Rectangle	Paper
C	Green	Triangle	Wood
D	Yellow	Triangle	Wood

$$d(B,A) = \frac{3-0}{3} = 1$$

$$d(C,A) = \frac{3-2}{3} = 0{,}333$$

Ordinal Variables

- Hampir sama dengan categorical variables, namun nilai-nilainya memiliki urutan yang berarti.
- Contoh: Temperature → high, medium, low.

Ordinal Variables

- Misalkan f adalah sebuah variabel dari himpunan ordinal variables yang mendeskripsikan n objek.
- Jarak atau dissimilarity terkait atribut f adalah sebagai berikut:
  - Nilai atribut f pada objek ke-i adalah x_{if} dan atribut f memiliki M_f urutan state, merepresentasikan ranking $1, \ldots, M_f$.
  - Ubah setiap nilai x_{if} dengan ranking yang bersesuaian, $r_{if} \in \{1, \ldots, M_f\}$.
  - Ubah jarak setiap variabel ke range $[0.0, 1.0]$ dengan cara mengubah ranking r_{if} dari objek ke-i pada variabel f dengan cara

$$z_{if} = \frac{r_{if} - 1}{M_f - 1}$$

  - Dissimilarity antar objek $\boldsymbol{x}$ dan $\boldsymbol{y}$ diukur berdasarkan nilai z_{if}menggunakan interval-scaled variables.

Object identifier	**Atribut 1 (ordinal)**	**Rank** r_{if} $(M_f = 3)$	z_{if}
1	Low	1	(1-1) / (3-1) = 0
2	High	3	(3-1) / (3-1) = 1
3	Medium	2	(2-1) / (3-1) = 0,5
4	Low	1	(1-1) / (3-1) = 0

Back to K-Means Algorithm

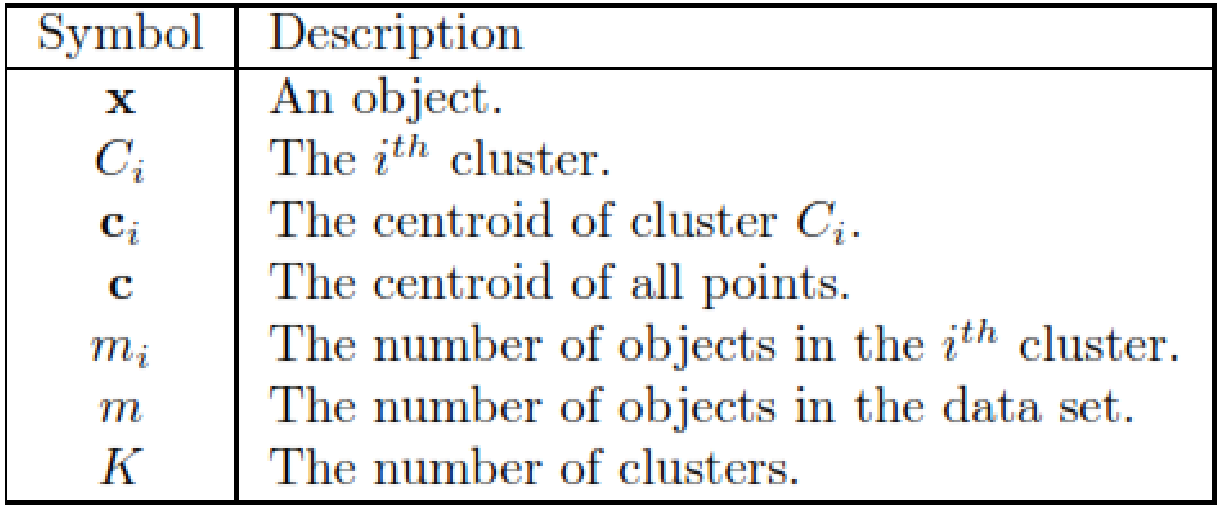
Centroids and Objective Functions

- Updating centroid depends on :
  - The proximity measure for the data
  - The objective function (the goal of the clustering), for example: minimize the **sum of the squared error (SSE)** of each point to its closest centroid.
- In other words, we calculate the error of each data point, i.e., its Euclidean distance to the closest centroid, and then compute the total sum of the squared errors.
- Given two different sets of clusters that are produced by two different runs of K-means, we prefer the one with the smallest SSE → this means that the prototypes (centroids) of this clustering area better representation of the points in their cluster.

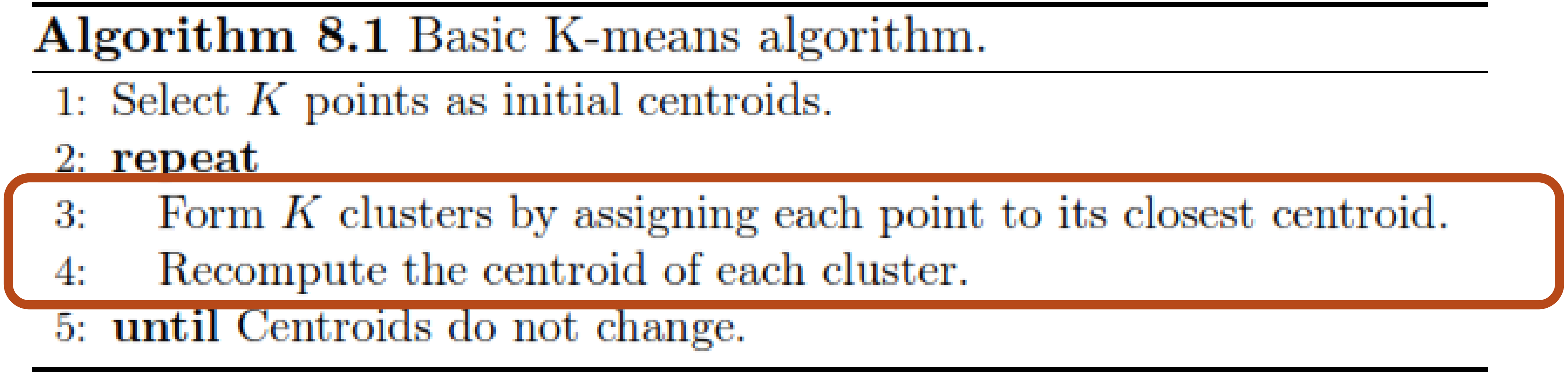
Centroids and Objective Functions (cont.)

$$\text{SSE} = \sum_{i=1}^{K} \sum_{\mathbf{x} \in C_i} dist(\mathbf{c}_i, \mathbf{x})^2$$

$$\mathbf{c}_i = \frac{1}{m_i} \sum_{\mathbf{x} \in C_i} \mathbf{x}$$



Symbol	Description
$\mathbf{x}$	An object.
C_i	The i^{th} cluster.
$\mathbf{c}_i$	The centroid of cluster C_i.
$\mathbf{c}$	The centroid of all points.
m_i	The number of objects in the i^{th} cluster.
m	The number of objects in the data set.
K	The number of clusters.

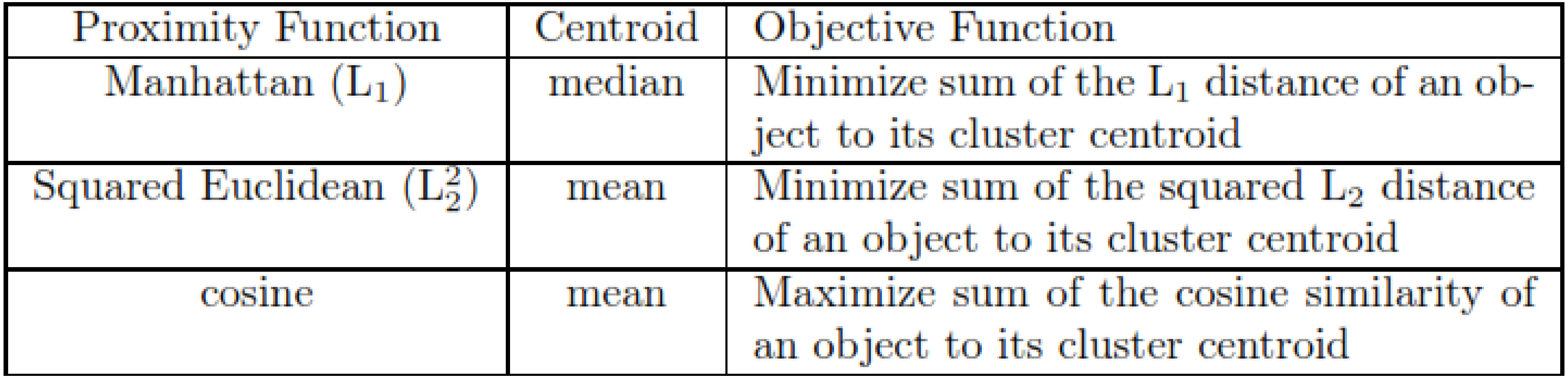


Algorithm 8.1 Basic K-means algorithm.

```
1: Select $K$ points as initial centroids.
2: repeat
3:     Form $K$ clusters by assigning each point to its closest centroid.
4:     Recompute the centroid of each cluster.
5: until Centroids do not change.
```

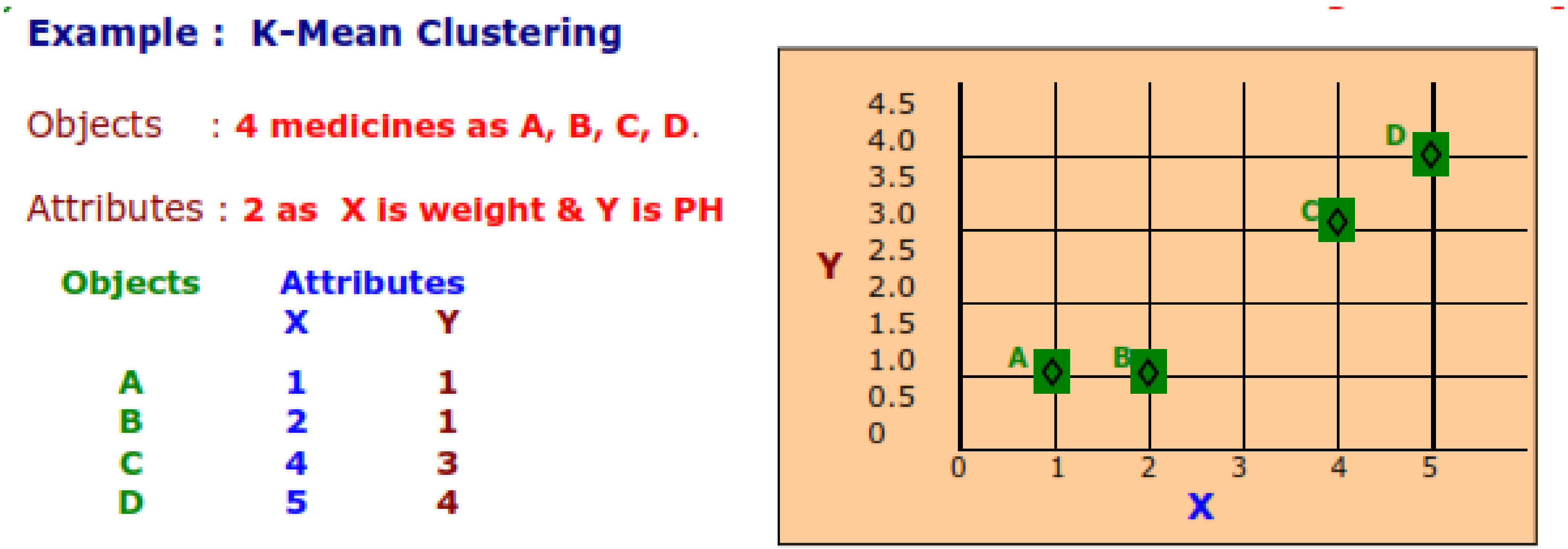
- Steps 3 and 4 of the K-means algorithm directly attempt to minimize objective function (i.e. SSE).
- However, the actions of K-means in Steps 3 and 4 are only guaranteed to find a **local minimum with respect to the SSE** since they are based on optimizing the SSE for specific choices of the centroids and clusters, rather than for all possible choices.

Centroids and Objective Functions (cont.)



Proximity Function	Centroid	Objective Function
Manhattan (L_1)	median	Minimize sum of the L_1 distance of an object to its cluster centroid
Squared Euclidean (L_2^2)	mean	Minimize sum of the squared L_2 distance of an object to its cluster centroid
cosine	mean	Maximize sum of the cosine similarity of an object to its cluster centroid

Example of K-Means Clustering



Example : K-Mean Clustering

Objects : **4 medicines as A, B, C, D.**

Attributes : **2 as X is weight & Y is PH**

Objects	Attributes	
	X	Y
A	1	1
B	2	1
C	4	3
D	5	4

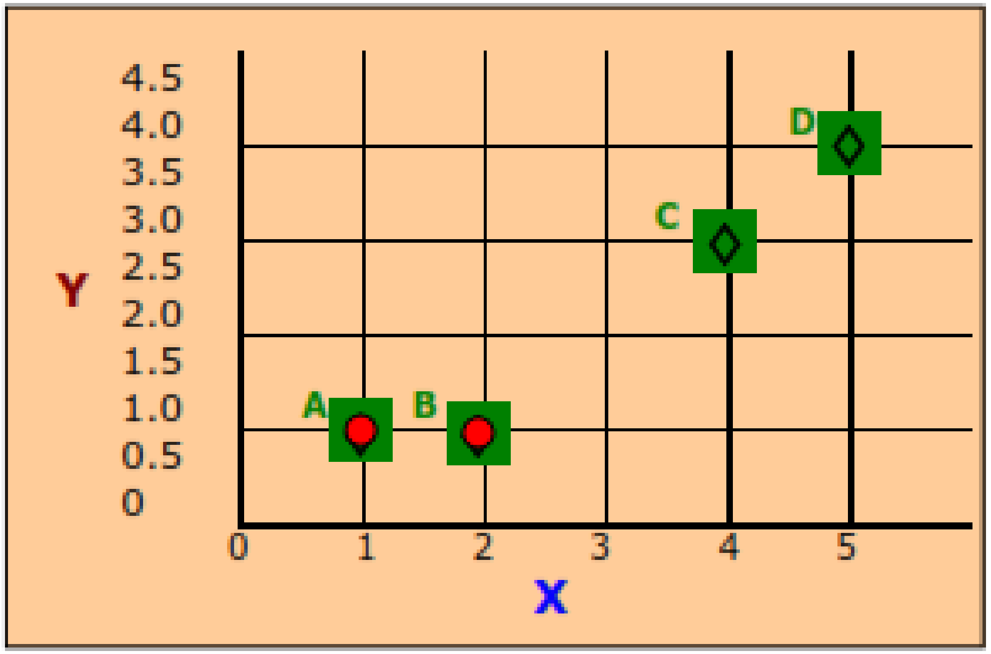
Initial value of centroids:

Suppose medicine **A** and medicine **B** be first centroids. If **P_1** and **P_2** denote coordinates of the centroids,

then **P_1= (1, 1)** and **P_2 = (2, 1)**.

Let centroid **P1** be for cluster group-1

Let centroid **P2** be for cluster group-2.



1. Iteraion 0

(a) Objects Clusters centers stated before :

Group-1 has center **$P_1 = (1, 1)$** ;

Group-2 has center **$P_2 = (2, 1)$** ;

Objects : A , B, C, D

Attributes : X and **Y**

	A	B	C	D
X	1	2	4	5
Y	1	1	3	4

(b) Calculate distances between cluster center to each object

- 1st, calculate the Euclidean distances from cetroid **P_1** to each point **A, B, C, D**. It is the 1st row of the distance matrix.
- 2nd, calculate the Euclidean distances from cetroid **P_2** to each point **A, B, C, D**. It is the 2nd row of the distance matrix.

The ways to calculate just two distance matrix elements **D_{13}** and **D_{23}** are :

$$D_{13} = \sqrt{(C_x - P_{1x})^2 + (C_y - P_{1y})^2} = \sqrt{(4-1)^2 + (3-1)^2} = 3.61$$

$$D_{23} = \sqrt{(C_x - P_{2x})^2 + (C_y - P_{2y})^2} = \sqrt{(4-2)^2 + (3-1)^2} = 2.83$$

Similarly calculate other elements **D_{11} , D_{12}, D_{14}, D_{21} , D_{22}, D_{24}**

(c) Distance matrix becomes

$$D^0 = \left\{ \begin{array}{cccc} 0 & 1 & 3.61 & 5 \\ 1 & 0 & 2.83 & 4,24 \\ A & B & C & D \end{array} \right.$$

1st row indicates **group-1** cluster

2nd row indicates **group-2** cluster

(d) Objects clustering into groups:

Assign group to each object based on the minimum distance. Thus,

medicine **A** is assigned to **group 1**;

medicine **B** is assigned to **group 2**,

medicine **C** is assigned to **group 2** , and

medicine **D** is assigned to **group 2**.

Group Matrix : matrix element is **1** if the object is assigned to that group

$$G^0 = \left\{ \begin{array}{cccc} 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 \\ A & B & C & D \end{array} \right.$$

1st row as **group 1**

2nd row as **group 2**

Objective function (SSE)

$$\begin{aligned} SSE &= D_{11} + D_{22} + D_{23} + D_{24} \\ &= 0 + 0 + 2.83 + 4.24 \\ &= 7.07 \end{aligned}$$

2. Iteration 1 :

The cluster groups have new members. Compute the new centroids of each group. **Repeat the process of iteration indicated below.**

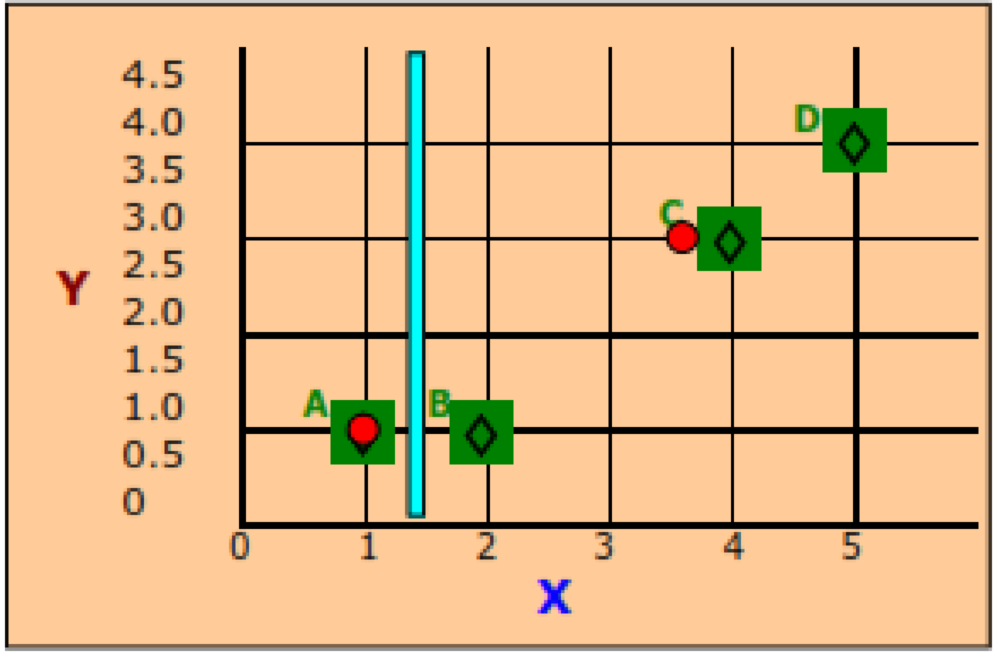
Group-1 has one member **A**, the centroid remains as **P_1 = (1, 1)**.

Group-2 now has 3 members **B, C, D,** so centroid is the average of their coordinates:

$$P_2 = \left(\frac{2+4+5}{3} , \frac{1+3+4}{3} \right)$$

$$= (11/3, 8/3)$$

$$= \mathbf{(3.67 , 2.67)}$$



(a) Objects Clusters centers stated above :

Group-1 has center **P_1 = (1, 1)** ;

Group-2 has center **P_2 = (3.67 , 2.67)** ;

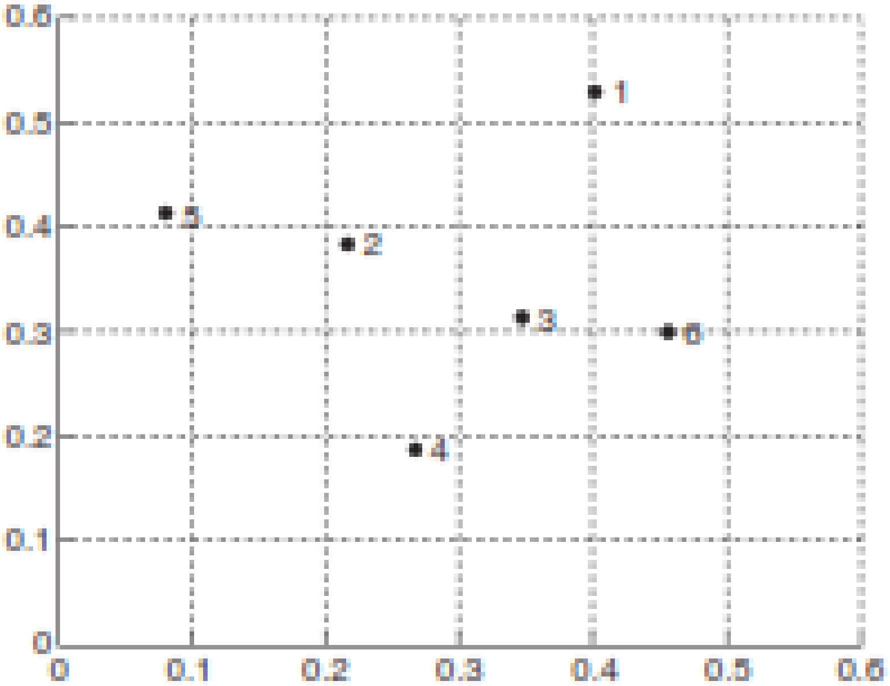
Objects : A , B, C, D

Attributes : X and Y

	A	B	C	D
X	1	2	4	5
Y	1	1	3	4

Tugas

- Lakukanlah clustering dari sample data berikut. Cobalah dengan jumlah cluster K=2 dan K=3.
- Cobalah juga dengan beberapa inisialisasi centroid yang berbeda.
- Analisis hasil clustering berdasarkan hasil percobaan yang telah dilakukan.
- Berikan kesimpulan dari hasil yang diperoleh.



Point	x Coordinate	y Coordinate
p1	0.40	0.53
p2	0.22	0.38
p3	0.35	0.32
p4	0.26	0.19
p5	0.08	0.41
p6	0.45	0.30