Chapter 8

Data mining

Data mining

- 8.1 Introduction
- 8.2 K-means Clustering
- 8.3 KNN Classification

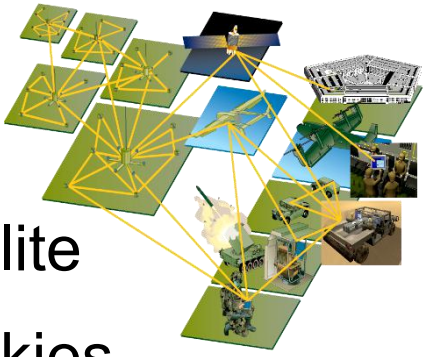
Why Mine Data? Commercial Viewpoint

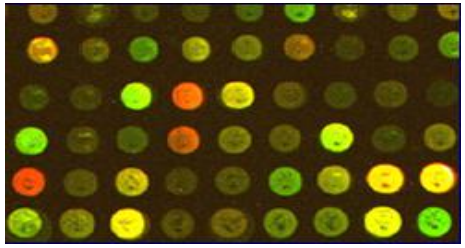
- Lots of data is being collected and warehoused
  - Web data, e-commerce
  - purchases at department/ grocery stores
  - Bank/Credit Card transactions

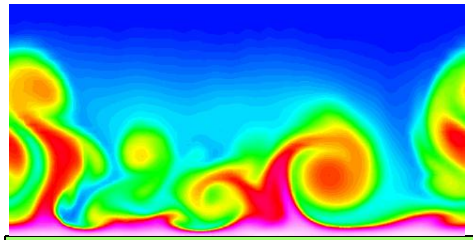
- Computers have become cheaper and more powerful
- Competitive Pressure is Strong
  - Provide better, customized services for an *edge* (e.g. in Customer Relationship Management)

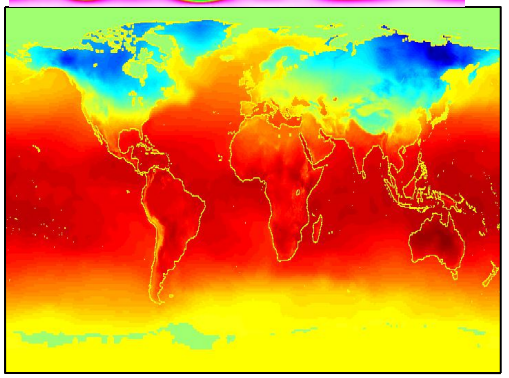
Why Mine Data? Scientific Viewpoint

- Data collected and stored at enormous speeds (GB/hour)
  - remote sensors on a satellite
  - telescopes scanning the skies
  - microarrays generating gene expression data
  - scientific simulations generating terabytes of data
- Traditional techniques infeasible for raw data
- Data mining may help scientists
  - in classifying and segmenting data
  - in Hypothesis Formation



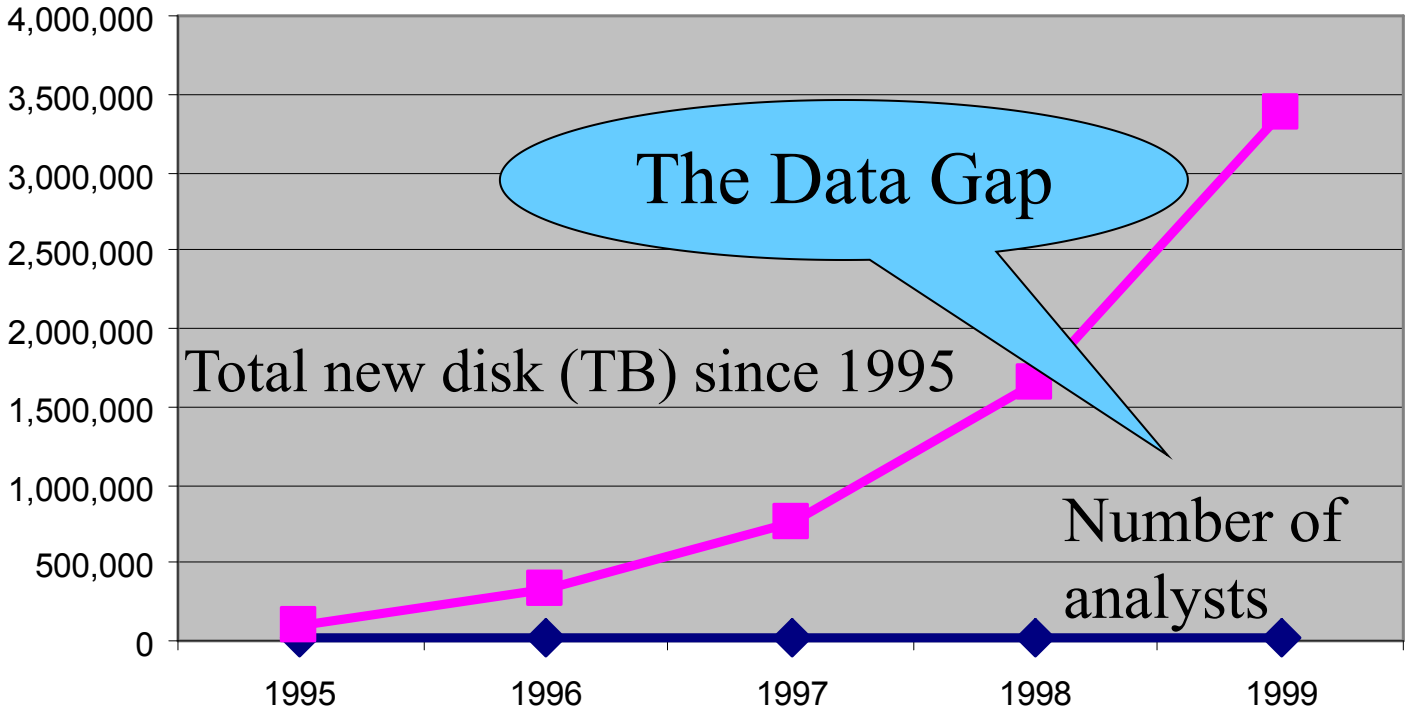






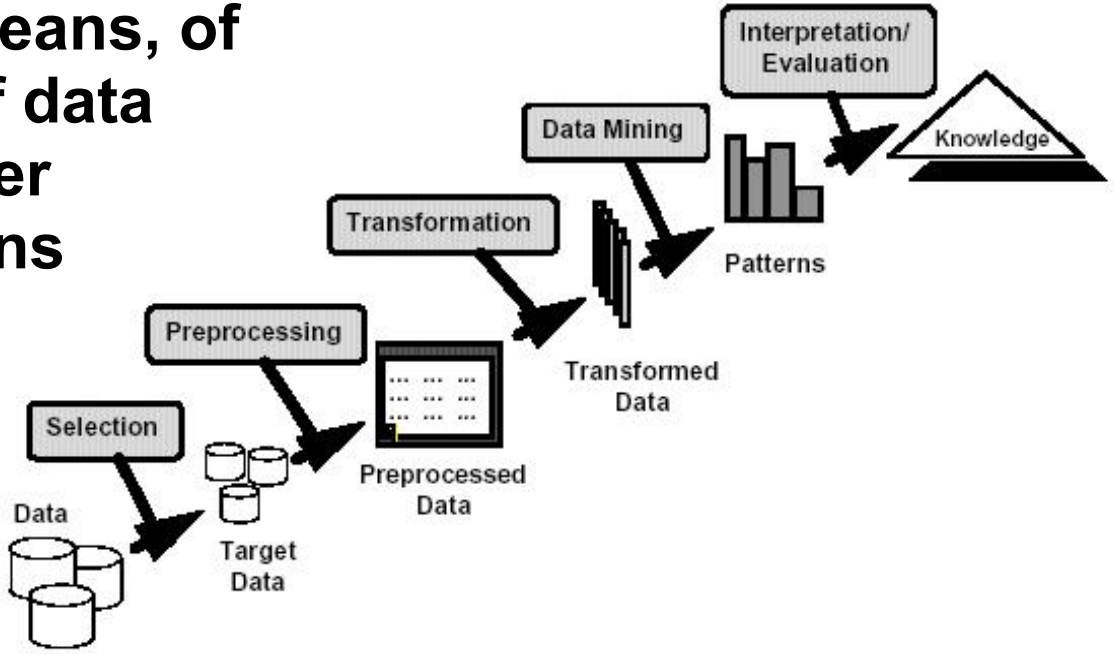
Mining Large Data Sets - Motivation

- **There is often information "hidden" in the data that is not readily evident**
- **Human analysts may take weeks to discover useful information**
- **Much of the data is never analyzed at all**



What is Data Mining?

- **Definitions**
  - **Non-trivial extraction of implicit, previously unknown and potentially useful information from data**
  - **Exploration & analysis, by automatic or semi-automatic means, of large quantities of data in order to discover meaningful patterns**



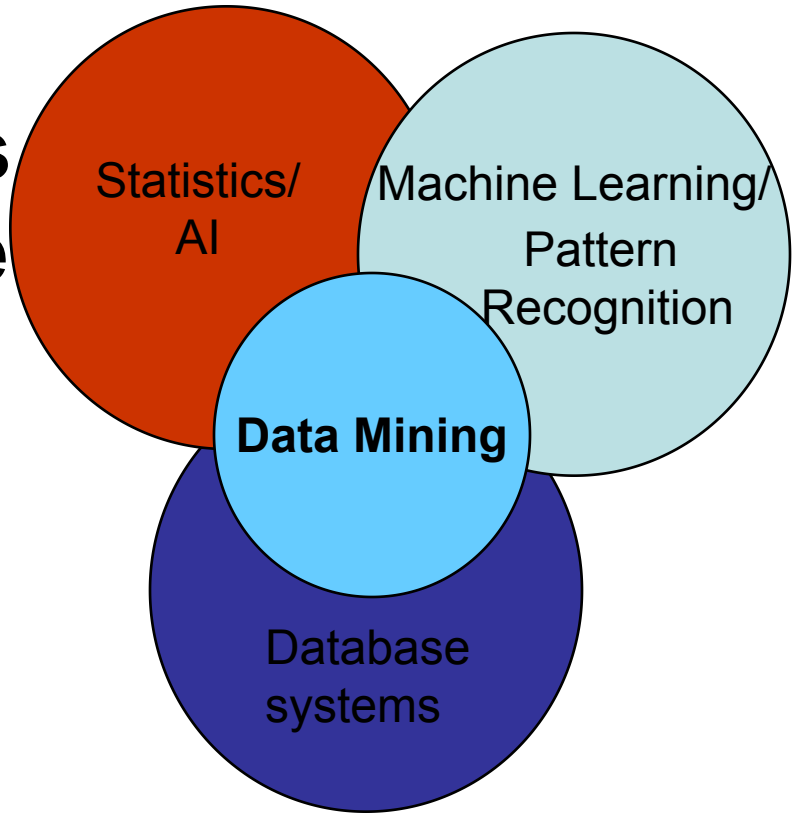
What is (not) Data Mining?

- **What is not Data Mining?**
  - Look up phone number in phone directory
  - Query a Web search engine for information about "Amazon"

- **What is Data Mining?**
  - Certain names are more prevalent in certain US locations (O'Brien, O'Rurke, O'Reilly… in Boston area)
  - Group together similar documents returned by search engine according to their context (e.g. Amazon rainforest, Amazon.com,)

Origins of Data Mining

- **Draws ideas from machine learning/AI, pattern recognition, statistics, and database systems**
- **Traditional Techniques may be unsuitable due**
  - **Enormity of data**
  - **High dimensionality of data**
  - **Heterogeneous, distributed nature of data**



Data Mining Tasks

- Prediction Methods
  - Use some variables to predict unknown or future values of other variables.

- Description Methods
  - Find human-interpretable patterns that describe the data.

Data Mining Tasks...

- Classification [Predictive]
- Clustering [Descriptive]
- Association Rule Discovery [Descriptive]
- Sequential Pattern Discovery [Descriptive]
- Regression [Predictive]
- Deviation Detection [Predictive]

Classification:

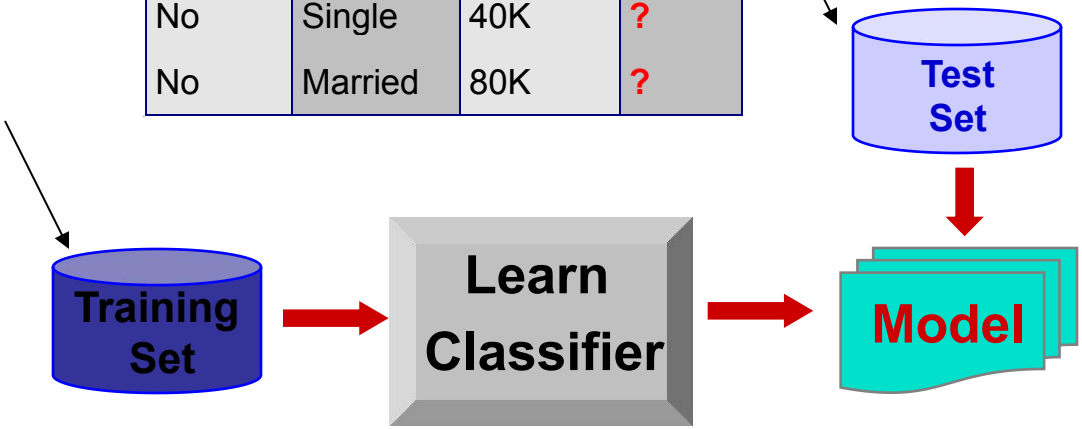
- Given a collection of records (*training set*)
  - Each record contains a set of *attributes*, one of the attributes is the *class*.
- Find a *model* for class attribute as a function of the values of other attributes.
- Goal: previously unseen records should be assigned a class as accurately as possible.
  - A *test set* is used to determine the accuracy of the model. Usually, the given data set is divided into training and test sets, with training set used to build the model and test set used to validate it.

Classification Example

categorical categorical continuous class

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Refund	Marital Status	Taxable Income	Cheat
No	Single	75K	?
Yes	Married	50K	?
No	Married	150K	?
Yes	Divorced	90K	?
No	Single	40K	?
No	Married	80K	?



Clustering:

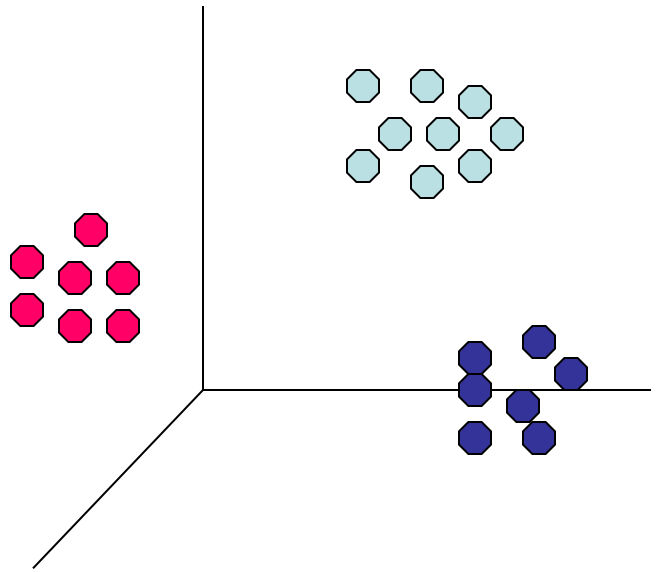
- Given a set of data points, each having a set of attributes, and a similarity measure among them, find clusters such that
  - Data points in one cluster are more similar to one another.
  - Data points in separate clusters are less similar to one another.
- Similarity Measures:
  - Euclidean Distance if attributes are continuous.
  - Other Problem-specific Measures.

Illustrating Clustering

- Euclidean Distance Based Clustering in 3-D space.

Intracluster distances are minimized

Intercluster distances are maximized



Association Rule Discovery: Definition

- Given a set of records each of which contain some number of items from a given collection;
  - Produce dependency rules which will predict occurrence of an item based on occurrences of other it

TID	*Items*
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

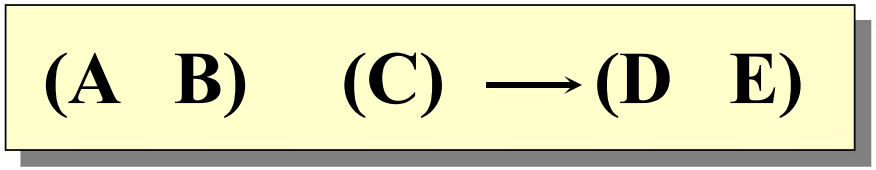
Rules Discovered:

{Milk} --> {Coke}

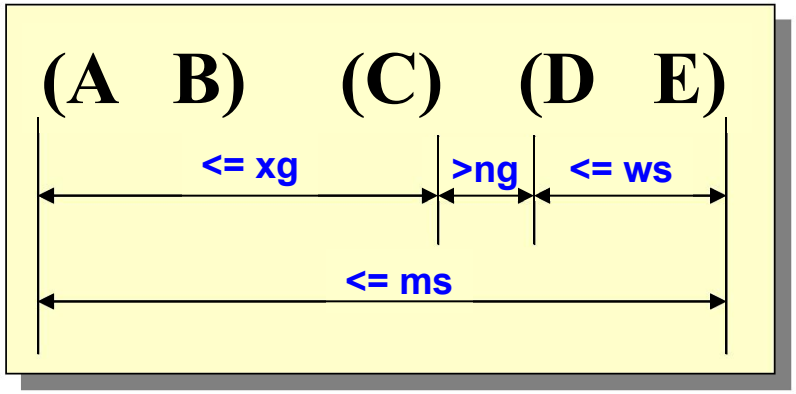
{Diaper, Milk} --> {Beer}

Sequential Pattern Discovery: Definition

- Given is a set of *objects*, with each object associated with its own *timeline of events*, find rules that predict strong sequential dependencies among different events.



- Rules are formed by first discovering patterns. Event occurrences in the patterns are governed by timing constraints.



Regression

- Predict a value of a given continuous valued variable based on the values of other variables, assuming a linear or nonlinear model of dependency.
- Greatly studied in statistics, neural network fields.
- Examples:
  - Predicting sales amounts of new product based on advetising expenditure.
  - Predicting wind velocities as a function of temperature, humidity, air pressure, etc.
  - Time series prediction of stock market indices.

Deviation/Anomaly Detection

- Detect significant deviations from normal behavior
- Applications:
  - Credit Card Fraud Detection

  - Network Intrusion Detection

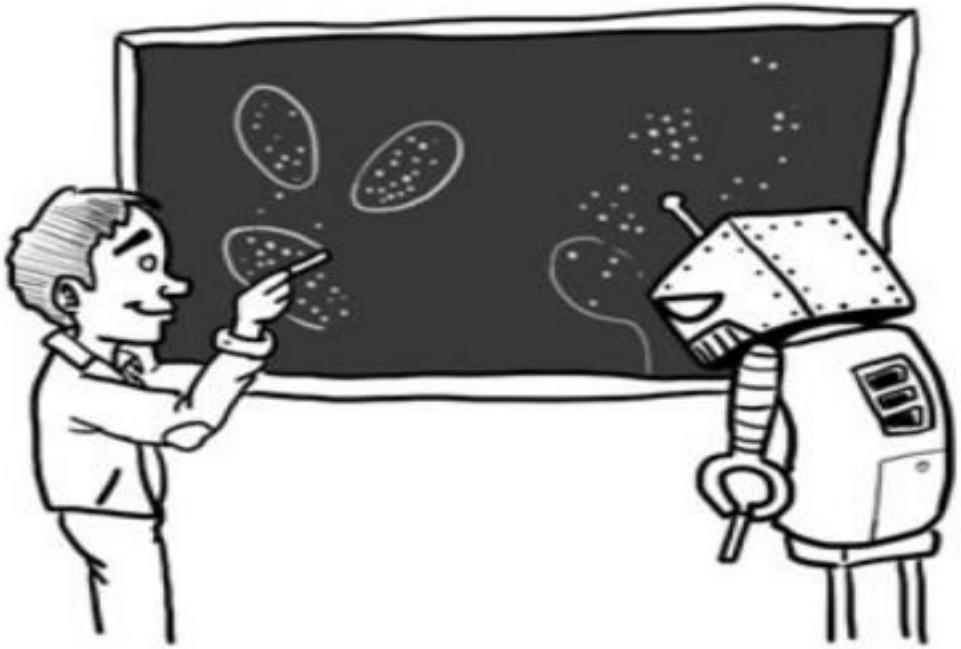
Typical network traffic at University level may reach over 100 million connections per day

Challenges of Data Mining

- Scalability
- Dimensionality
- Complex and Heterogeneous Data
- Data Quality
- Data Ownership and Distribution
- Privacy Preservation
- Streaming Data

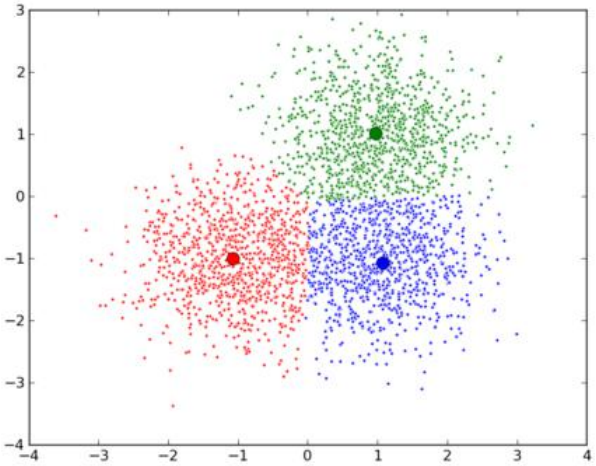
8-2 K-means Clustering

algorithm



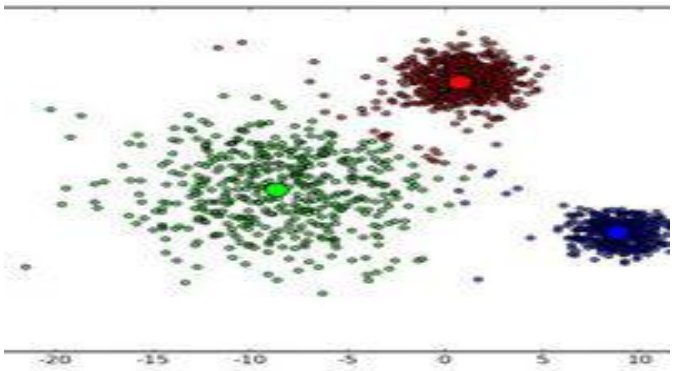
K-means Clustering

K-means clustering is a sort of clustering algorithm and it is a method of vector quantization, originally from signal processing, that is popular for cluster analysis in data mining. K-means clustering aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster. --From Wikipedia

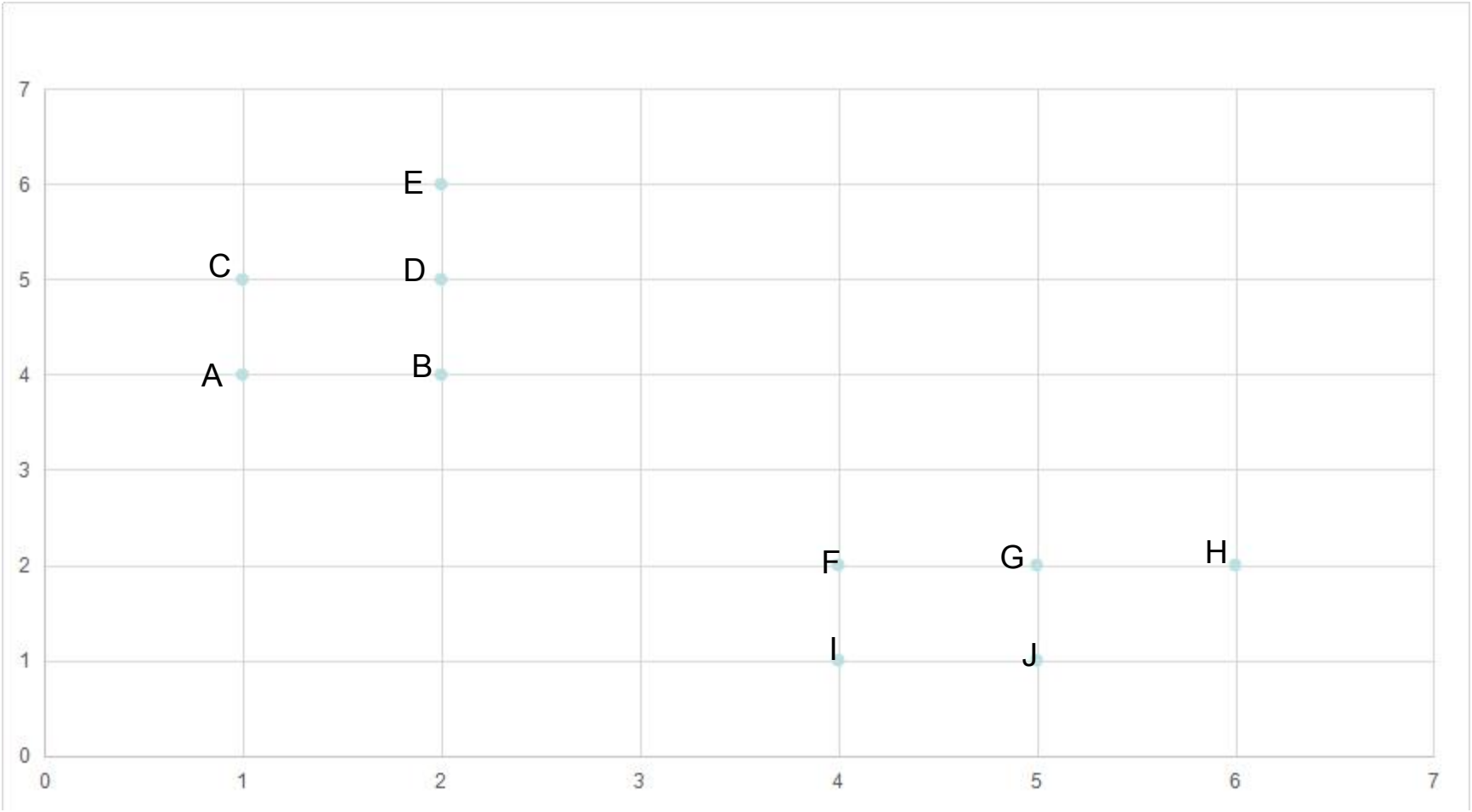


Algorithm Procedure

1. Randomly select K points from complete samples as the initial center.(That's what k means in K-means)
2. Each point in the dataset is assigned to the closed cluster,based upon the Euclidean distance between each point and each cluster center. $S=\sqrt{(X_1-X_2)^2+(Y_1-Y_2)^2}$
3. Each cluster's center is recomputed as the average of the points in that cluster.
4. Iterate step 2 or more until the new center of cluster equals to the original center of cluster or less than a specified threshold,then clustering finished.

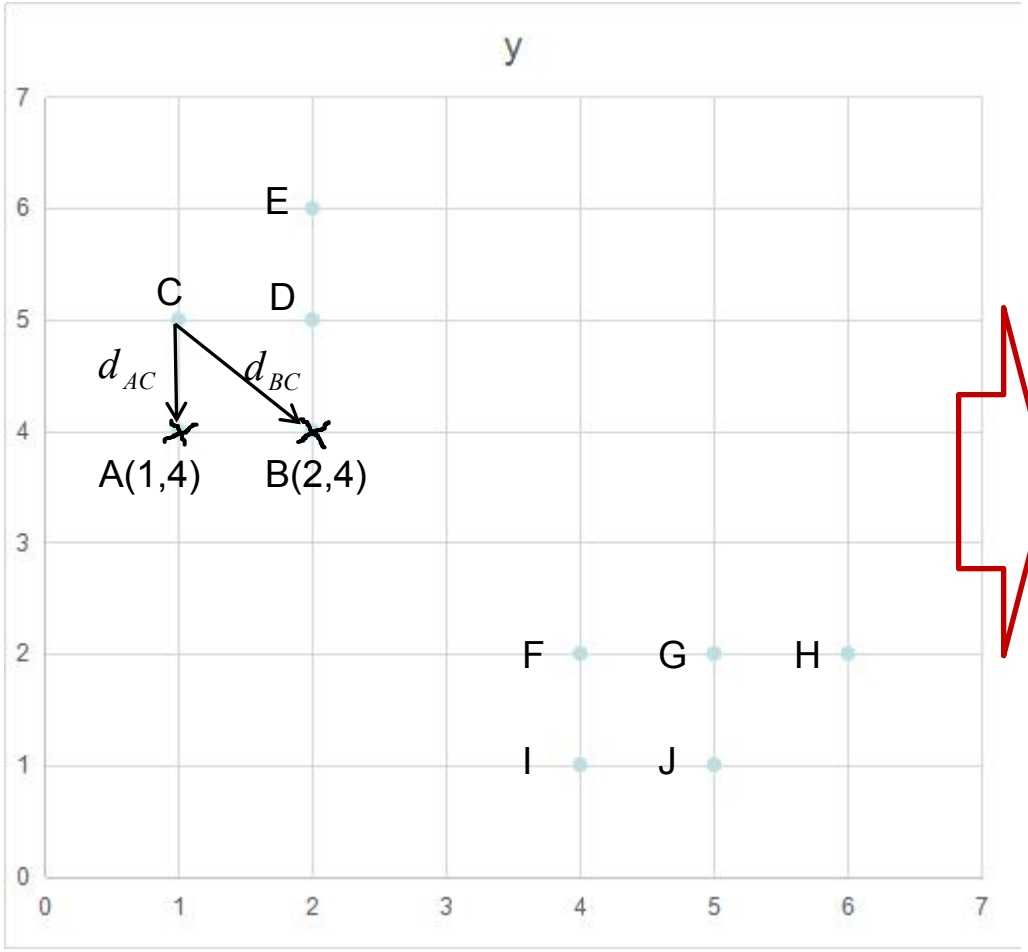


Example



How to cluster A,B...H,J into two clusters?

Example



Randomly choose A,B as the centre and K=2.

$$S=\sqrt{(X_1-X_2)^2+(Y_1-Y_2)^2}$$

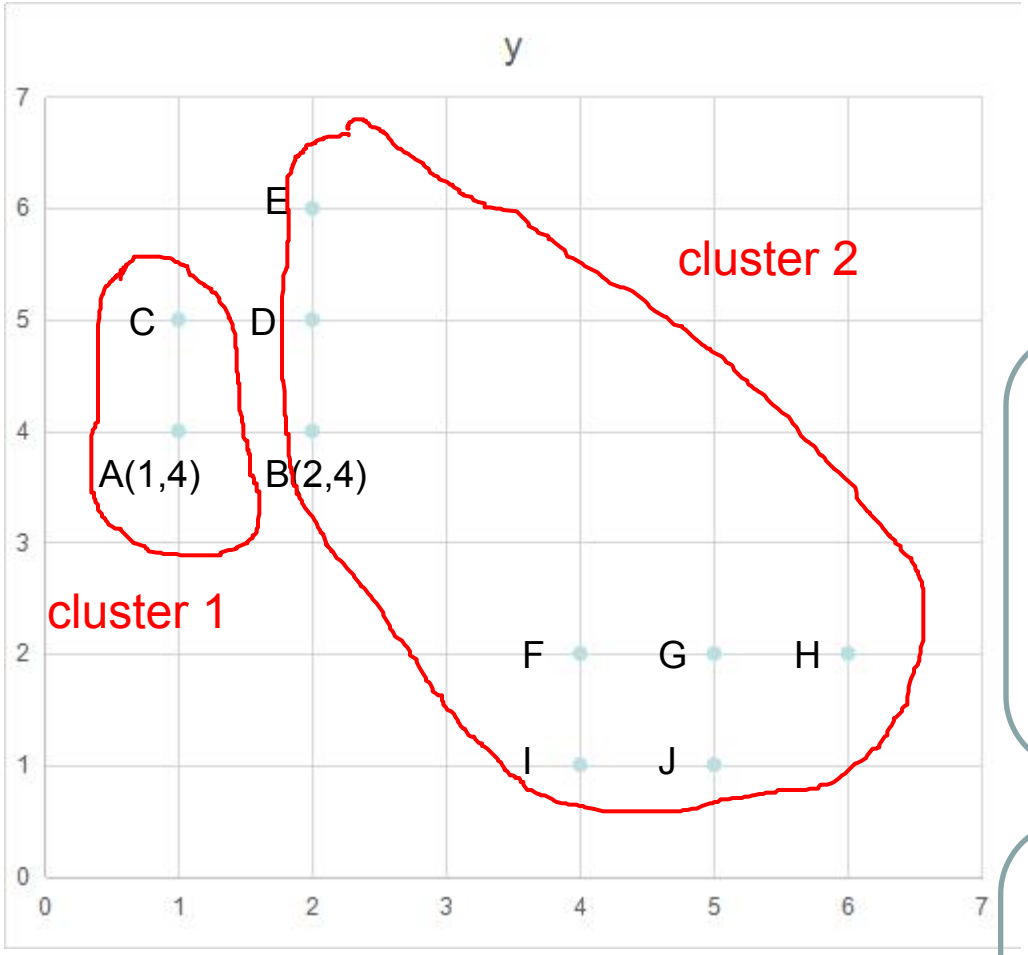
A	$d_{AA}=0$	<	$d_{BA}=1$
B	$d_{AB}=1$	>	$d_{BB}=0$
C	$d_{AC}=1$	<	$d_{BC}=1.41$
D	$d_{AD}=1.41$	>	$d_{BD}=1$
E	$d_{AE}=2.24$	>	$d_{BE}=2$
F	$d_{AF}=3.61$	>	$d_{BF}=2.83$
G	$d_{AG}=4.47$	>	$d_{BG}=3.61$
H	$d_{AH}=5.39$	>	$d_{BH}=4.47$
I	$d_{AI}=4.24$	>	$d_{BI}=3.61$
J	$d_{AJ}=5$	>	$d_{BJ}=4.24$

d_{AB} means distance A→B

Step 1 and 2.

So,we classify A,C as a cluster and B,E,D,F,G,H,I and J as another cluster.

Example



Randomly choose A,B as the centre and K=2.

$$center = (\frac{\sum x_i}{i}, \frac{\sum y_j}{j})$$

$$\alpha_{A,C} = (\frac{1+1}{2}, \frac{4+5}{2}) = (1,4.5)$$

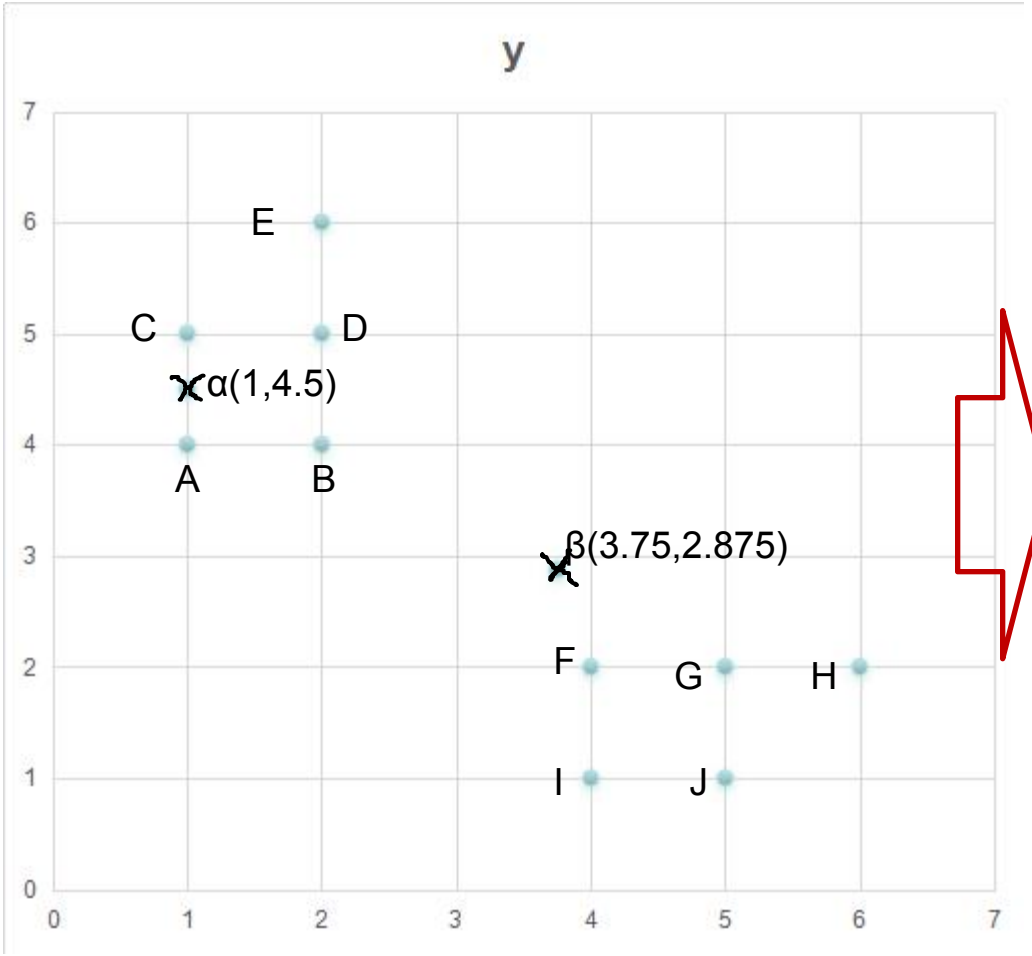
new center

$$\beta_{B,D,E,F,G,H,I,J} = (3.75,2.875)$$

The new centers of the two clusters are (1,4.5) and (3.75,2.875)

Step 3.

Example



$$S=\sqrt{(X_1-X_2)^2+(Y_1-Y_2)^2}$$

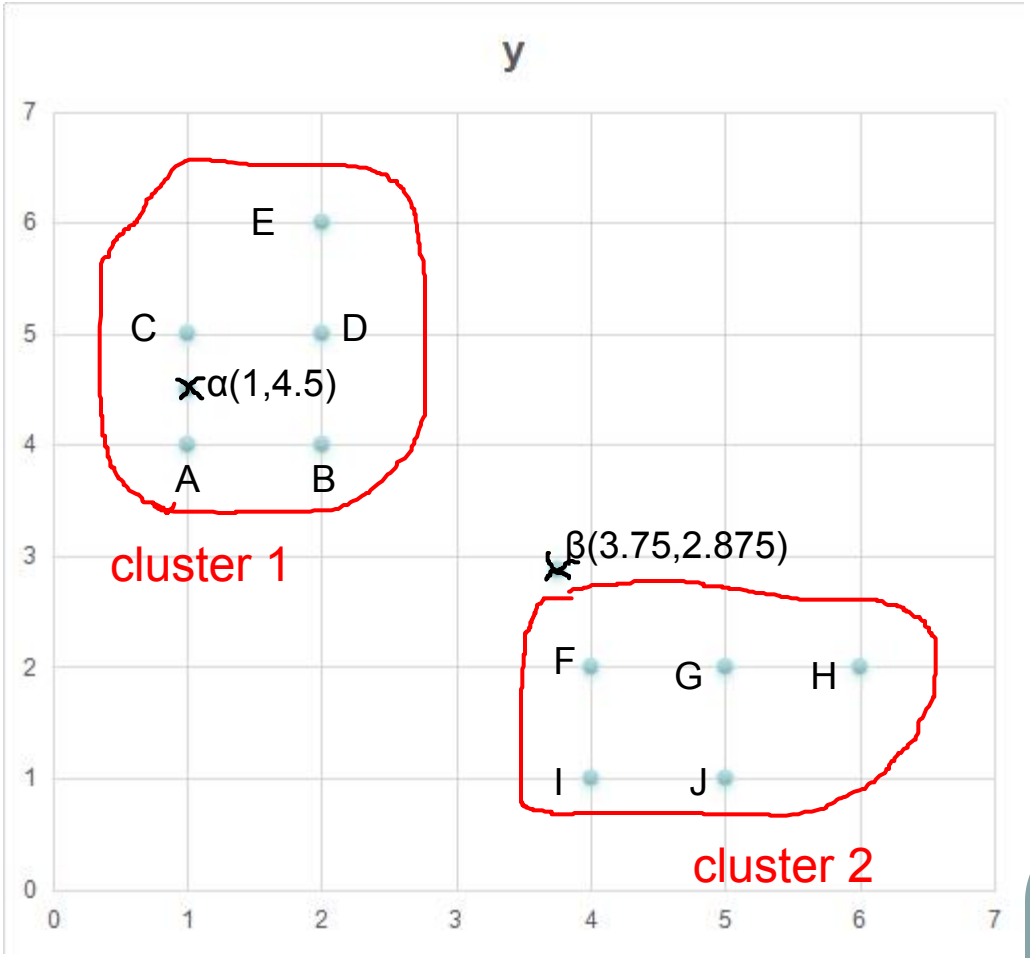
A	$d_{\alpha A}=0.5$	<	$d_{\beta A}=2.97$
B	$d_{\alpha B}=1.12$	<	$d_{\beta B}=2.08$
C	$d_{\alpha C}=0.5$	<	$d_{\alpha C}=3.48$
D	$d_{\alpha D}=1.12$	<	$d_{\beta D}=2.75$
E	$d_{\alpha E}=1.8$	<	$d_{\beta E}=3.58$
F	$d_{\alpha F}=3.91$	>	$d_{\beta F}=0.91$
G	$d_{\alpha G}=4.72$	>	$d_{\beta G}=1.53$
H	$d_{\alpha H}=5.59$	>	$d_{\beta H}=2.41$
I	$d_{\alpha I}=4.61$	>	$d_{\beta I}=1.89$
J	$d_{\alpha J}=5.32$	>	$d_{\beta J}=2.25$

α , β as the centre and K=2.

Step 2 again.

So,we classify A,B,C,D,E as a cluster and F,G,H,I,J as another cluster.

Example



α , β as the centre and K=2.

Step 3 again.

$$center = (\frac{\sum x_i}{i}, \frac{\sum y_j}{j})$$

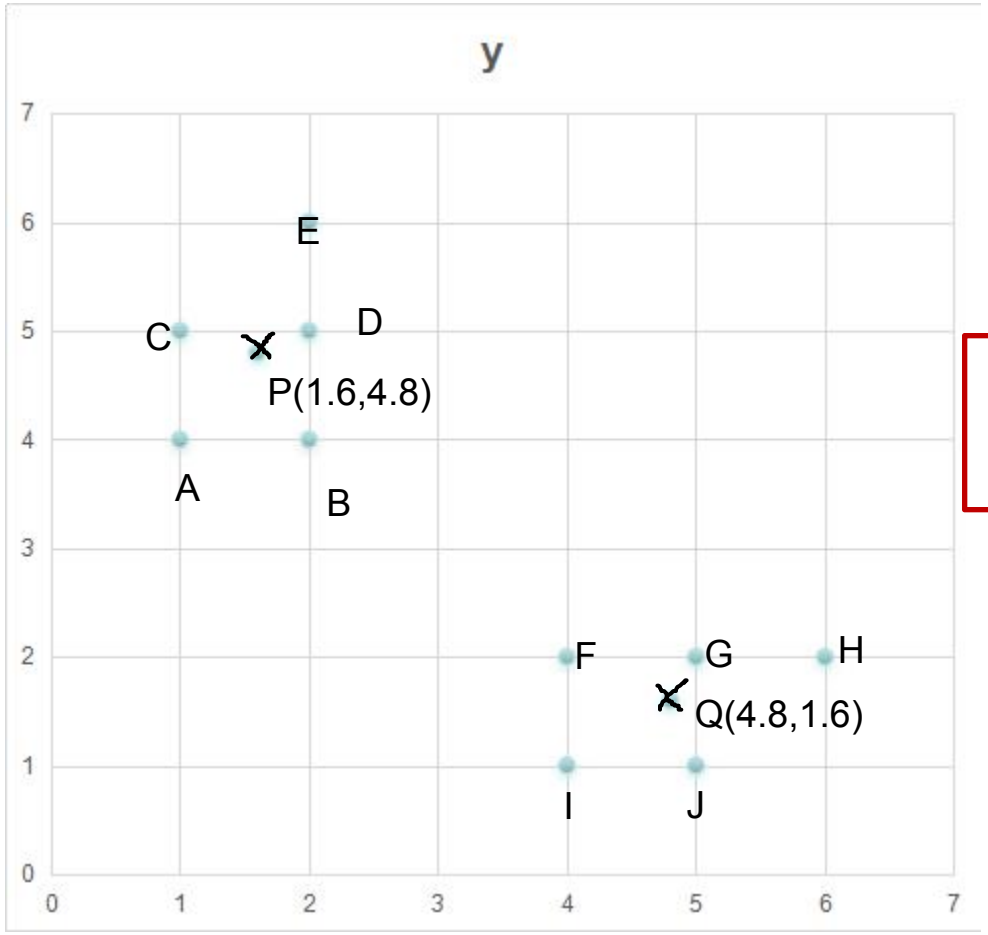
$$P_{A,B,C,D,E} = (1.6,4.8)$$

new center

$$Q_{F,G,H,I,J} = (4.8,1.6)$$

The new centers of the two clusters are P(1.6,4.8) and Q(4.8,1.6)

Example



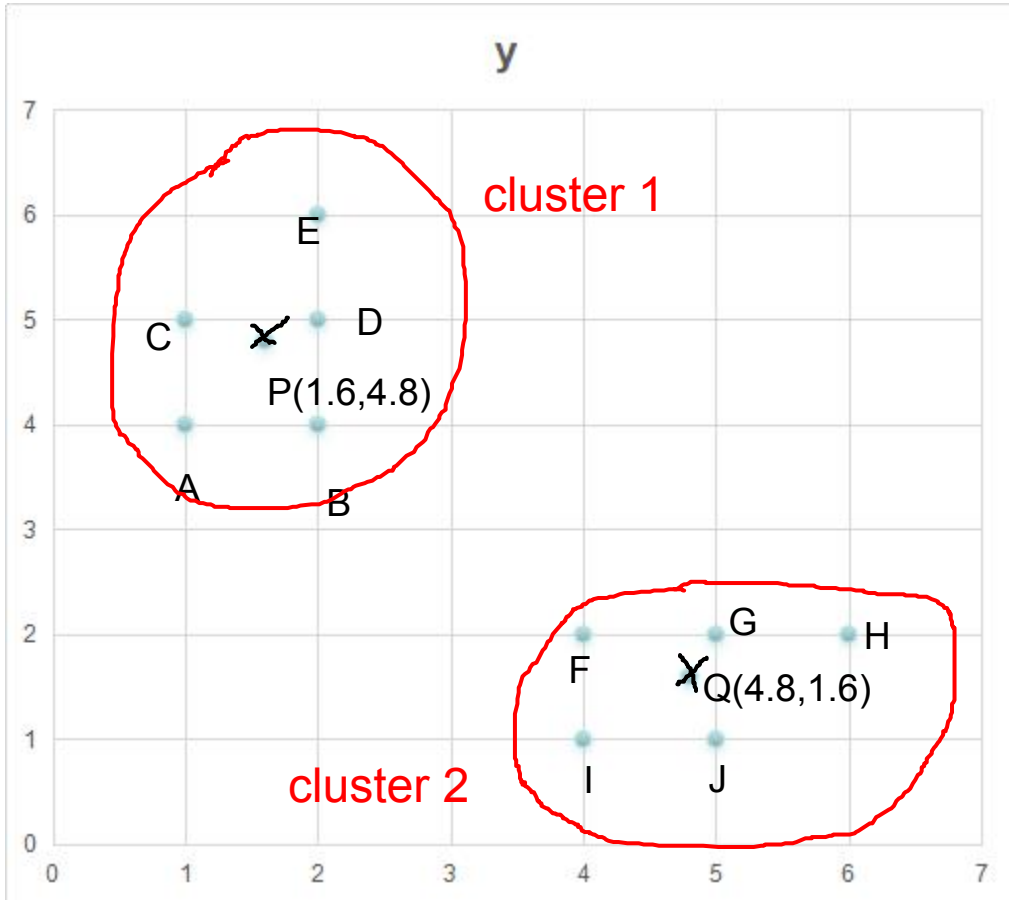
$$S = \sqrt{(X_1 - X_2)^2 + (Y_1 - Y_2)^2}$$

A	$d_{PA} = 1$	<	$d_{QA} = 4.49$
B	$d_{PB} = 0.89$	<	$d_{QB} = 3.69$
C	$d_{PC} = 0.63$	<	$d_{QC} = 5.10$
D	$d_{PD} = 0.45$	<	$d_{QD} = 4.4$
E	$d_{PE} = 1.26$	<	$d_{QE} = 5.22$
F	$d_{PF} = 3.69$	>	$d_{QF} = 0.89$
G	$d_{PG} = 4.40$	>	$d_{QG} = 0.45$
H	$d_{PH} = 5.22$	>	$d_{QH} = 1.26$
I	$d_{PI} = 4.49$	>	$d_{QI} = 1$
J	$d_{PJ} = 5.10$	>	$d_{QJ} = 0.63$

Step 2 again.

So,we classify A,B,C,D,E as a cluster and F,G,H,I,J as another cluster.

Example



$$center = (\frac{\sum x_i}{i}, \frac{\sum y_j}{j})$$

$$M_{A,B,C,D,E} = (1.6,4.8)$$

new center

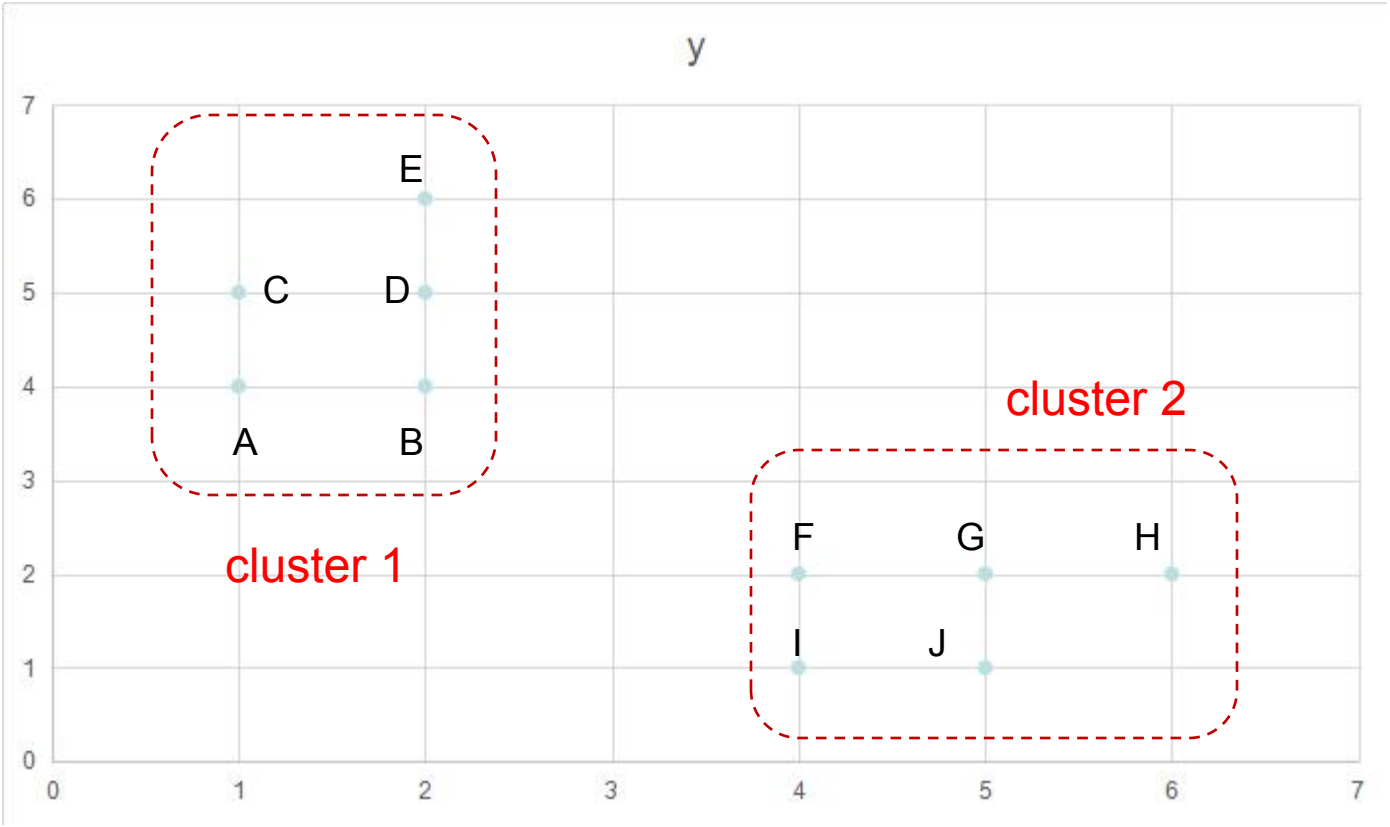
$$N_{F,G,H,I,J} = (4.8,1.6)$$

The new centers of the two clusters are equal to the original P(1.6,4.8) and Q(4.8,1.6)

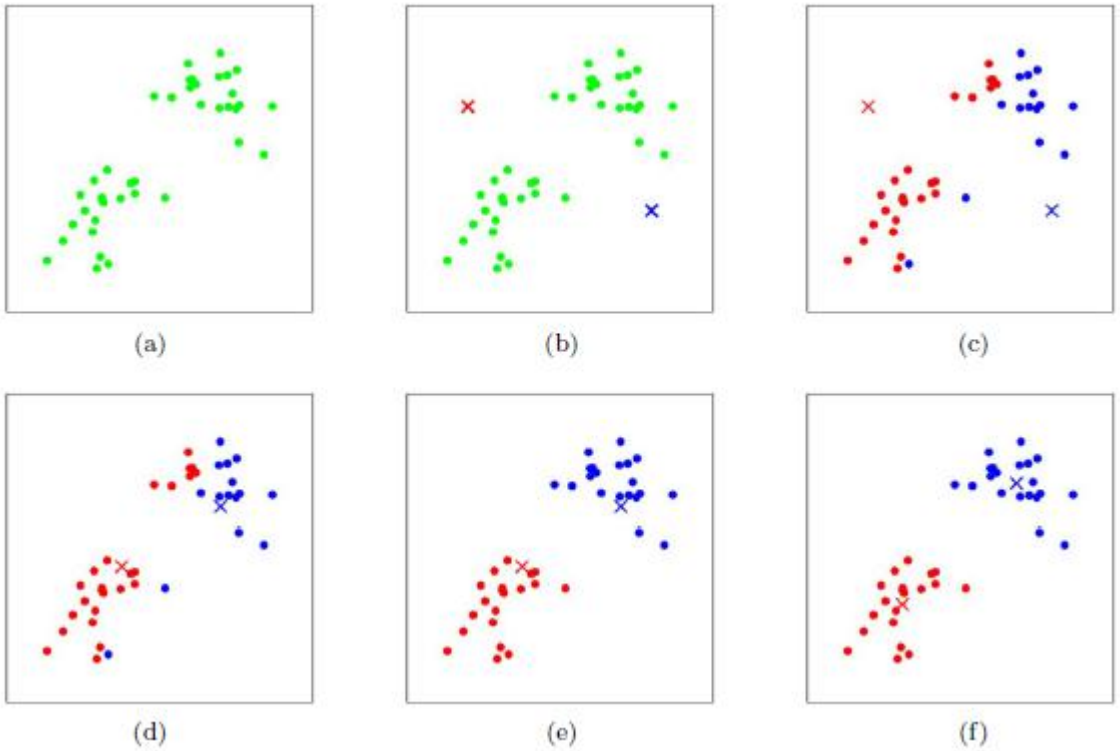
P , Q as the centre and K=2.

Step 3 again.

Final



Clustering finished !



主要步骤：

（1）选择k个中心点。

（2）计算每个点到中心点的距离，选择距离最近的中心点将其归类。

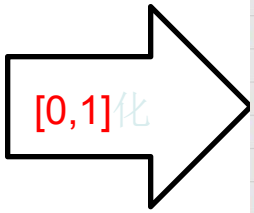
（3）更新中心点为每类的均值。

（4）重复（2）（3）迭代更新，直至新的中心点与之前中心点的距离小于某个值。

算法应用举例

亚洲15只球队在2005年-2010年间大型杯赛的战绩：

	A	B	C	D
1	中国	50	50	9
2	日本	28	9	4
3	韩国	17	15	3
4	伊朗	25	40	5
5	沙特	28	40	2
6	伊拉克	50	50	1
7	卡塔尔	50	40	9
8	阿联酋	50	40	9
9	乌兹别克斯坦	40	40	5
10	泰国	50	50	9
11	越南	50	50	5
12	阿曼	50	50	9
13	巴林	40	40	9
14	朝鲜	40	32	17
15	印尼	50	50	9



	A	B	C	D
1	中国	1	1	0.5
2	日本	0.3	0	0.19
3	韩国	0	0.15	0.13
4	伊朗	0.24	0.76	0.25
5	沙特	0.3	0.76	0.06
6	伊拉克	1	1	0
7	卡塔尔	1	0.76	0.5
8	阿联酋	1	0.76	0.5
9	乌兹别克斯坦	0.7	0.76	0.25
10	泰国	1	1	0.5
11	越南	1	1	0.25
12	阿曼	1	1	0.5
13	巴林	0.7	0.76	0.5
14	朝鲜	0.7	0.68	1
15	印尼	1	1	0.5

算法应用举例

用k-means算法进行聚类：

（1）设N=3，将15个球队分成3个等级

（2）随机抽取三个球队作为中心点

3x4 double

	1	2	3	4
1	3	0	0.1500	0.1300
2	4	0.2400	0.7600	0.2500
3	14	0.7000	0.6800	1
4				

抽取结果

本次选择韩国{0，0.15,0.13}、

伊朗{0.24,0.76,0.25}、

朝鲜{0.7,0.68,1}作为中心点。

算法应用举例

（3）计算所有球队与中心点的距离

	1	2	3	4
1	1	0.6651	0.6651	0.6651
2	2	1.1307	1.1307	1.1307
3	3	1.2360	1.2360	1.2360
4	4	0.8835	0.8835	0.8835
5	5	1.0247	1.0247	1.0247
6	6	1.0920	1.0920	1.0920
7	7	0.5886	0.5886	0.5886
8	8	0.5886	0.5886	0.5886
9	9	0.7543	0.7543	0.7543
10	10	0.6651	0.6651	0.6651
11	11	0.8688	0.8688	0.8688
12	12	0.6651	0.6651	0.6651
13	13	0.5064	0.5064	0.5064
14	14	0	0	0
15	15	0.6651	0.6651	0.6651

各球队与中心点的距离

根据最小的距离将球队分到各自的类别

求出均值得到新的中心点

根据新的中心点再次分类

算法应用举例

	1	2	3	4	5
1	1	1	1	0.5000	3
2	2	0.3000	0	0.1900	1
3	3	0	0.1500	0.1300	1
4	4	0.2400	0.7600	0.2500	2
5	5	0.3000	0.7600	0.0600	2
6	6	1	1	0	2
7	7	1	0.7600	0.5000	3
8	8	1	0.7600	0.5000	3
9	9	0.7000	0.7600	0.2500	2
10	10	1	1	0.5000	3
11	11	1	1	0.2500	2
12	12	1	1	0.5000	3
13	13	0.7000	0.7600	0.5000	3
14	14	0.7000	0.6800	1	3
15	15	1	1	0.5000	3

球队分类结果

迭代3次后得到：

第一类：日本、韩国

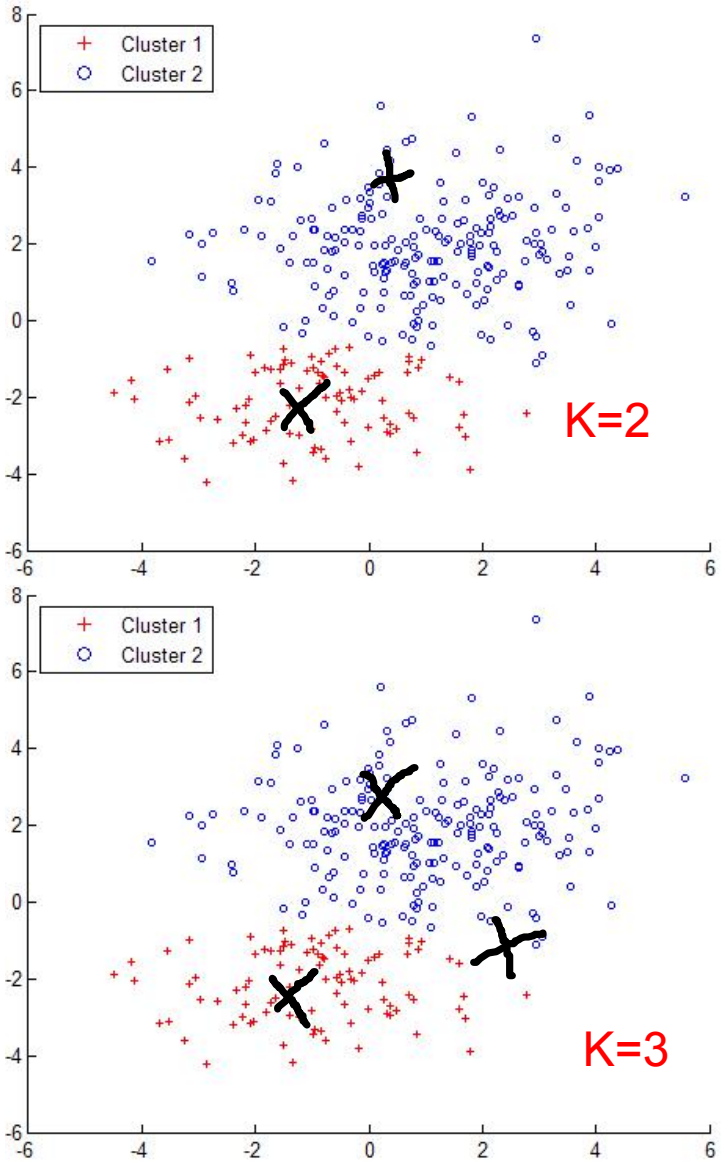
第二类：伊朗、沙特、伊拉克、乌兹别克斯坦、越南

第三类：中国、卡塔尔、阿联酋、泰国、阿曼、巴林、朝鲜、印尼

以上结果可以看出我国足球队处于亚洲三流水平。

Disadvantages

one of the main disadvantages to k-means is the fact that you must specify the number of clusters(K) as an input to the algorithm.As designed,the algorithm is not capable of determining the appropriate number of clusters and depends upon the user to identify this in advance.



初始中心点的选取

- 选择距离尽可能远的**K**个点。
- ①随机选择一个点作为初始中心点。
- ②选择距离该点最远的点作为第二个初始中心点。
- ③再选择距离前两个点的最近距离最大的点作为第三个初始的中心点。
- ④以此类推，直至选出**K**个中心点。