Cluster Analysis

Nathan Poslusny, Shanshan Li
Spring 2014
Instructor: Anita Wasilewska
Stony Brook University

References

1) Han Jiawei and Kamber Micheline. Data Mining: Concepts and Techniques, 2nd Edition. Morgan Kaufmann Publishers: San Francisco, CA, 2006

2) Eisen MB, Spellman PT, Brown PO, and Botstein D (1998) **Cluster analysis and display of genome-wide expression patterns**. *Proceedings of the National Academy of Sciences,* Vol. 95: 14863-14868.

3) Iyer VR, Eisen MB, Ross DT, Schuler G, Moore T, Lee JC, Trent JM, Staudt LM, Hudson J, Boguski MS, Lashkari D, Shalon D, Botstein D, Brown PO (1999) **The transcriptional program in the response of human fibroblasts to serum.** *Science* Vol 283: 83-87.

4) Iris flower data set. http://en.wikipedia.org/wiki/Iris_flower_data_set#cite_note-fisher36-1

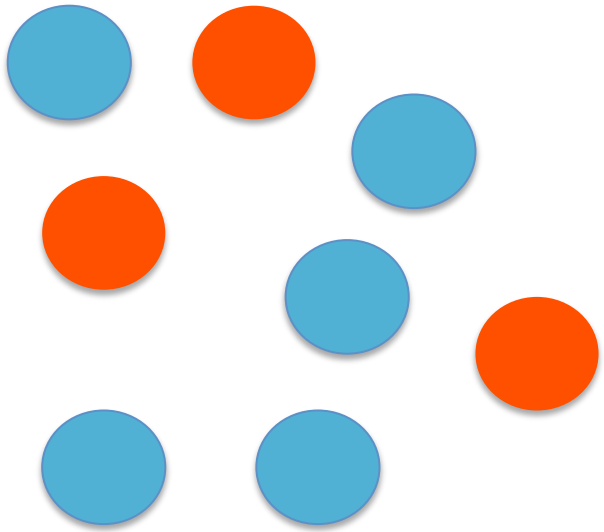
5) Saey TH (2010) More than a chicken, fewer than a grape. https://www.sciencenews.org/article/more-chicken-fewer-grape

6) Gene expression. http://en.wikipedia.org/wiki/Gene_expression
7) Nandita Das. Hedge Fund Classification using K-means Clustering Method.

Contents

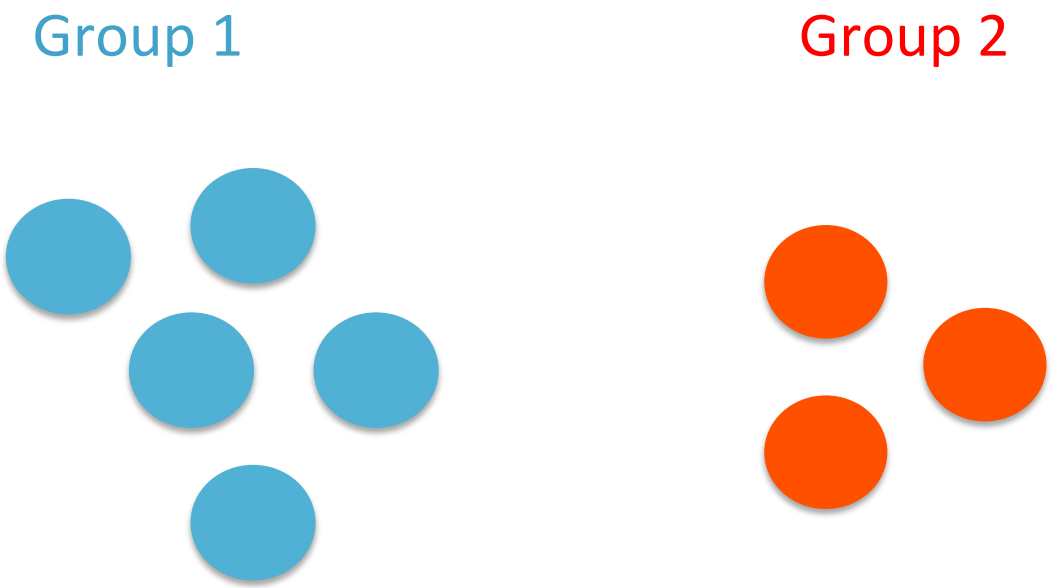
What is Cluster Analysis?

Clustering: The process of grouping a set of physical or abstract objects into classes of similar objects



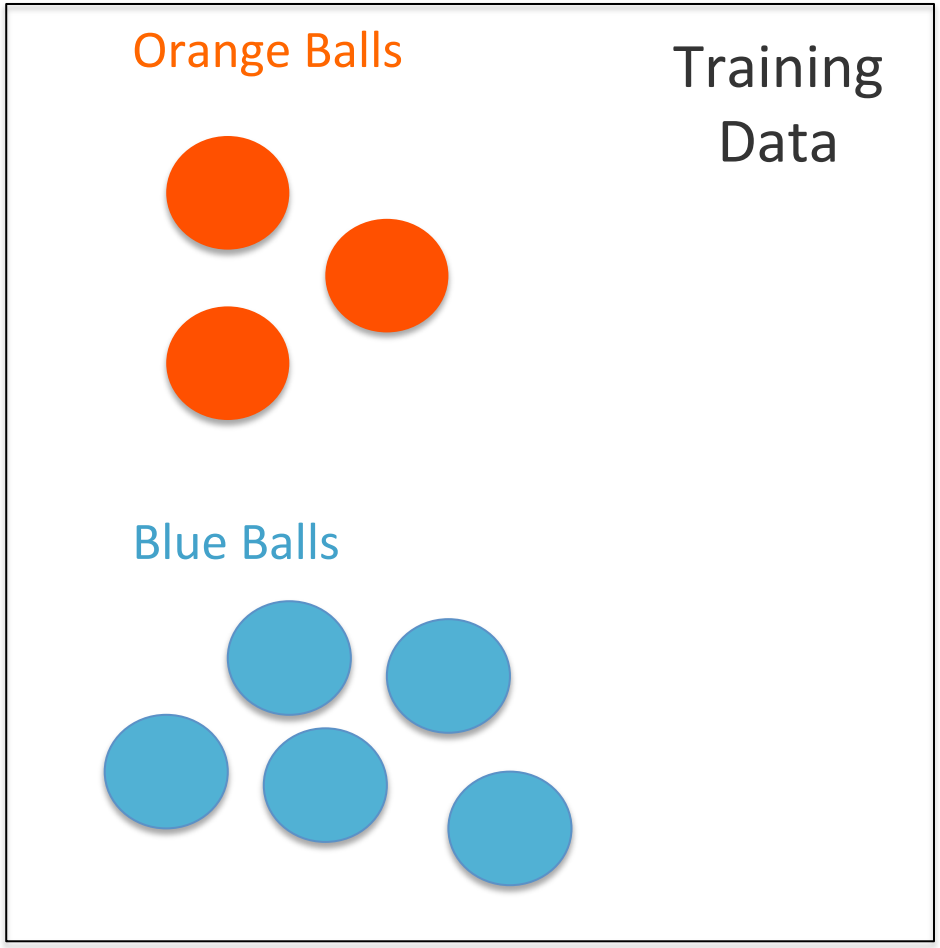
What is Cluster Analysis?

Clustering: The process of grouping a set of physical or abstract objects into classes of similar objects



Clustering versus Classification

Classification is supervised learning with pre-defined classes and class-labeled training data.

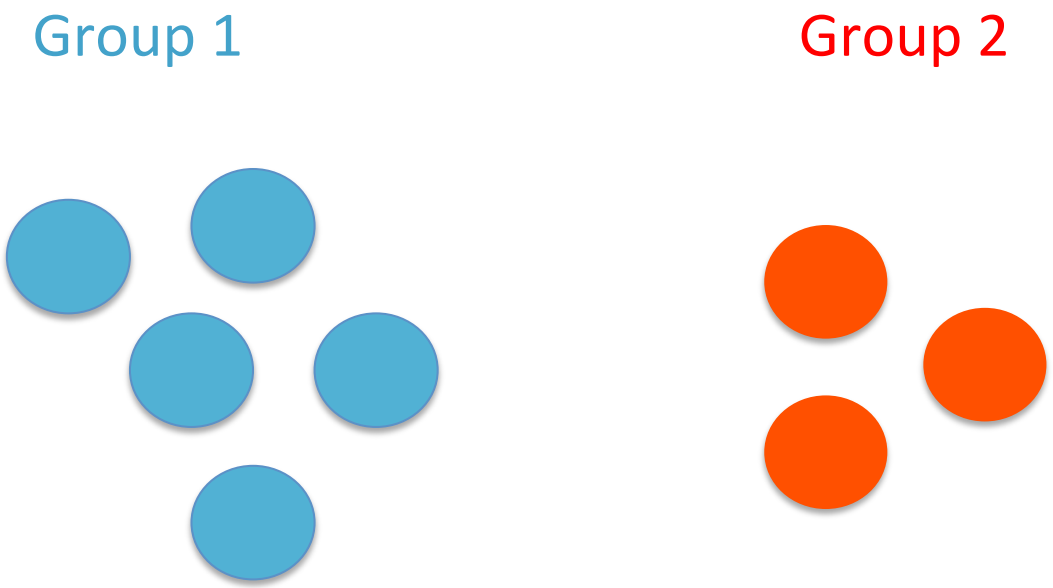


Test Data

Orange or Blue?

What is Cluster Analysis?

Clustering is unsupervised learning with no labeled training data or pre-defined classes

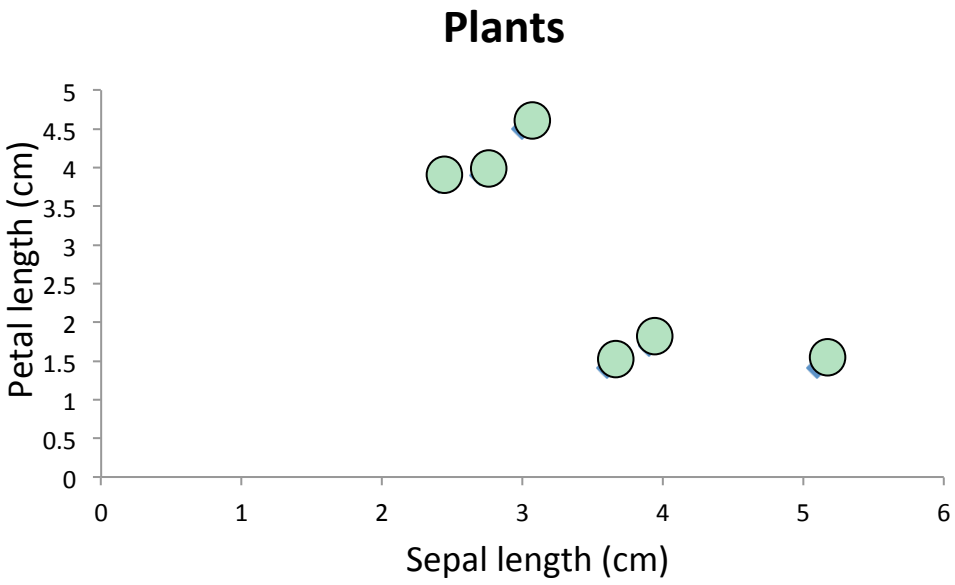


Usual clustering data

Notice no class attribute!

Plant ID	Sepal Length	Petal Length
Plant #1	5.1	1.4
Plant #2	3.6	1.4
Plant #3	3.9	1.7
Plant #4	2.4	3.8
Plant #5	2.7	3.9
Plant #6	3.0	4.5

Data from [4]

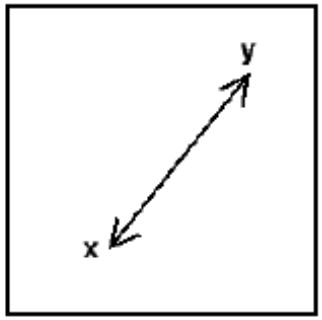


How to compute similarity or dissimilarity between objects/records?

Distance Measurements

Euclidean distance:

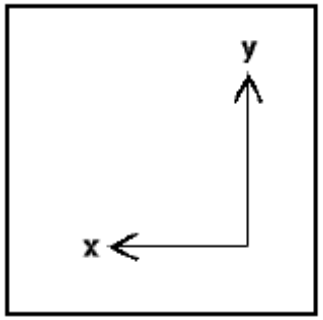
$$d(i,j) = \sqrt{\left((x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2 + \ldots (x_{jn} - x_{nj})^2\right)}$$



Euclidean

Manhattan distance:

$$d(i,j) = |x_{i1} - x_{j1}| + |x_{i2} - x_{j2}| + \ldots |x_{in} - x_{jn}|$$



Manhattan

Requirements of distance function

1) d(i,j) >= 0: Distance is a nonnegative number

2) d(i,i) = 0: The distance of an object to itself is 0

3) d(i,j) = d(j,i): Distance is a symmetric function

4) d(i,j) <= d(i,h) + d(h,j): Going directly from object i to object j in space is no more than making a detour over any other object h (triangular inequality)

What is clustering good for?

- Reduce dimensionality of data by forming clusters (pre-processing step)
- Used for outlier detection (detect values far from a cluster)
- Stand-alone tool to gain insight into
  - Distribution of data
  - Observe characteristics of each cluster
  - Find particularly interesting clusters for follow-up analysis

Major Clustering Methods

- Partitioning Methods
- Hierarchical Methods
- Density-based Methods
- Model-based Methods

Partitioning Methods

- It classifies the database of n objects into k groups, which together satisfy the following requirements:

  (1) Each group must contain at least one object, and

  (2) Each object must belong to exactly one group.

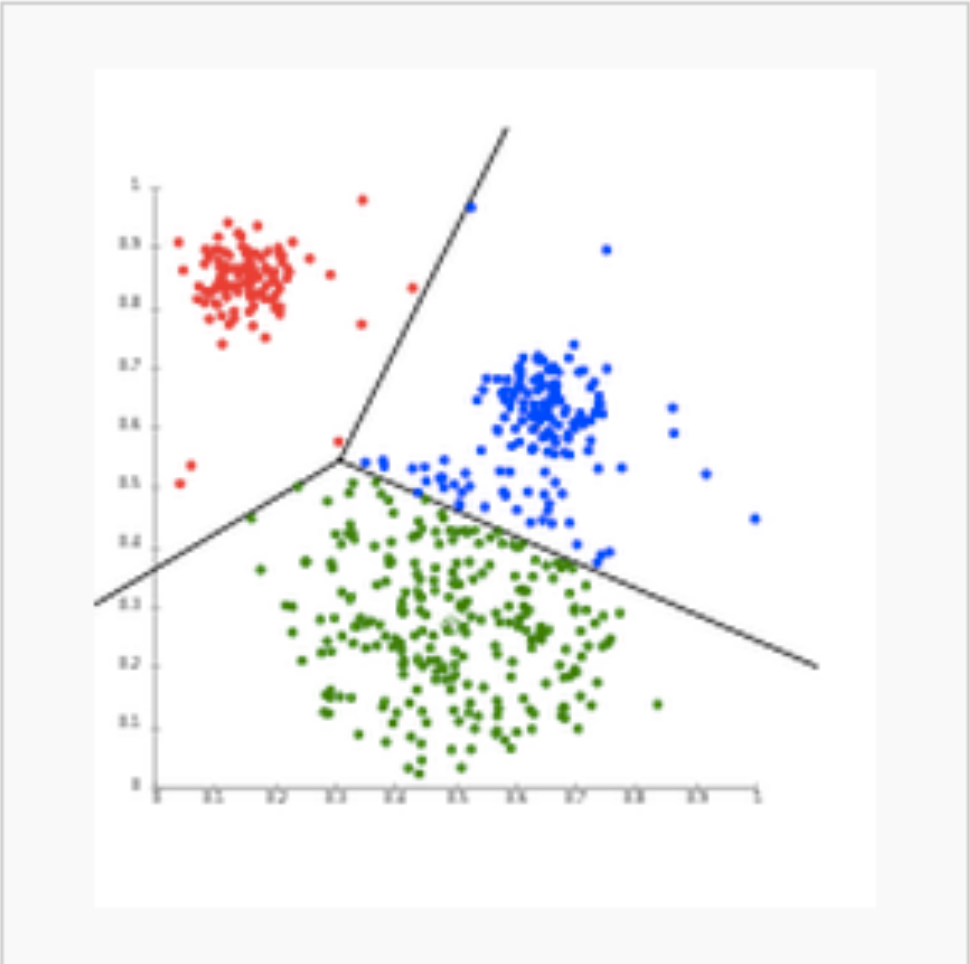
Partitioning Methods

- Given k , a partitioning method creates an initial partitioning. It then uses an iterative relocation technique that attempts to improve the partitioning by moving objects from one group to another
- The general criterion of a good partitioning is that objects in the same cluster are “close” or related to each other, whereas objects of different clusters are “far apart” or very different.

Partitioning Methods

- To achieve global optimality, there are two popular methods
  (1) the k-means algorithm,
  where each cluster is represented by the mean value of the objects in the cluster, and
  (2) the k-medoids algorithm,
  where each cluster is represented by one of the objects located near the center of the cluster.

Partitioning Methods



K-means separate data (not adequate here)

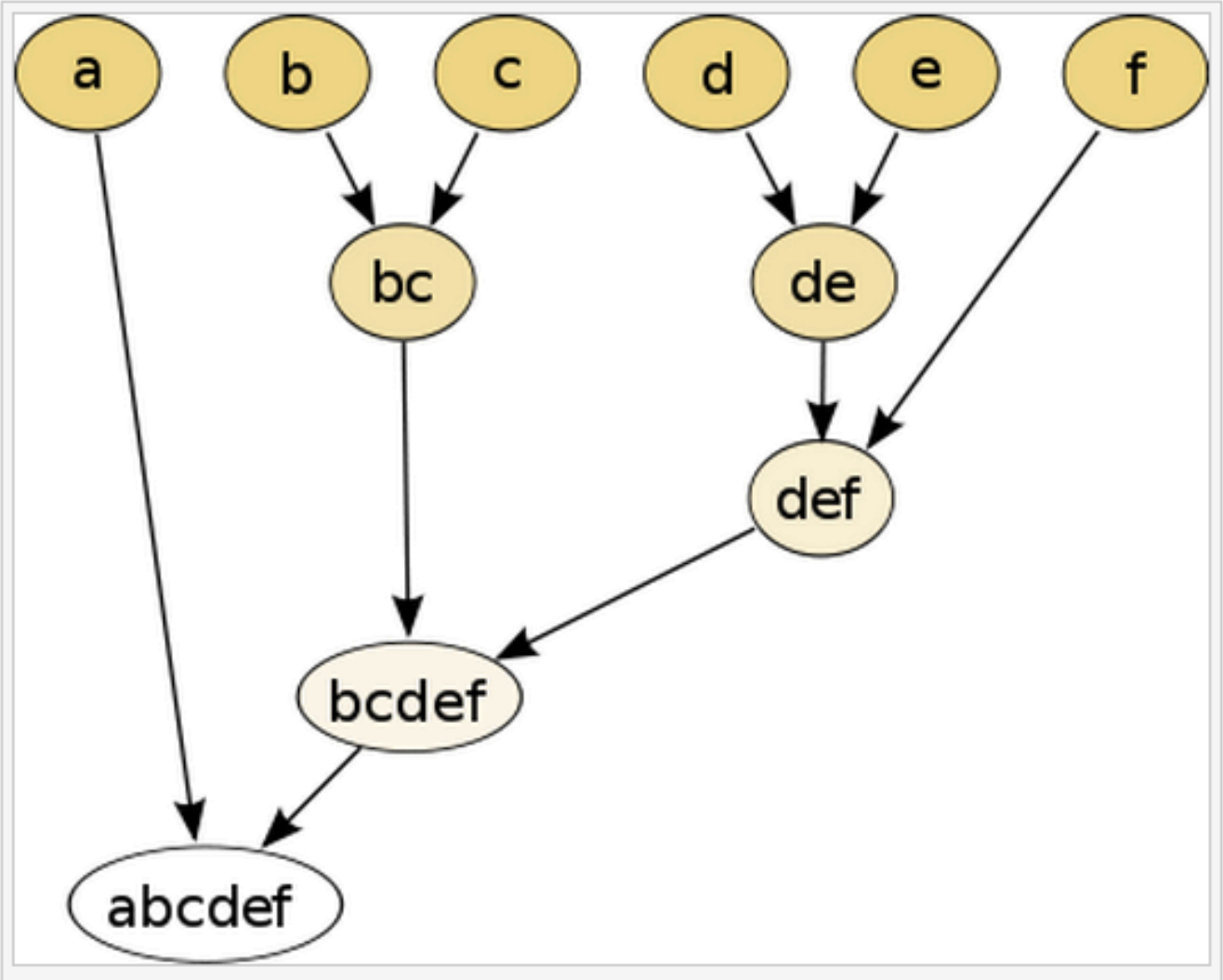
Hierarchical Methods

- A hierarchical method creates a hierarchical decomposition of the given set of data objects
- A hierarchical method can be classified as being either agglomerative or divisive , based on how the hierarchical decomposition is formed.

Hierarchical Methods

- The agglomerative approach , also called the bottom-up approach, starts with each object forming a separate group. It successively merges the objects or groups that are close to one another, until all of the groups are merged into one (the topmost level of the hierarchy), or until a termination condition holds.

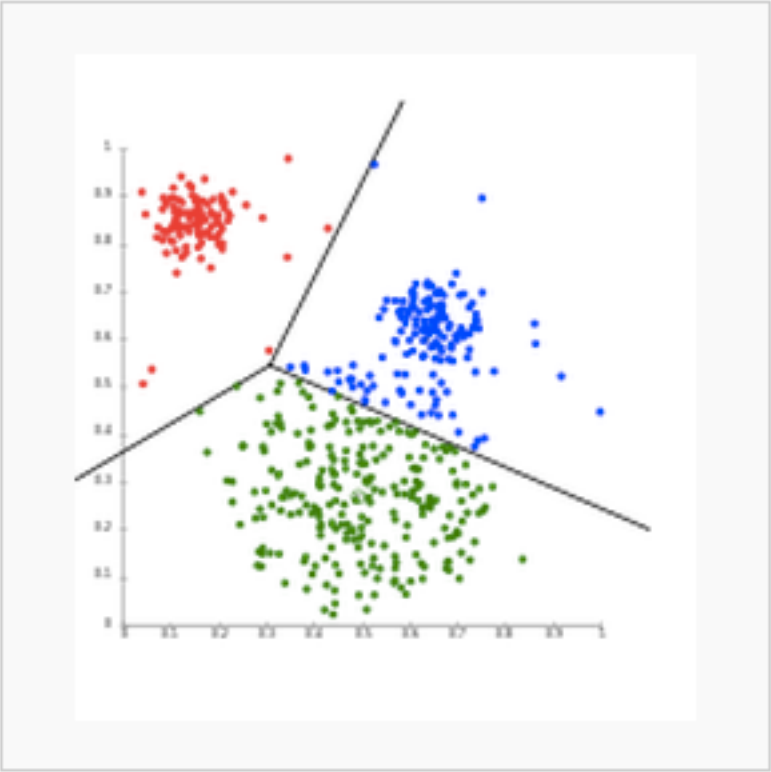
Hierarchical Methods

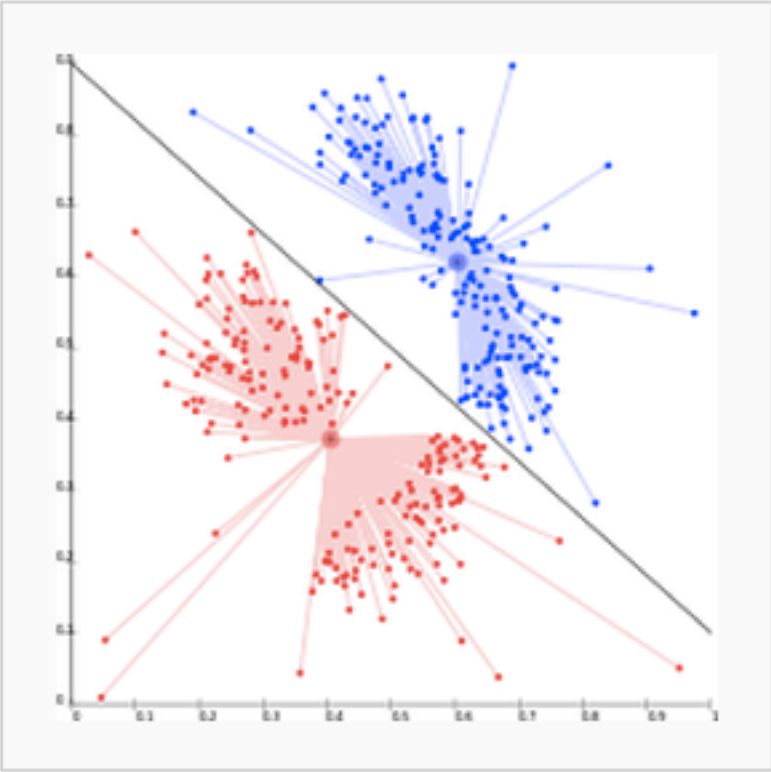


Hierarchical Methods

- The divisive approach , also called the top-down approach, starts with all of the objects in the same cluster. In each successive iteration, a cluster is split up into smaller clusters, until eventually each object is in one cluster, or until a termination condition holds.

Density-based Methods



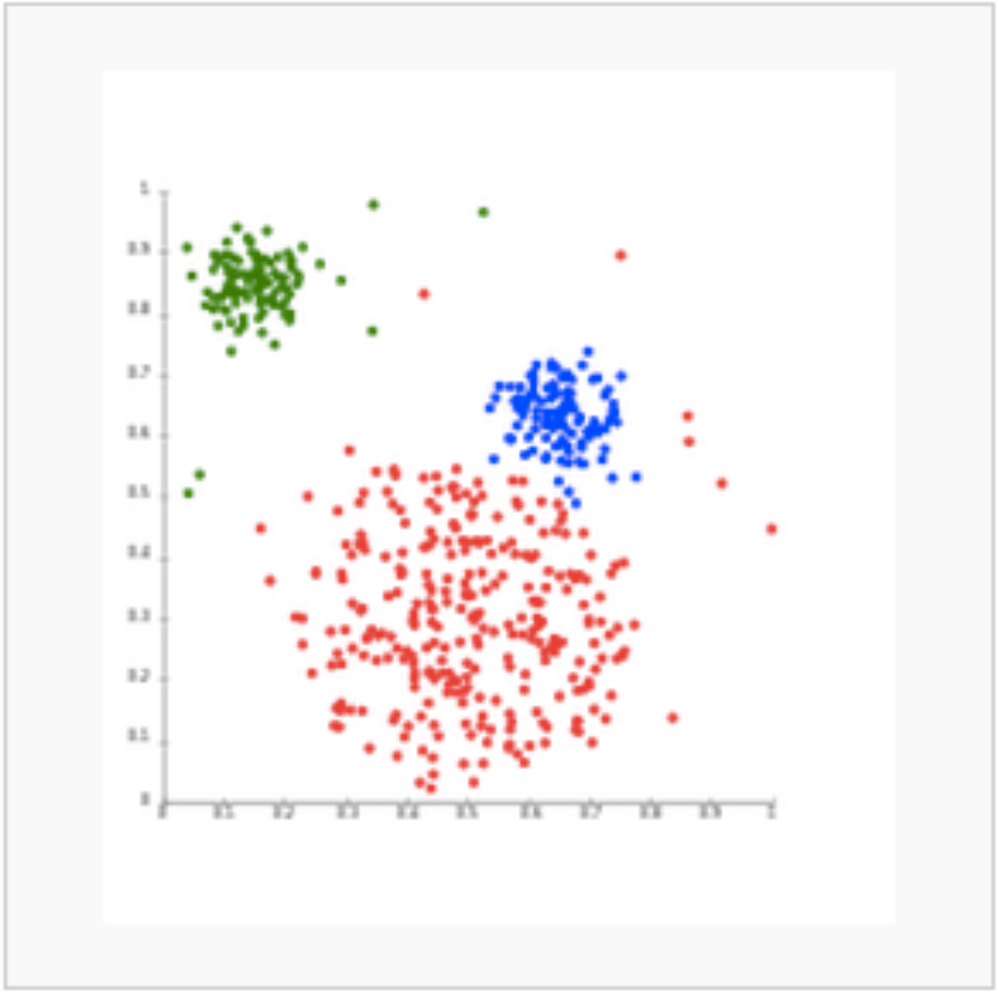


K-means can not represent density-based models

Density-based Methods

- The general idea is to continue growing the given cluster as long as the density (number of objects or data points) in the "neighborhood" exceeds some threshold
- That is, for each data point within a given cluster, the neighborhood of a given radius has to contain at least a minimum number of points
- Can be used to filter out noise (outliers) and discover clusters of arbitrary shape

Density-based Methods



Keep different densities

Model-based Methods

- Most closely related to statistics
- Clusters can then easily be defined as objects belonging most likely to the same distribution
- This methods hypothesize a model for each of the clusters and find the best fit of the data to the given model.

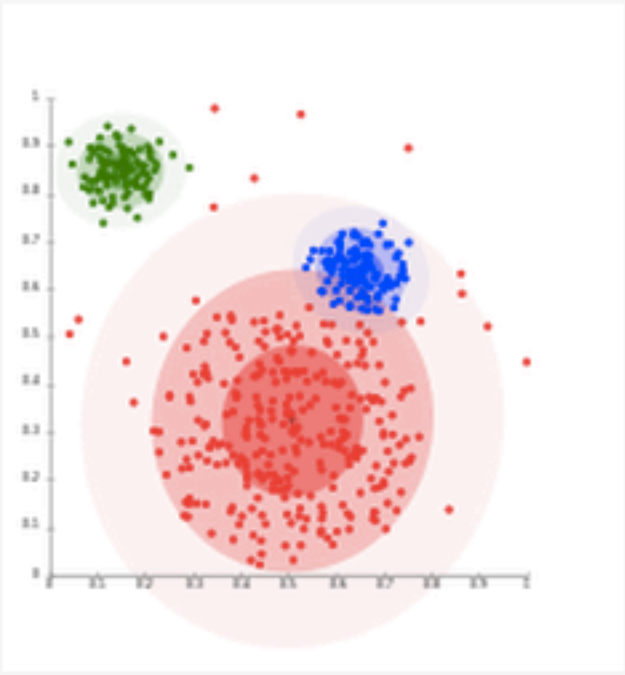
Model-based Methods

- Most closely related to statistics
- Clusters can then easily be defined as objects belonging most likely to the same distribution
- This methods hypothesize a model for each of the clusters and find the best fit of the data to the given model

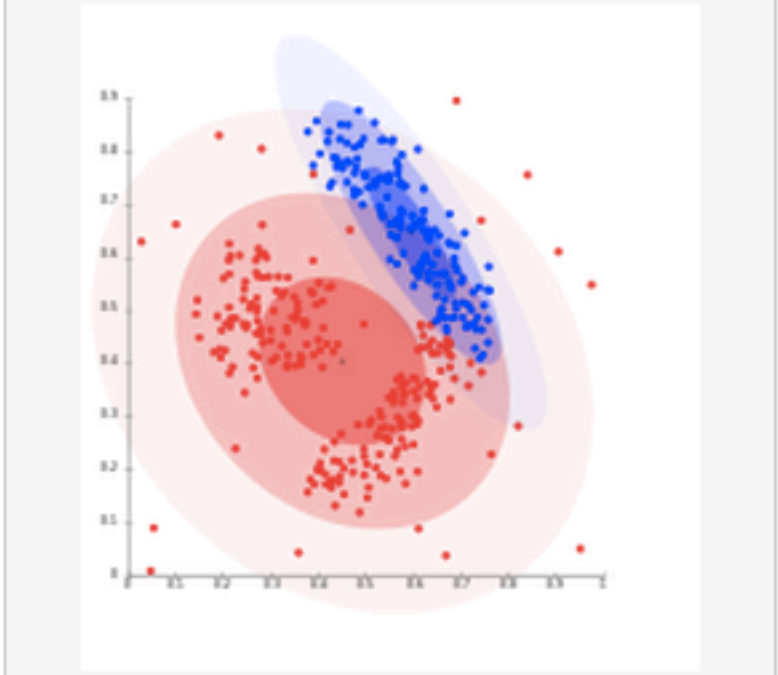
Model-based Methods

- One prominent method is known as Gaussian mixture models（using the expectation–maximization algorithm)
- The data set is usually modelled with a fixed number of Gaussian Distributions that are initially randomly
- The parameter of the models are iteratively optimized to fit better to the data set
- This will converge to a local optimum

Model-based Methods



On Gaussian-distributed data, EM works well, since it uses Gaussians for modelling clusters



Density-based clusters cannot be modeled using Gaussian distributions

Choice of Clustering Algorithms

- The choice of clustering algorithm depends both on the type of data available and on the particular purpose of the application
- If cluster analysis is used as a descriptive or exploratory tool, it is possible to try several algorithms on the same data to see what the data may disclose.

Choice of Clustering Algorithms

- Some clustering algorithms integrate the ideas of several clustering methods, so that it is sometimes difficult to classify a given algorithm as uniquely belonging to only one clustering method category.
- Some applications may have clustering criteria that require the integration of several clustering techniques

Contents

Description of K-means algorithm

Objective: Given a data set of n objects, and k, number of clusters to form, organize the objects into k partitions or clusters.

The clusters are formed to minimize an objective partitioning critieron.

The objects within a cluster are “similar” whereas objects of different clusters are “dissimiliar” in terms of set data attributes.

Description of K-means algorithm

Objective: Given a data set of n objects, and k, number of clusters to form, organize the objects into k partitions or clusters.

The clusters are formed to minimize an objective partitioning critieron.

The objects within a cluster are “similar” whereas objects of different clusters are “dissimiliar” in terms of set data attributes.

Square-error criterion

$$E = \sum_{i=1}^{k} \sum_{p \in Ci} |p - m_i|^2$$

E = sum of the square error for all objects
p = point in space representing a given object
m = mean of cluster Ci

Description of K-means algorithm

Objective: Given a data set of n objects, and k, number of clusters to form, organize the objects into k partitions or clusters.

The clusters are formed to minimize an objective partitioning critieron.

The objects within a cluster are "similar" whereas objects of different clusters are "dissimiliar" in terms of set data attributes.

Square-error criterion

$$E = \sum_{i=1}^{k} \sum_{p \in Ci} |p - m_i|^2$$

E = sum of the square error for all objects
p = point in space representing a given object
m = mean of cluster Ci

In Words
For each object in each cluster, the distance from the object to its cluster center is squared, and the distances are summed.

K-means algorithm

Input:
k: number of clusters
D: a data set containing n objects

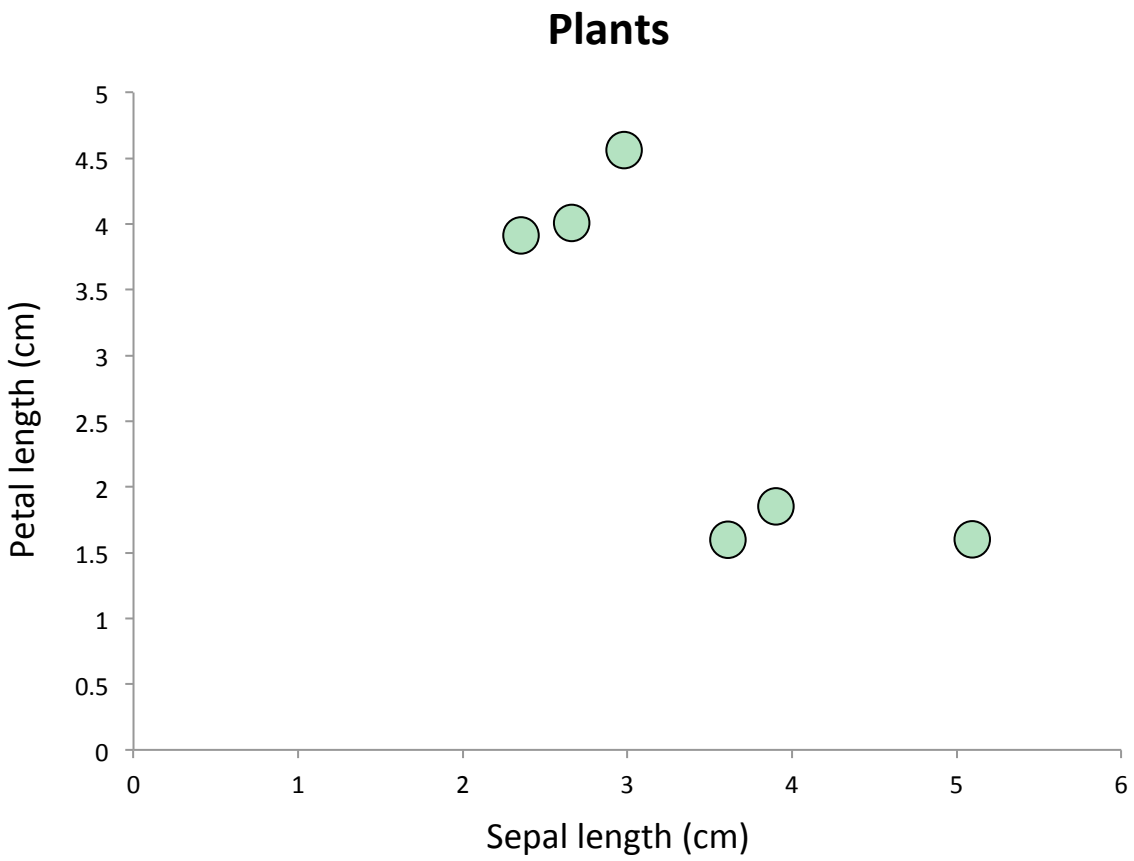
Output:
A set of k clusters

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 most similar (based on euclidean distance between object and cluster center)
4) update the cluster means, i.e. calculate the mean value of the objects for each cluster
5) **Until no change;**

K-means worked example

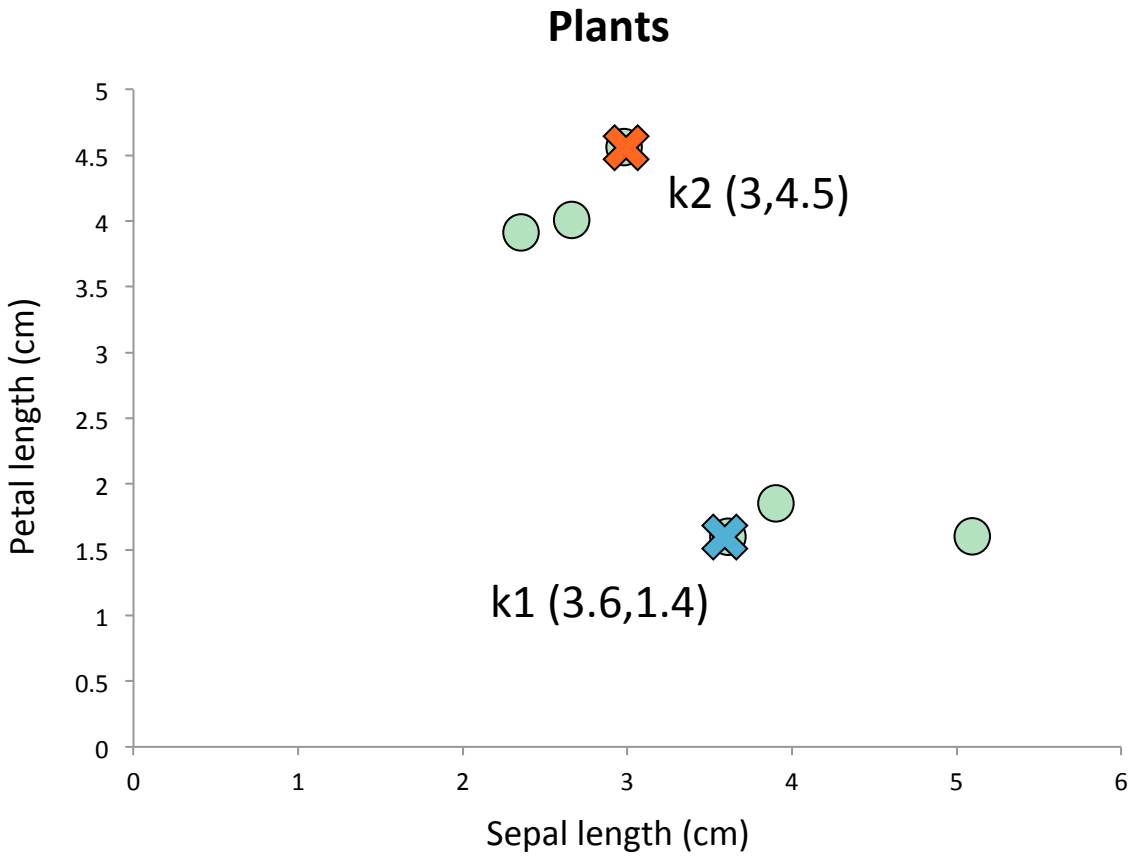
Plant ID	Sepal Length	Petal Length
Plant #1	5.1	1.4
Plant #2	3.6	1.4
Plant #3	3.9	1.7
Plant #4	2.4	3.8
Plant #5	2.7	3.9
Plant #6	3.0	4.5

K-means worked example



K-means worked example

Step 1) Arbitrarily choose k objects from D as the initial cluster centers



K-means worked example

Step 2) Assign each object to the cluster to which the object is most similar based on Euclidean distance between object and cluster center.

Cluster centers
k_1: (3.6,1.4)
k_2: (3,4.5)

Plant	Sepal Length	Petal Length	dist(k_1)	dist(k_2)
#1	5.1	1.4	1.50	3.75
#2	3.6	1.4	0.00	3.16
#3	3.9	1.7	0.62	2.94
#4	2.4	3.8	1.96	0.92
#5	2.7	3.9	1.82	0.67
#6	3.0	4.5	1.86	0.00

K-means worked example

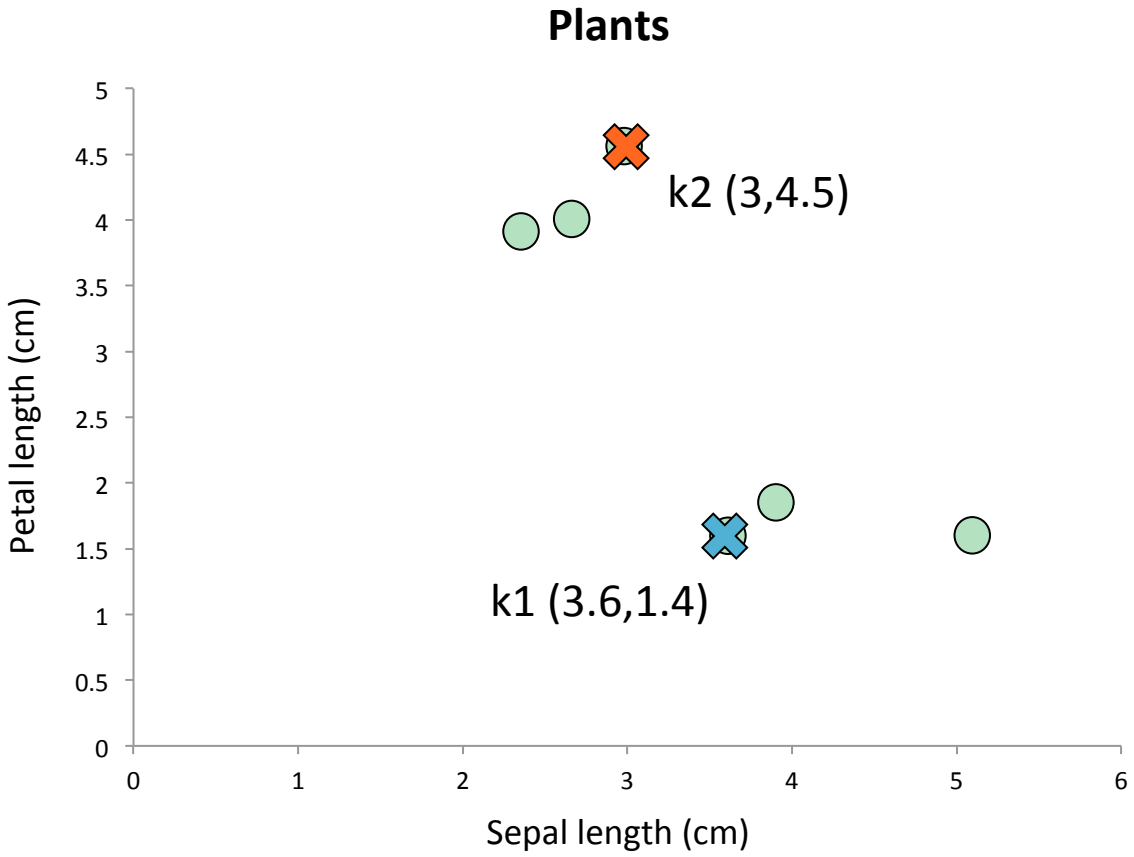
Step 2) Assign each object to the cluster to which the object is most similar based on Euclidean distance between object and cluster center.

Cluster centers
k_1: (3.6,1.4)
k_2: (3,4.5)

Plant	Sepal Length	Petal Length	dist(k_1)	dist(k_2)
#1	5.1	1.4	1.50	3.75
#2	3.6	1.4	0.00	3.16
#3	3.9	1.7	0.62	2.94
#4	2.4	3.8	1.96	0.92
#5	2.7	3.9	1.82	0.67
#6	3.0	4.5	1.86	0.00

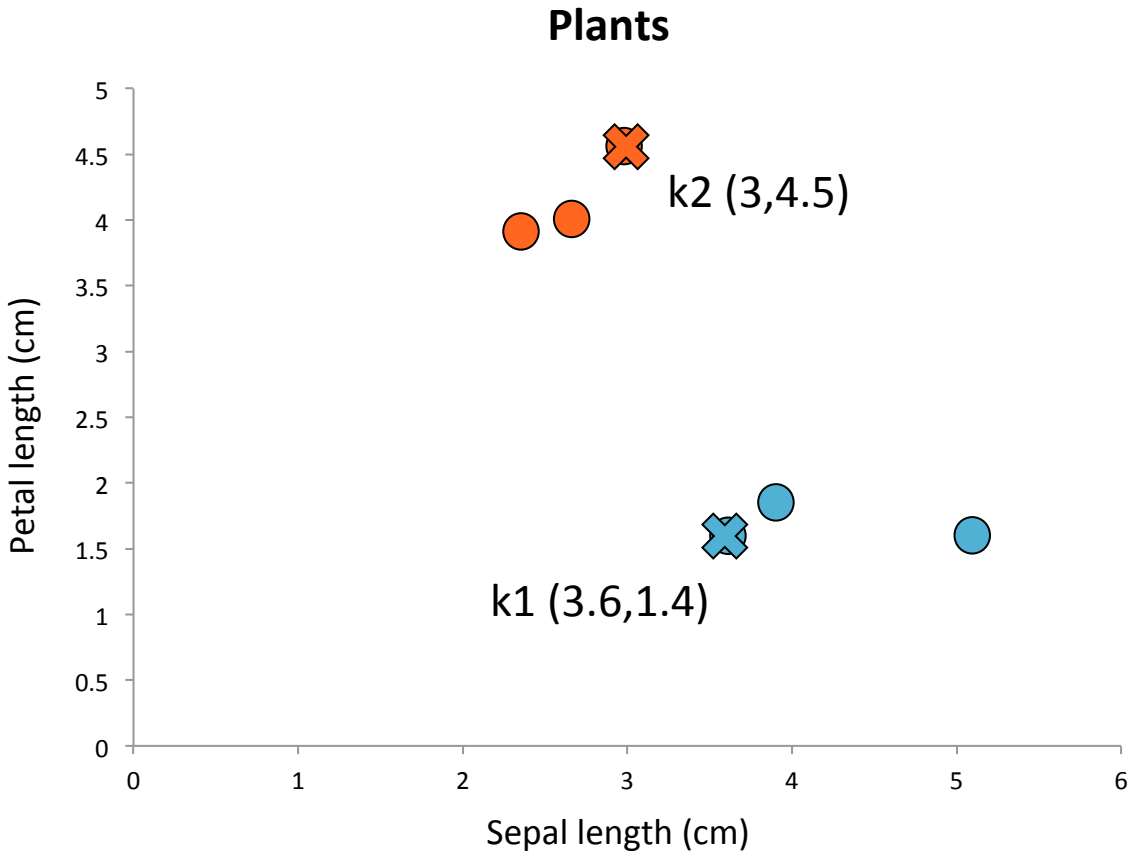
K-means terminates at local optimum

Step 2) Assign each object to the cluster to which the object is most similar based on Euclidean distance between object and cluster center.



K-means terminates at local optimum

Step 2) Assign each object to the cluster to which the object is most similar based on Euclidean distance between object and cluster center.



K-means worked example

Step 3) Update the cluster means, i.e. calculate the mean value of the objects for each cluster

Group #1	Sepal	Petal
#1	5.1	1.4
#2	3.6	1.4
#3	3.9	1.7
Mean	4.2	1.5

Group #2	Sepal	Petal
#1	2.4	3.8
#2	2.7	3.9
#3	3.0	4.5
Avg	2.7	4.1

K-means worked example

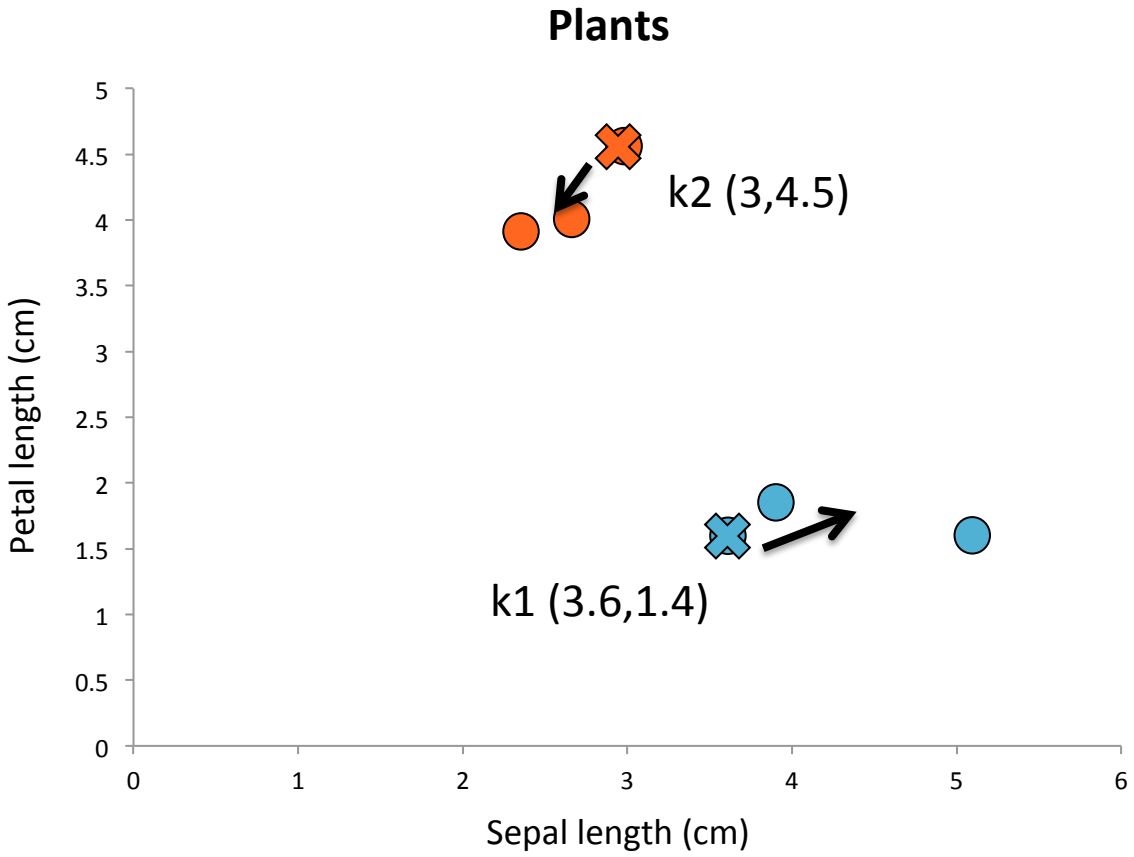
Step 3) Update the cluster means, i.e. calculate the mean value of the objects for each cluster

Group #1	Sepal	Petal
#1	5.1	1.4
#2	3.6	1.4
#3	3.9	1.7
Mean	4.2	1.5

Group #2	Sepal	Petal
#1	2.4	3.8
#2	2.7	3.9
#3	3.0	4.5
Avg	2.7	4.1

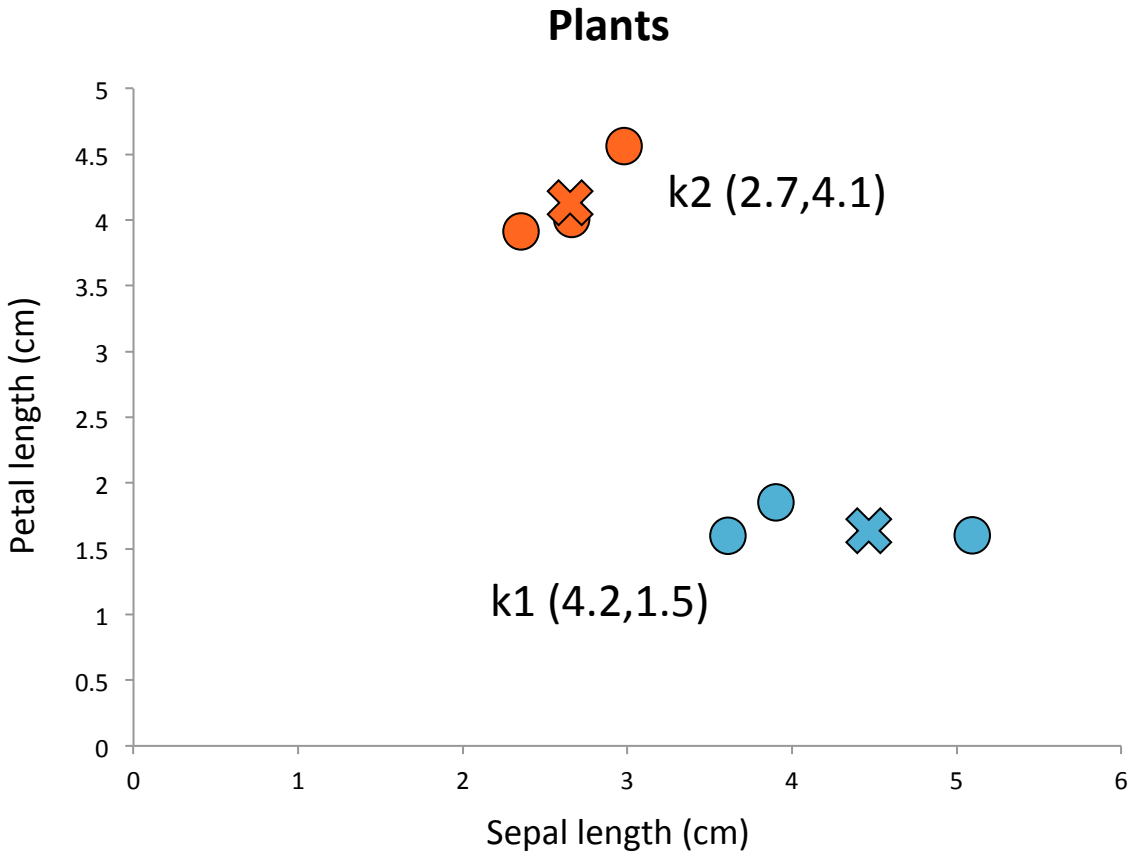
K-means terminates at local optimum

Step 3) Update the cluster means, i.e. calculate the mean value of the objects for each cluster



K-means terminates at local optimum

Step 3) Update the cluster means, i.e. calculate the mean value of the objects for each cluster



K-means worked example

Step 2) Re-assign each object to the cluster to which the object is most similar based on Euclidean distance between object and cluster center.

Cluster centers
k_1: (4.1,1.5)
k_2: (2.7,4.1)

Plant	Sepal Length	Petal Length	dC1	dC2
#1	5.1	1.4	0.91	3.58
#2	3.6	1.4	0.61	2.81
#3	3.9	1.7	0.36	2.65
#4	2.4	3.8	2.92	0.40
#5	2.7	3.9	2.83	0.17
#6	3.0	4.5	3.23	0.52

K-means worked example

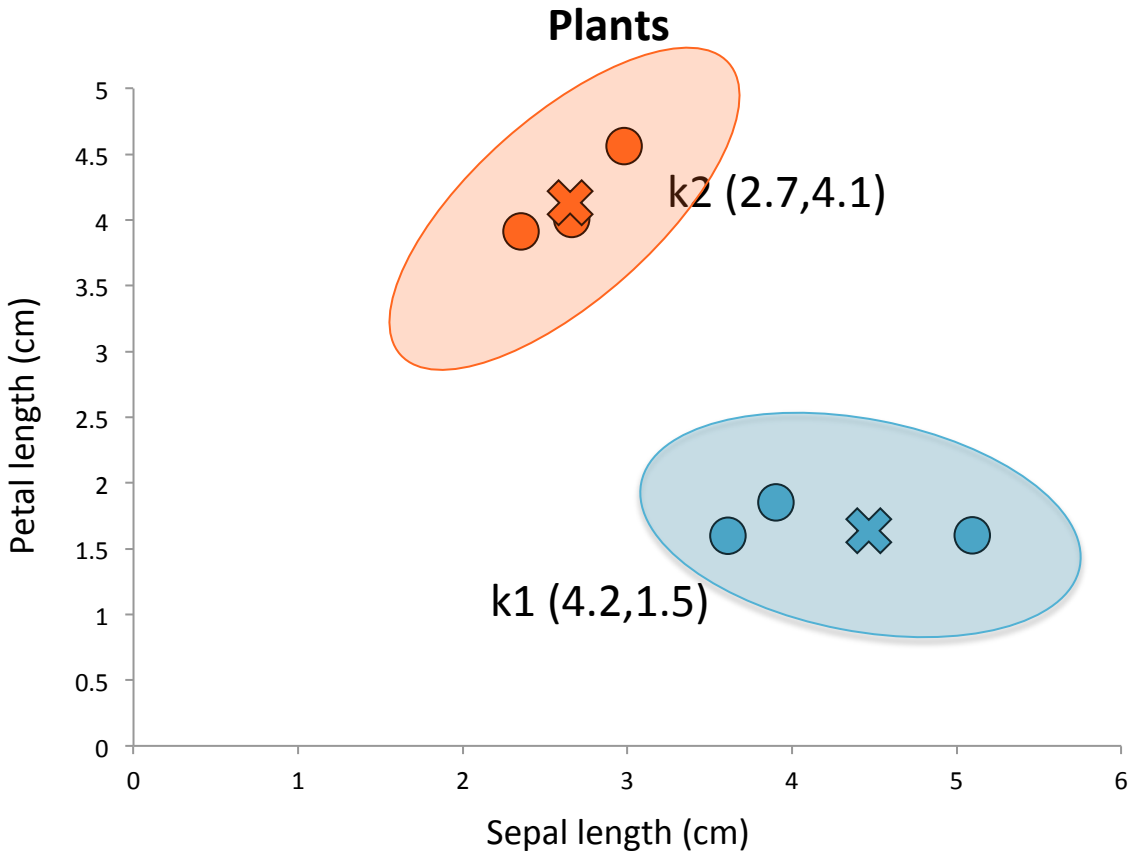
Step 3) Update the cluster means, i.e. calculate the mean value of the objects for each cluster

Group #1	Sepal	Petal
#1	5.1	1.4
#2	3.6	1.4
#3	3.9	1.7
Mean	4.2	1.5

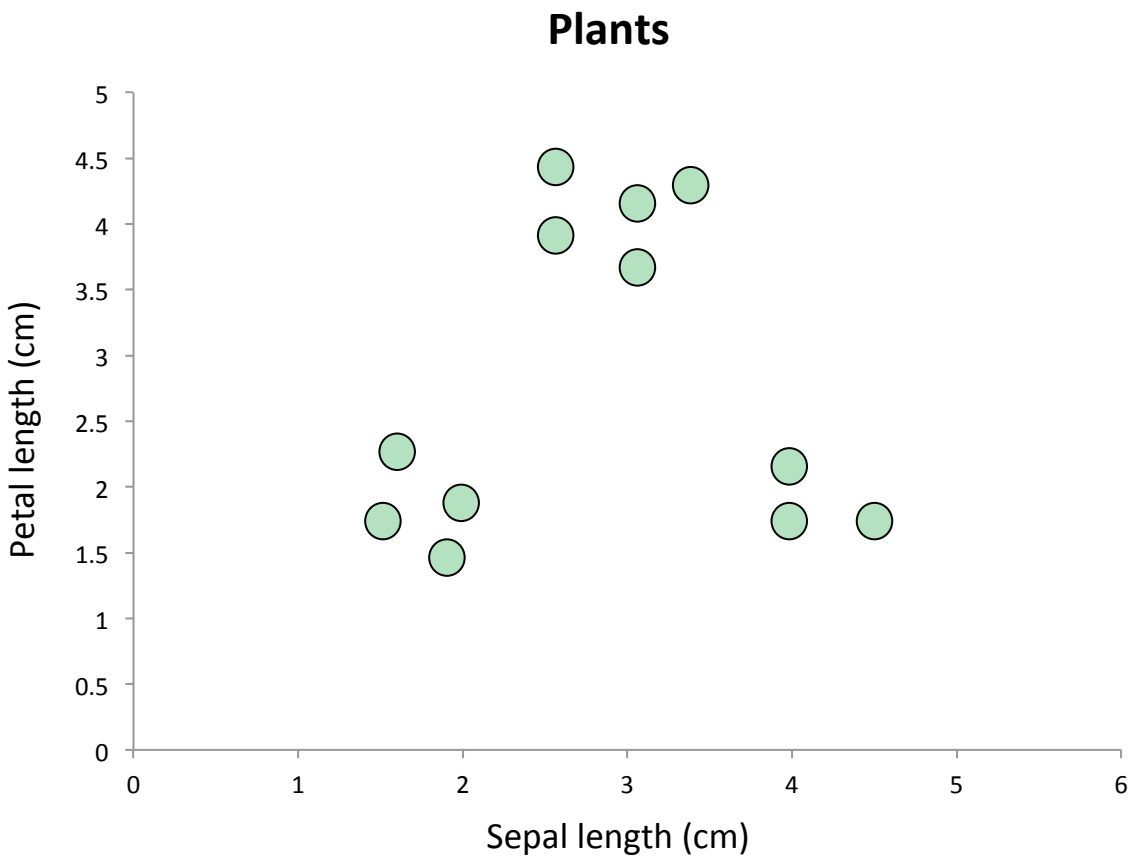
Group #2	Sepal	Petal
#1	2.4	3.8
#2	2.7	3.9
#3	3.0	4.5
Avg	2.7	4.1

K-means worked example

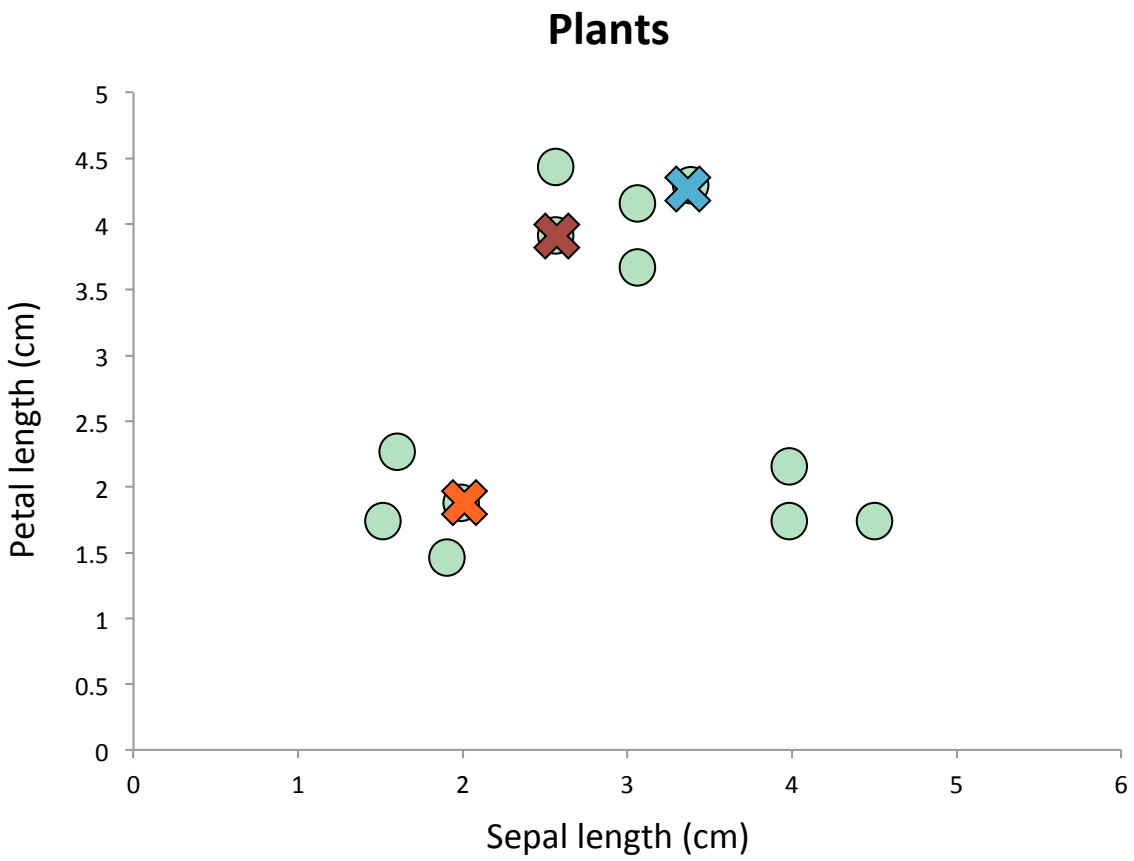
Step 3) Update the cluster means, i.e. calculate the mean value of the objects for each cluster



K-means terminates at local optimum

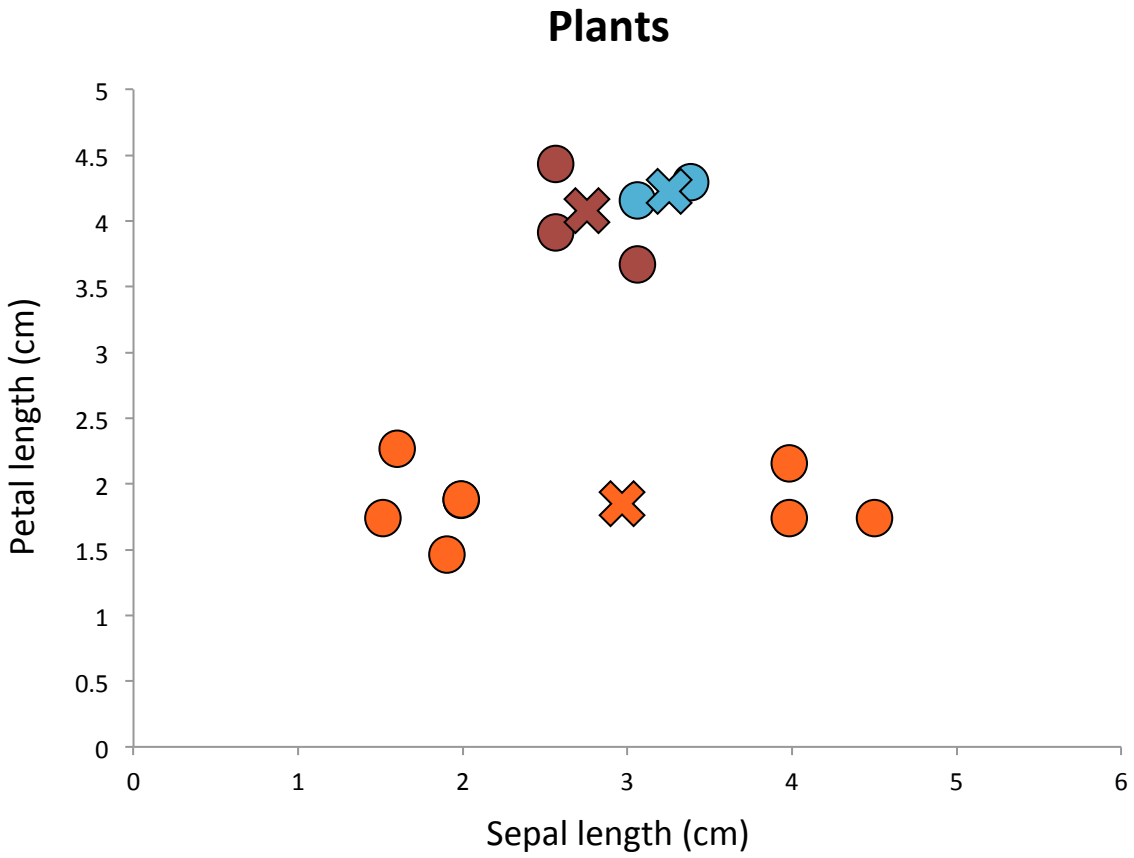


K-means terminates at local optimum



K-means terminates at local optimum

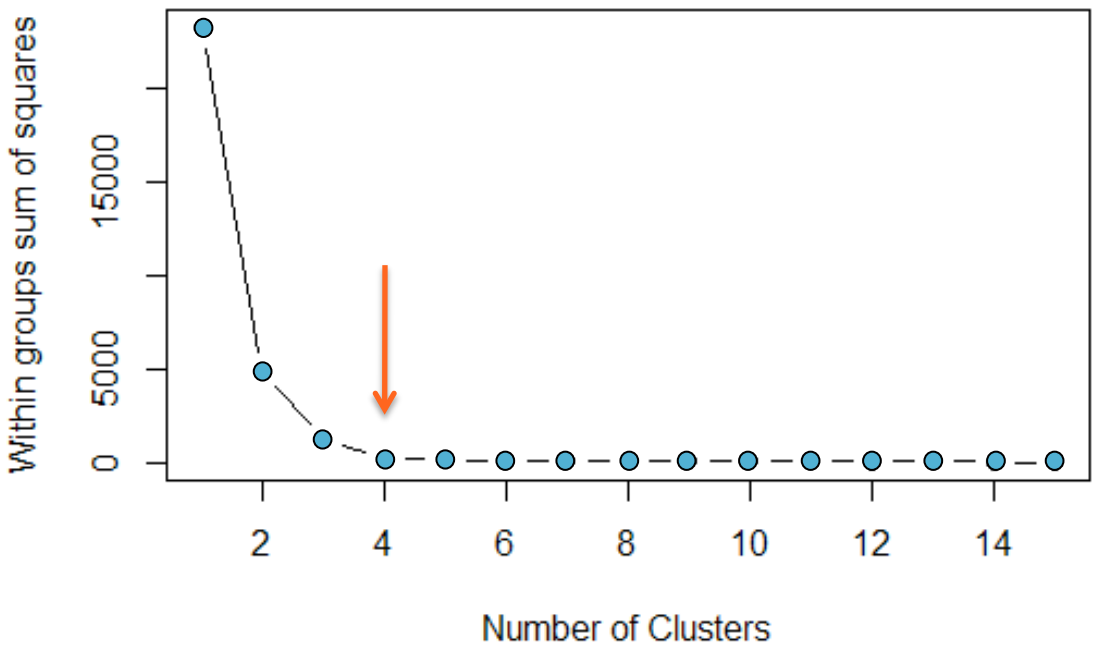
Solution: Run k-means multiple times with random starting points



How many clusters?

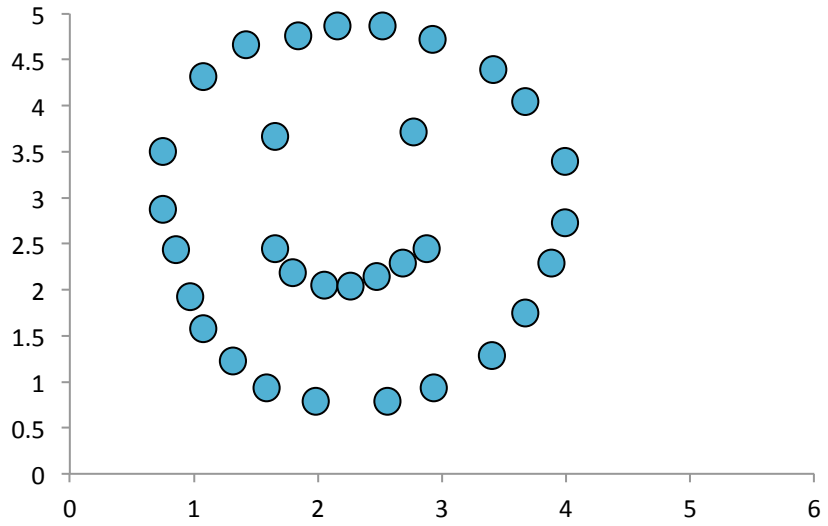
- The correct choice of k is often ambiguous and may require a priori knowledge

- One solution is called the "elbow" approach



Other K-means weaknesses

- Applicable only when mean is defined, what about categorical data?
- Unable to handle noisy data and outliers
- Not suitable to discover clusters with non-convex shapes



Contents

K-means Application in Finance

"Hedge Fund Classification Using K-means Clustering Method"

Nandita Das

9th International Conference on Computing in Economics and Finance

K-means Application in Finance

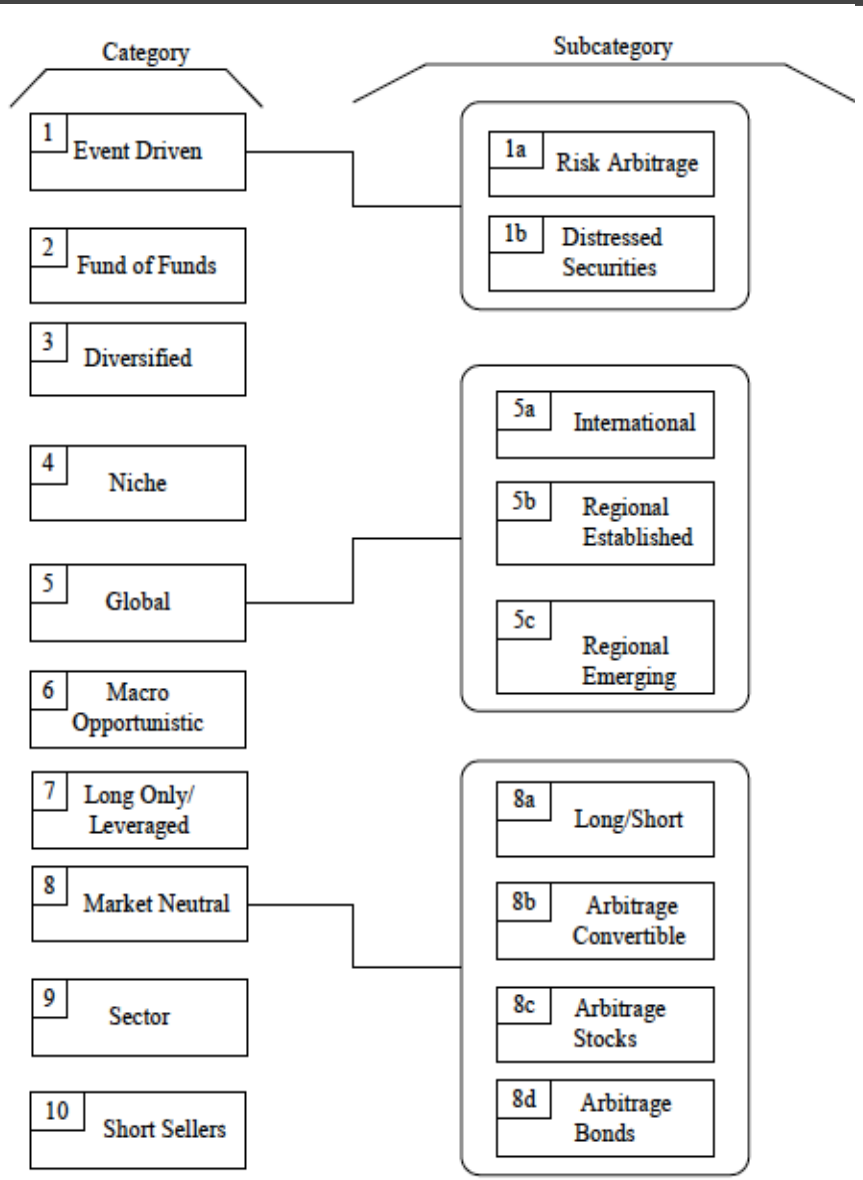
- What is Hedge Fund?
  A wide range of investments vehicles that can vary substantially in terms of size, trading strategy, and organizational structures

- Existing Classification of Hedge Fund
  Zurich Capital Markets(ZCM/Hedge) database provides a comprehensive coverage of global hedge funds. The database classifies hedge funds into four general classes and ten broad categories of investment styles.

K-means Application in Finance

- Onshore hedge fund (HF-US)
- Offshore hedge fund (HF-NON)
- Onshore fund-of-funds (FOF-US)
- Offshore fund-of-funds (FOF-NON)

K-means Application in Finance



K-means Application in Finance

- There is a need for "alternative approach" to hedge fund classifications because of the lack of "pure" hedge fund types that exist in the industry
- Why?
  Performance comparison of various hedge funds with the existing hedge fund return index is not appropriate since a particular hedge fund could be classified in two or more classes

K-means Application in Finance

Choice of Attributes:

- Asset Class

  Stocks, bonds, currency and derivatives
- Size

  The net asset value could be a measure of size
- Fee

  The hedge fund manager's share in the fund's profit
- Leverage

  Varies from zero to 70 times the asset value
- Liquidity

  Redemption frequency is considered as a measure of liquidity

K-means Application in Finance

Convert all the quantitative attributes into qualitative attributes using multi-valued logic

Values	Number of Funds	Grouping	Logical State	Number of Funds
Panel A. Incentive Fee				
0%	140	Less than equal to 10%	1	351
0.2-8%	28			
9-10%	183			
12-15%	11	Greater than 10% and less than equal to 20%	2	2,253
15.01-20%	2,242			
20.02-60%	189	Greater than 20%	3	189

K-means Application in Finance

Convert all the quantitative attributes into qualitative attributes using multi-valued logic

Panel B. Leverage				
No leverage	109	Less than equal to 1X		
Less than equal to 0.8X	77		1	1,223
Greater than 0.8X and less than equal to 1X	1,037			
Greater than 1X and less than equal to 1.25X	345	Greater than 1X and less than equal to 2X	2	1,013
Greater than 1.25X and less than equal to 1.5X	348			
Greater than 1.5X and less than equal to 2X	320			
Greater than 2X and less than equal to 9X	196	Greater than 2X and less than equal to 10X	3	225
Greater than 9X and less than equal to 10X	29			
Greater than 10X and less than equal to 25X	30	Greater than 10X and less than equal to 30X	4	44
Greater than 25X and less than equal to 30X	14			
Greater than 30X and less than equal to 35X	6	Greater than 30X and less than equal to 50X	5	14
Greater than 35X and less than equal to 50X	3			
Greater than 50X and less than equal to 70X	11	Greater than 50X and less than equal to 70X	6	3
Not known	273	Not declared	7	273

K-means Application in Finance

Convert all the quantitative attributes into qualitative attributes using multi-valued logic

Panel C. Redemption Frequency				
Daily Weekly Bimonthly Monthly	61 89 13 1,003	Less than equal to monthly	1	1,166
Quarterly Semiannually	1,084 175	Greater than monthly and less than equal to semiannually	2	1,259
Annually More than Annual	321 51	Greater than semiannually	3	372

K-means Application in Finance

Convert all the quantitative attributes into qualitative attributes using multi-valued logic

Panel D. Minimum Purchase				
<=$100	180	Less than equal to $25,000	1	386
$101-$5,000	17			
$5,001-$25,000	189			
$25,001-$50,000	87	Less than equal to $50,000	2	87
$50,001 - $100,000	383	Less than equal to $100,000	3	383
$100,001-$500,000	1,179	Less than equal to $500,000	4	1,179
$500,001-$1 million	649	Less than equal to $1 million	5	649
$1 million -$5 million	103	Less than equal to $25 million	6	111
$5 million -$25 million	8			
More than $25 million	2	Greater than $25 million	7	2

K-means Application in Finance

Two Distance Measure:

- Squared Euclidean

$$d_{xy}^2 = (x_1 - y_1)^2 + (x_2 - y_2)^2 + \ldots (x_p - y_p)^2$$
$$= (\mathrm{X} - \mathrm{Y})' (\mathrm{X} - \mathrm{Y})$$

- City-block distance measure

$$d_{xy} = (x_1 - y_1) + (x_2 - y_2) + \ldots (x_p - y_p)$$
$$= \sum_{j=1}^{p} (x_j - y_j)$$

K-means Application in Finance

The results

ZCM/Hedge Category	Number of Funds in Cluster						Total Funds
	1	2	3	4	5	6	
Event Driven	8	14	10	8	33	**0**	**73**
Global International	15	10	5	3	**8**	**4**	**45**
Global Regional Established	25	78	42	15	12	1	**173**
Global Regional Emerging	2	10	7	20	24	19	**82**
Global US	6	15	16	20	13	10	**80**
Global Macro	3	13	4	35	13	19	**87**
US Opportunistic	**0**	3	3	**0**	2	**0**	**8**
Long Only/Leveraged	**0**	1	1	1	3	1	7
Market Neutral	8	45	31	30	46	71	**231**
Sector	**0**	29	18	1	5	1	**54**
Short Sellers	**0**	9	2	**0**	1	**0**	**12**
Total Funds	**67**	**227**	**139**	**133**	**160**	**126**	**852**

K-means Application in Finance

ZCM/Hedge Category	Number of Funds in Cluster							Total Funds
	1	2	3	4	5	6	7	
Event Driven	8	14	10	8	33	**0**	**0**	**73**
Global International	15	10	5	3	**1**	**8**	**3**	**45**
Global Regional Established	25	78	42	15	12	1	**0**	**173**
Global Regional Emerging	2	10	7	20	24	5	14	**82**
Global US	6	15	16	20	13	5	5	**80**
Global Macro	3	13	4	35	13	7	12	**87**
US Opportunistic	**0**	3	3	**0**	2	**0**	**0**	**8**
Long Only/Leveraged	**0**	1	1	1	3	1	0	7
Market Neutral	8	45	31	30	46	51	20	**231**
Sector	**0**	29	18	1	5	1	**0**	**54**
Short Sellers	**0**	9	2	0	1	**0**	**0**	**12**
Total Funds	**67**	**227**	**139**	**133**	**153**	**79**	**54**	**852**

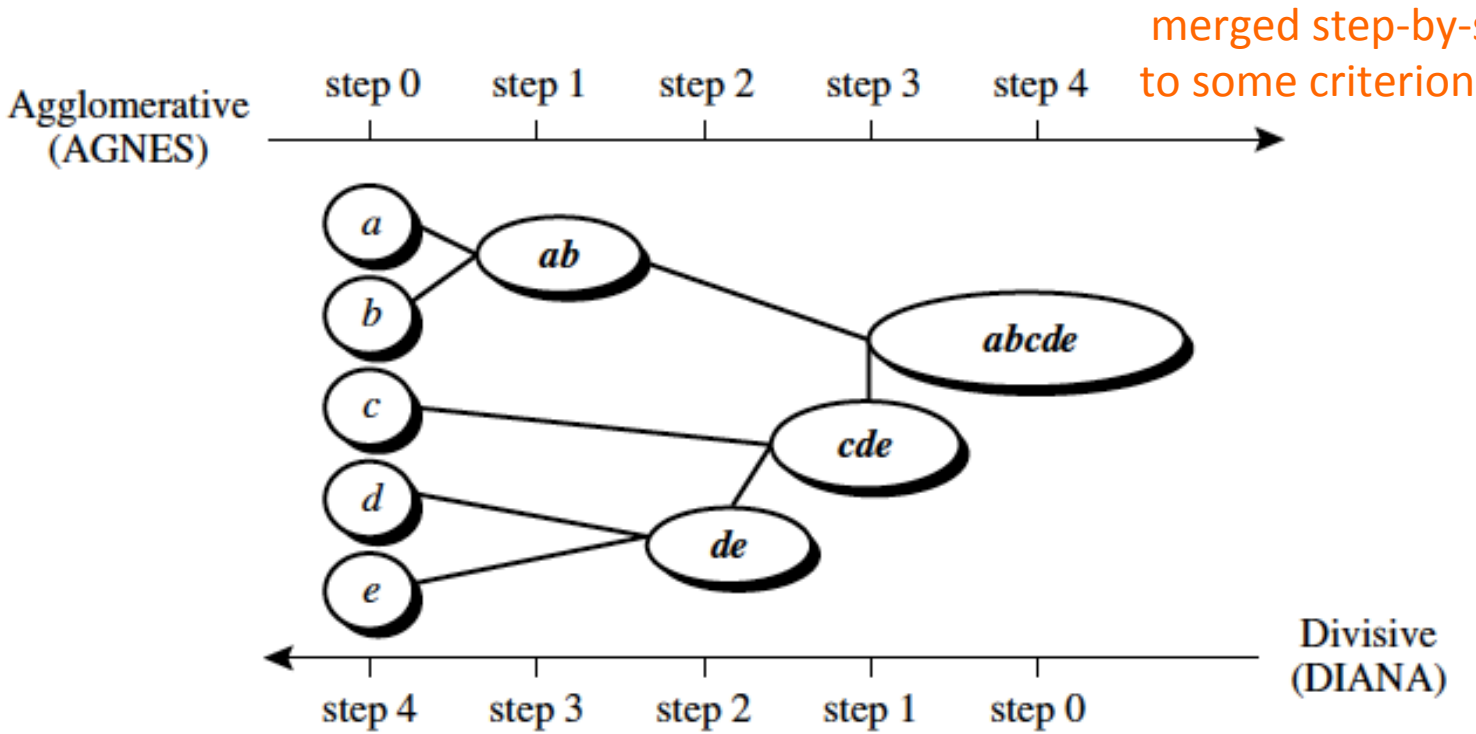
Contents

Hierarchical Methods

- A hierarchical clustering method works by grouping data objects into a tree of clusters
- Inability to perform adjustment once a merge or split decision has been executed. if a particular merge or split decision later turns out to have been a poor choice, the method cannot backtrack and correct it
- The integration of hierarchical agglomeration with iterative relocation methods

Hierarchical Methods

Agglomerative versus divisive hierarchical clustering



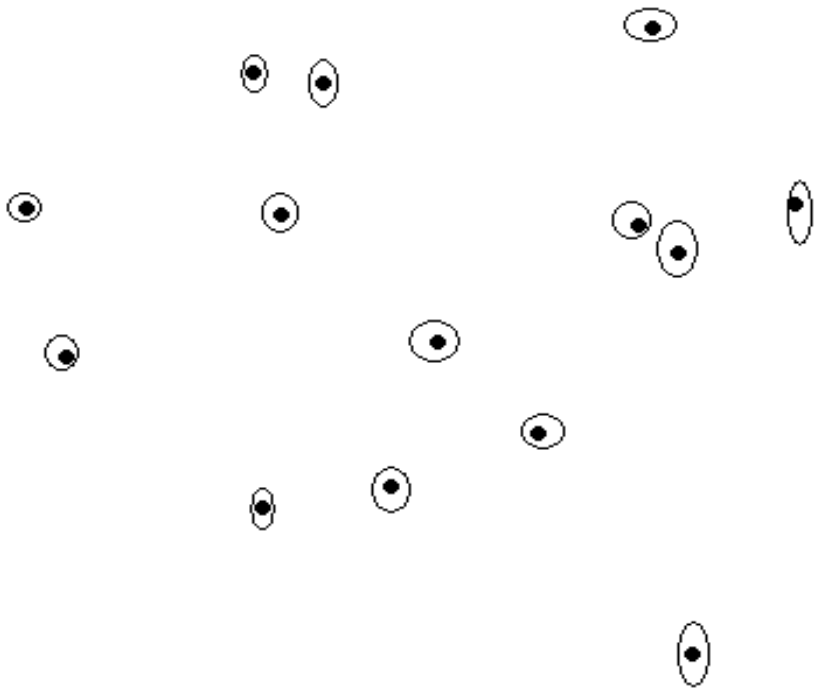
merged step-by-step according to some criterion

repeats until each new cluster contains only a single object

Hierarchical Methods

How objects are grouped together step by step.

Step 1: Every object is a cluster:



Hierarchical Methods

How objects are grouped together step by step.

Step 2: Compute the distance:

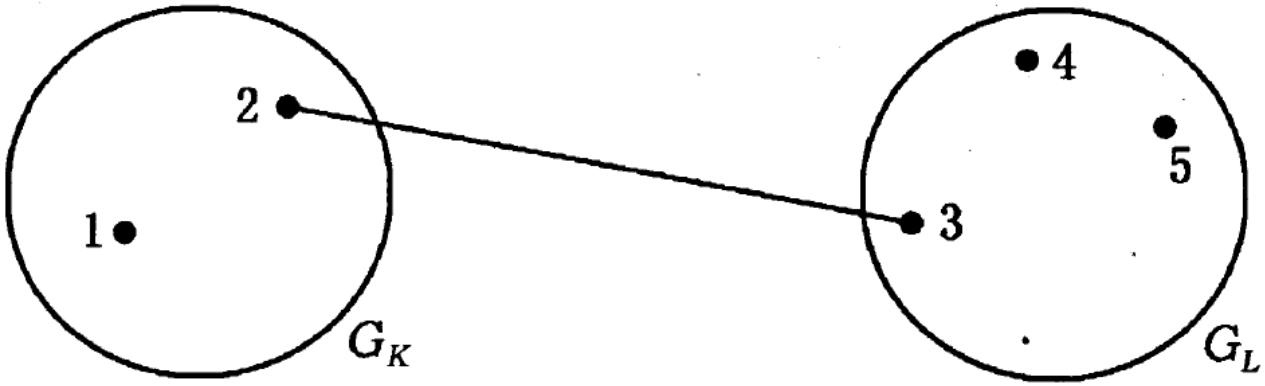
- Minimum Distance (Single-linkage Clustering)
- Maximum Distance (Complete-linkage Clustering)
- Mean Distance (Average-linkage Clustering)
- Centroid-linkage Clustering

Hierarchical Methods

How objects are grouped together step by step.

Minimum Distance (Single-linkage Clustering)

$$D_{KL} = \min_{i \in G_K, j \in G_L} d_{ij}$$



$$D_{KL} = d_{23}$$

Hierarchical Methods

How objects are grouped together step by step.

Suppose that we have 5 samples, $G_1=\{1\}$, $G_2=\{2\}$, $G_3=\{6\}$, $G_4=\{8\}$, $G_5=\{11\}$

(1) Compute the distance matrix $D_{(0)}$, which is Symmetric

	G_1	G_2	G_3	G_4	G_5
G_1	0				
G_2	1	0			
G_3	5	4	0		
G_4	7	6	2	0	
G_5	10	9	5	3	0

Hierarchical Methods

How objects are grouped together step by step.

(2) Find the minimum distance of the $D_{(0)}$, define it like D_{KL}, then G_K and G_L combine as a new cluster G_M, $G_M = G_K \cup G_L$

(3) Compute the distance between the new cluster and the other clusters, and then we have $D_{(1)}$

$$D_{MJ} = \min_{i \in G_M, j \in G_J} d_{ij} = \min\left\{ \min_{i \in G_K, j \in G_J} d_{ij}, \min_{i \in G_L, j \in G_J} d_{ij} \right\}$$

$$= \min\left\{ D_{KJ}, D_{LJ} \right\}$$

Hierarchical Methods

How objects are grouped together step by step.

	G_1	G_2	G_3	G_4	G_5
G_1	0				
G_2	1	0			
G_3	5	4	0		
G_4	7	6	2	0	
G_5	10	9	5	3	0

$G_6 = G_1 \cup G_2$

$D_{(0)} \rightarrow D_{(1)}$

	G_6	G_3	G_4	G_5
G_6	0			
G_3	4	0		
G_4	6	2	0	
G_5	9	5	3	0

Hierarchical Methods

How objects are grouped together step by step.

(4) Repeat until it merge to one cluster

$G_7 = G_3 \cup G_4$ $D_{(2)}$

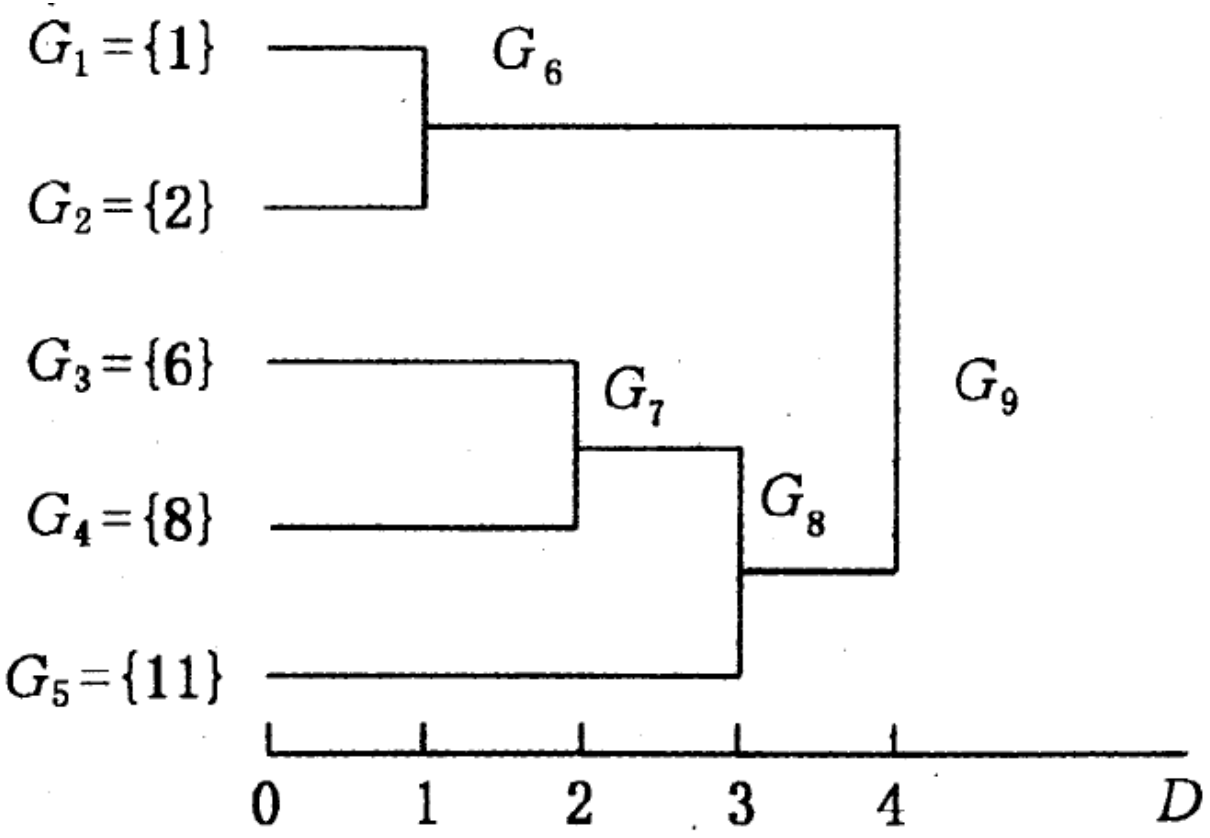
	G_6	G_7	G_5
G_6	0		
G_7	4	0	
G_5	9	3	0

$G_6 = G_1 \cup G_2$ $D_{(3)}$

	G_6	G_8
G_6	0	
G_8	4	0

Hierarchical Methods

How objects are grouped together step by step.



Hierarchical Methods

How objects are grouped together step by step.

Step 2: Compute the distance:

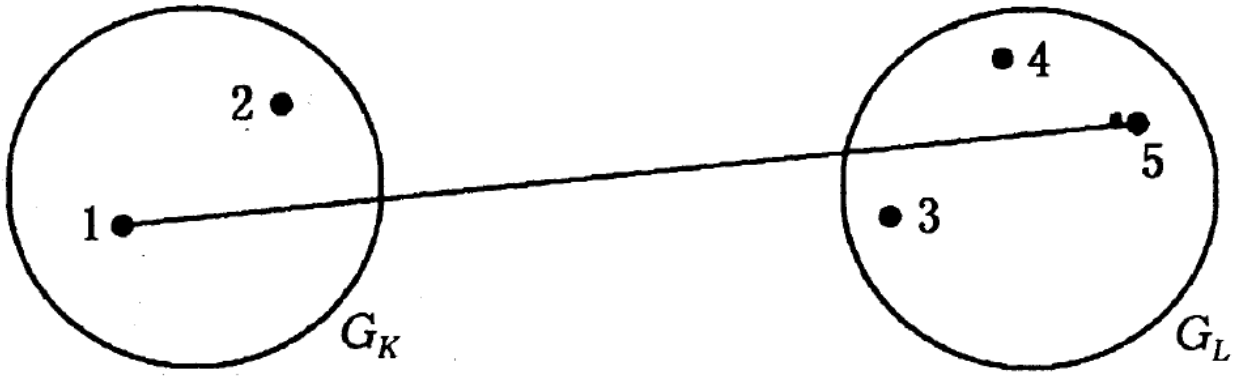
- Minimum Distance (Single-linkage Clustering)
- Maximum Distance (Complete-linkage Clustering)
- Mean Distance (Average-linkage Clustering)
- Centroid-linkage Clustering

Hierarchical Methods

How objects are grouped together step by step.

Maximum Distance (Complete-linkage Clustering)

$$D_{KL} = \max_{i \in G_K, j \in G_L} d_{ij}$$

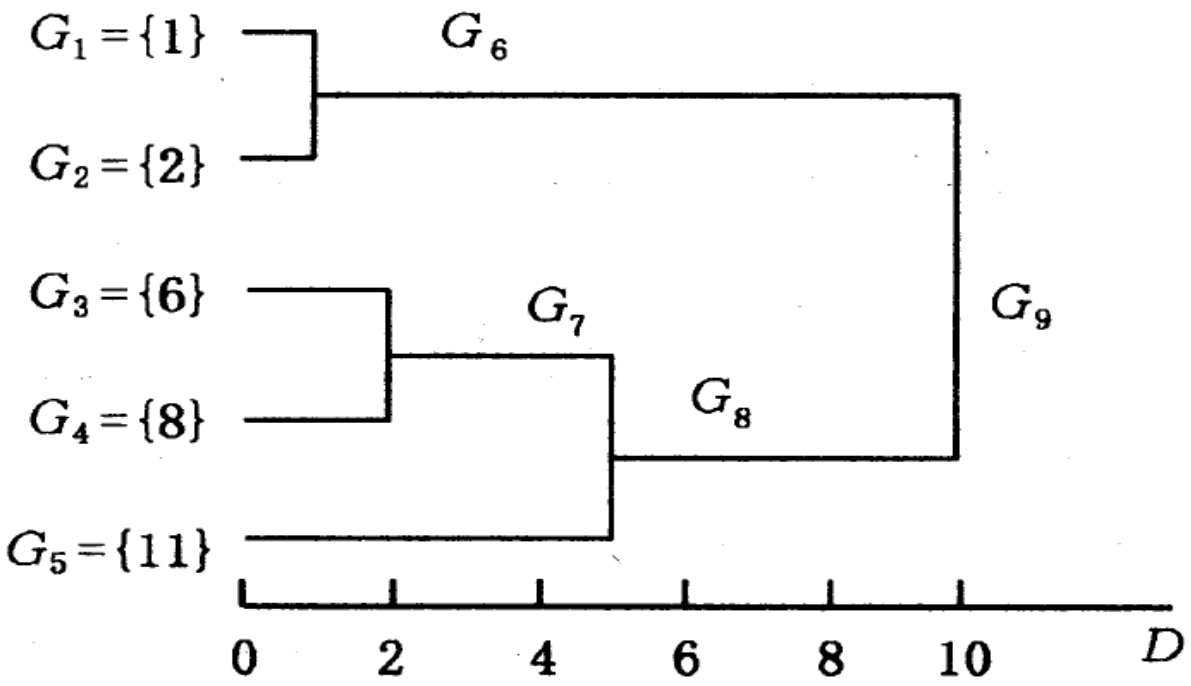


$$D_{KL} = d_{15}$$

Hierarchical Methods

How objects are grouped together step by step.

It is the same method with Single-linkage Clustering, expect that $D_{MJ} = \max\{D_{KJ}, D_{LJ}\}$



Hierarchical Methods

How objects are grouped together step by step.

It might suffer from the outliers. One efficient solution is that pick up all the outliers at first, and then do the clustering.

Hierarchical Methods

How objects are grouped together step by step.

Step 2: Compute the distance:

- Minimum Distance (Single-linkage Clustering)
- Maximum Distance (Complete-linkage Clustering)
- Mean Distance (Average-linkage Clustering)
- Centroid-linkage Clustering

Hierarchical Methods

How objects are grouped together step by step.

Mean Distance (Average-linkage Clustering)

(1) Define the mean distance as

$$D_{KL} = \frac{1}{n_K n_L} \sum_{i \in G_K, j \in G_L} d_{ij}$$

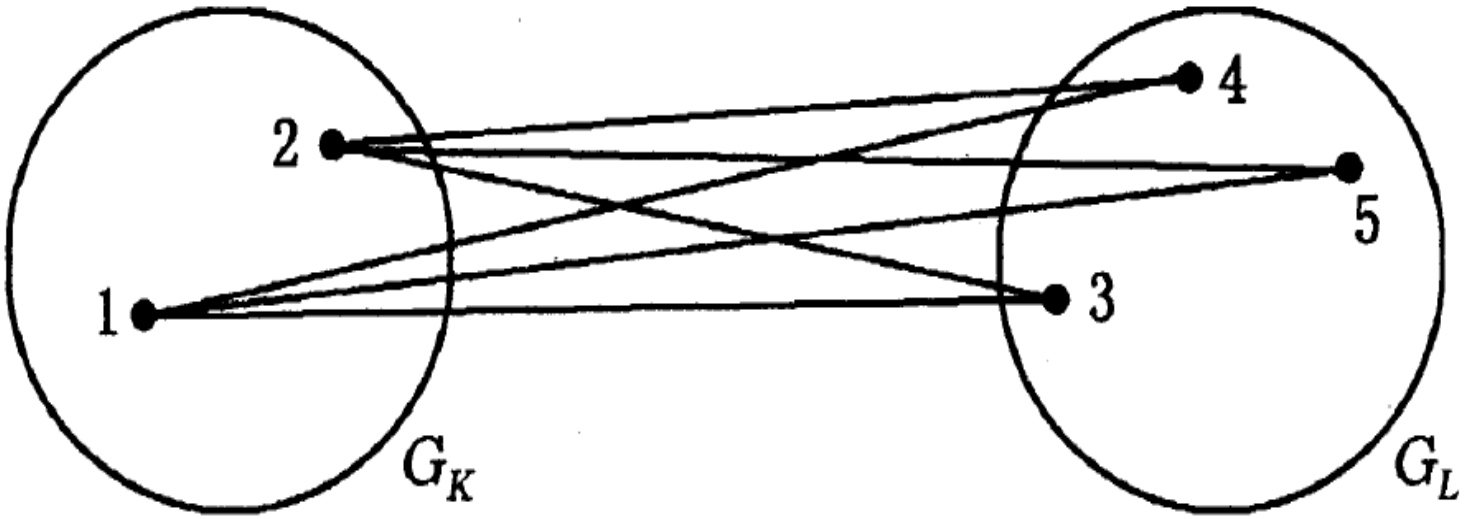
And then recurrence formula:

$$D_{MJ} = \frac{1}{n_M n_J} \sum_{i \in G_M, j \in G_J} d_{ij} = \frac{1}{n_M n_J} \left(\sum_{i \in G_K, j \in G_J} d_{ij} + \sum_{i \in G_L, j \in G_J} d_{ij} \right)$$

$$= \frac{n_K}{n_M} D_{KJ} + \frac{n_L}{n_M} D_{LJ}$$

Hierarchical Methods

How objects are grouped together step by step.



$$D_{KL}=(d_{13}+d_{14}+d_{15}+d_{23}+d_{24}+d_{25})/\ 6$$

Hierarchical Methods

How objects are grouped together step by step.

Mean Distance (Average-linkage Clustering)

(2) Define the mean distance as

$$D_{KL}^2 = \frac{1}{n_K n_L} \sum_{i \in G_K, j \in G_L} d_{ij}^2$$

And then recurrence formula:

$$D_{MJ}^2 = \frac{n_K}{n_M} D_{KJ}^2 + \frac{n_L}{n_M} D_{LJ}^2$$

Hierarchical Methods

How objects are grouped together step by step.

$D^2_{(0)}$

	G_1	G_2	G_3	G_4	G_5
G_1	0				
G_2	1	0			
G_3	25	16	0		
G_4	49	36	4	0	
G_5	100	81	25	9	0

Hierarchical Methods

How objects are grouped together step by step.

$$\boldsymbol{D}^2_{(1)}$$

	G_6	G_3	G_4	G_5
G_6	0			
G_3	20.5	0		
G_4	42.5	4	0	
G_5	90.5	25	9	0

Hierarchical Methods

How objects are grouped together step by step.

$\boldsymbol{D}^2_{(2)}$

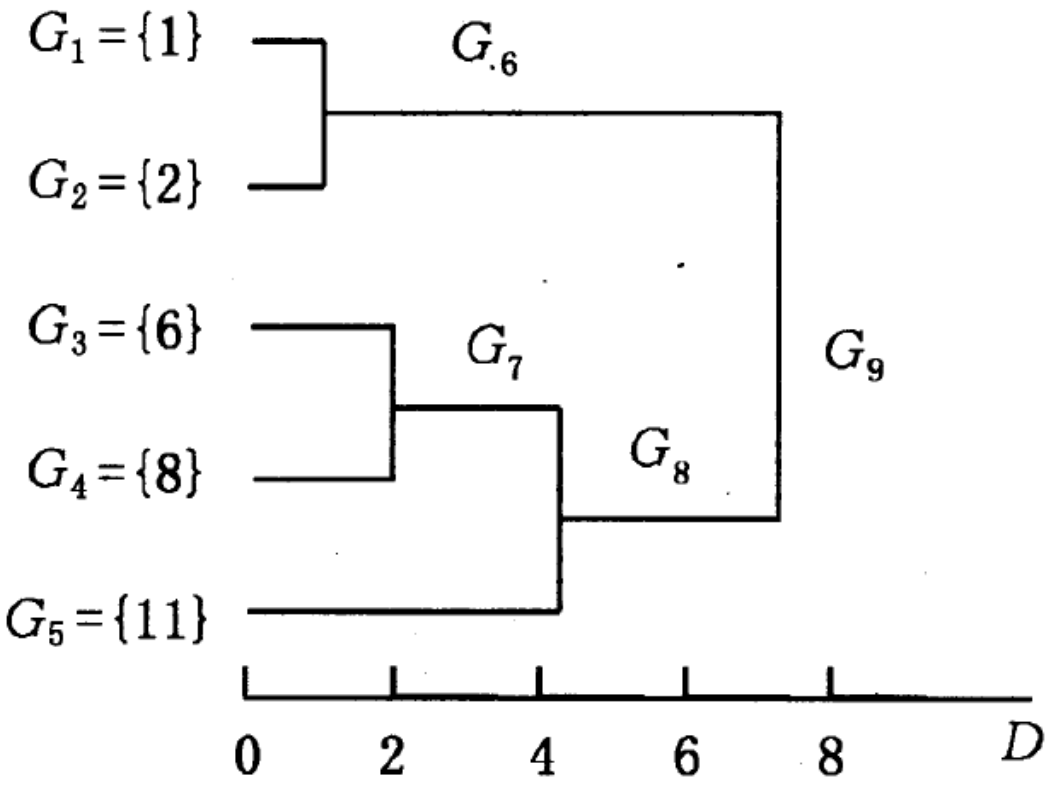
	G_6	G_7	G_5
G_6	0		
G_7	31.5	0	
G_5	90.5	17	0

$\boldsymbol{D}^2_{(3)}$

	G_6	G_8
G_6	0	
G_8	51.17	0

Hierarchical Methods

How objects are grouped together step by step.



Hierarchical Methods

How objects are grouped together step by step.

Step 2: Compute the distance:

- Minimum Distance (Single-linkage Clustering)
- Maximum Distance (Complete-linkage Clustering)
- Mean Distance (Average-linkage Clustering)
- Centroid-linkage Clustering

Hierarchical Methods

How objects are grouped together step by step.

Centroid-linkage Clustering

Suppose that the mean value of G_k and G_L are $\bar{\boldsymbol{x}}_K and \bar{\boldsymbol{x}}_L$

Define the distance between the two objects as

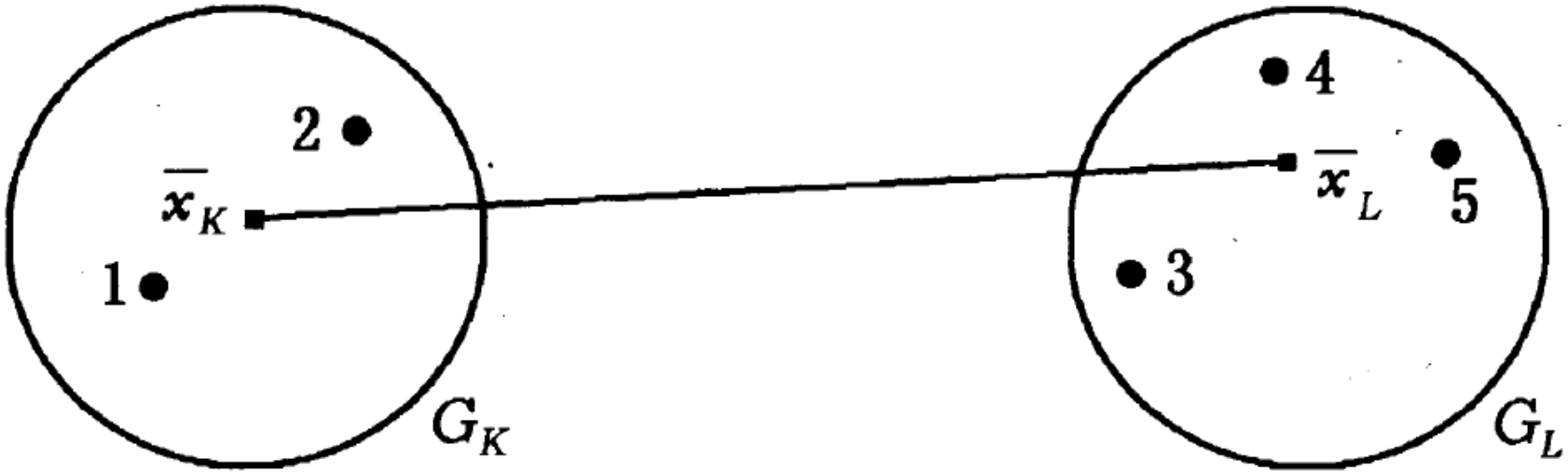
$$D_{KL}^2 = d_{\bar{\boldsymbol{x}}_K \bar{\boldsymbol{x}}_L}^2 = \left(\bar{\boldsymbol{x}}_K - \bar{\boldsymbol{x}}_L\right)' \left(\bar{\boldsymbol{x}}_K - \bar{\boldsymbol{x}}_L\right)$$

Then the recurrence formula:

$$D_{MJ}^2 = \frac{n_K}{n_M} D_{KJ}^2 + \frac{n_L}{n_M} D_{LJ}^2 - \frac{n_K n_L}{n_M^2} D_{KL}^2$$

Hierarchical Methods

How objects are grouped together step by step.



$$D_{KL} = d_{\bar{x}_K \bar{x}_L}$$

Contents

Cluster Analysis in Biology

Method Paper

Eisen MB, Spellman PT, Brown PO, and Botstein D (1998) **Cluster analysis and display of genome-wide expression patterns**. *Proceedings of the National Academy of Sciences,* Vol. 95: 14863-14868.

Why did I choose these paper:

- One of first papers to statistically analyze a genomic dataset (regarded as a classic)
- 3rd most cited Biology paper of all-time (13,975 citations)
- Direct application of hierarchical clustering to biological data
- My background is in genomics/bioinformatics

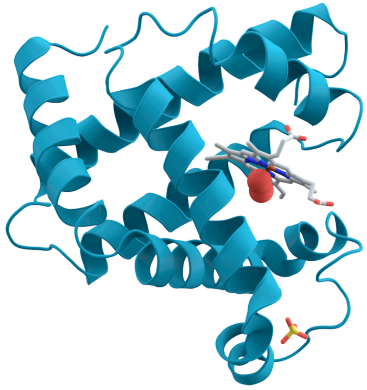
DNA, Genes and Proteins

DNA: molecule that encodes genetic instructions

AGCGATGCCCAATGTAACAA

Gene: a segment of DNA that codes for a protein

Simplifying process dramatically since there are several immediate steps

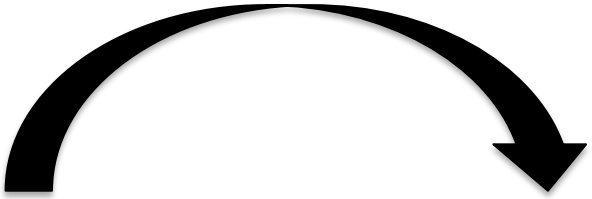


Protein: molecules that perform vast array of biological functions

Gene Expression

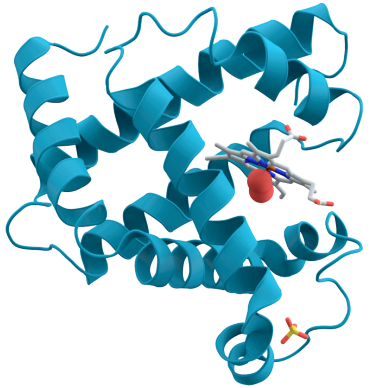
Gene Expression: the process by which information from a gene is used in the synthesis of a functional gene product, usually protein.

Measure of gene activity



AGCGATGCCCAATGTAACAA

Gene A



Protein A

Key goal of biology

Number of genes in model organisms

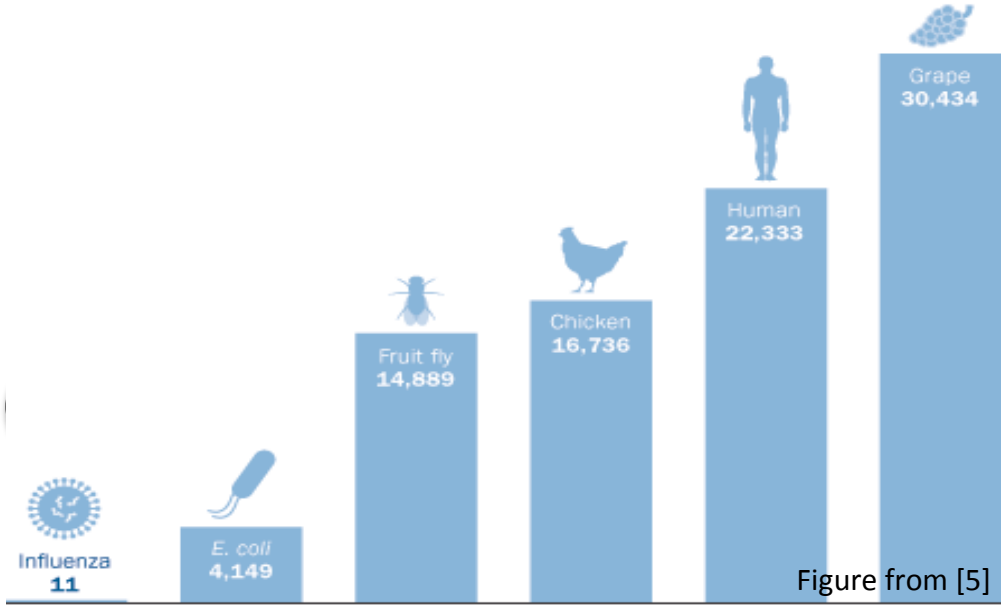


Figure from [5]

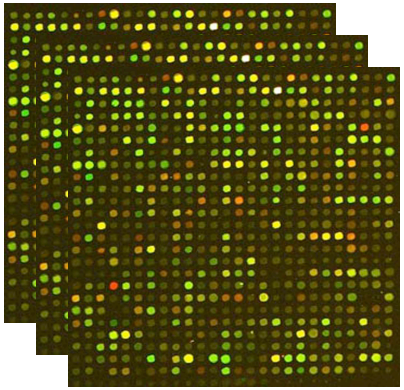
- While we can identify almost all genes, still don't know the function of a large percentage even in fruit fly

- New genomic technologies that measure gene expression and algorithms are helping to achieve this goal

Goal of study

Overall goal: Develop a hierarchical clustering method that can be applied to genomic expression data to be able to group functionally similar genes and assign functions to genes for which little information is available.

Genomic dataset



DNA microarray

	Expression levels	
Gene	**Condition1**	**Condition2**
geneA	1.2	2.2
geneB	-1.3	-0.5
……		
geneZ	-0.4	-0.7

Cluster

Group 1

Gene	Function
GeneA	Stress resp.
GeneC	unknown
…….	……

Assign function to uncharacterized genes

Group 2

Gene	Function
GeneF	Growth
GeneD	unknown
…….	……

Experimental Data and Cluster Analysis

- Measured response of genes in mammalian fibroblast cells to the addition of growth serum
- Measured gene expression in a time series (13 time-points) over the course of 24 hr
- Used paired-wise average-linkage hierarchical clustering with the Pearson Correlation Coefficient as distance metric

Hierarchical Clustering: Distance Metric

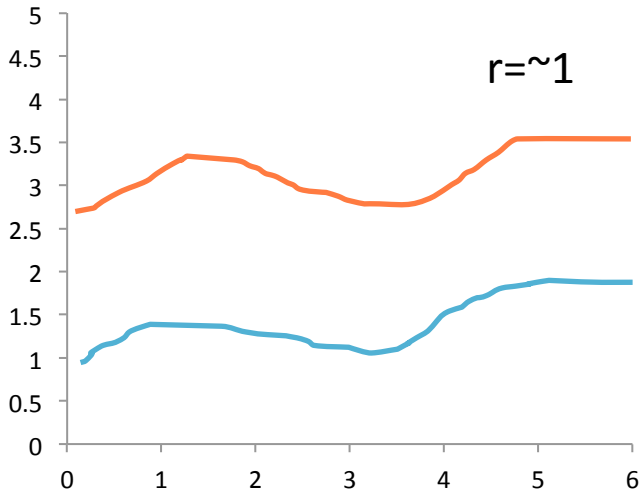
Pearson Correlation Coefficient

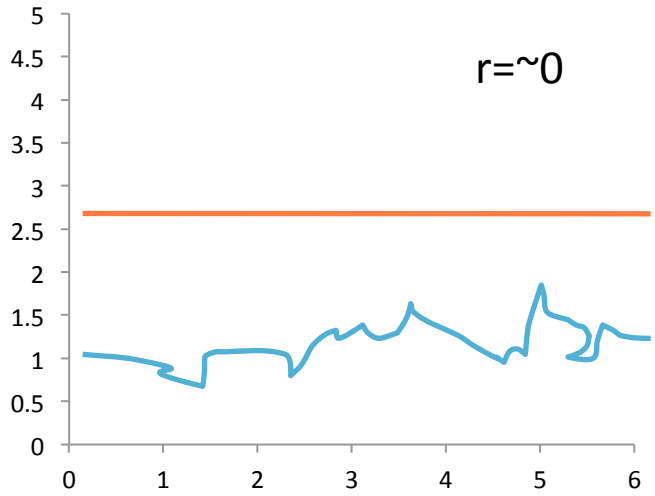
$$S(X,Y) = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{(X_i - \overline{X})}{\sigma_X}\right)\left(\frac{(Y_i - Y)}{\sigma_Y}\right)$$

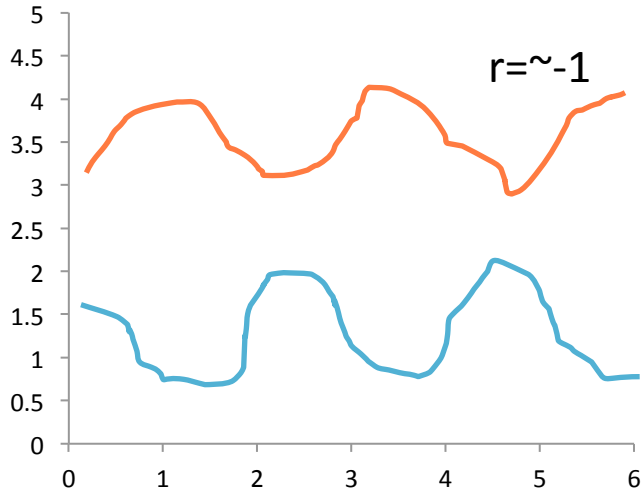
where

X_i = primary data for gene X in condition i
Y_i = primary data for gene Y in condition i
σ = standard deviation of X or Y

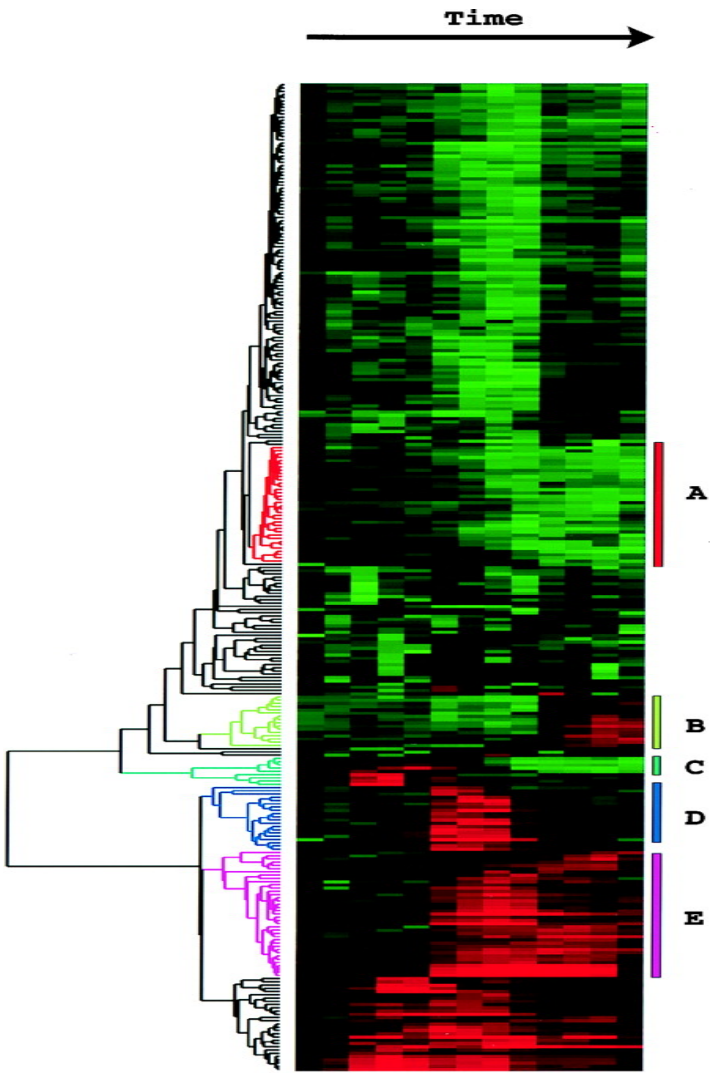
Pearson Correlation Coefficient







Results



Genes of known similar function tended to cluster together

Co-expression of genes with known function with poorly characterized genes helped assign functions to these less understood genes

Functional Groups

(A) Cholesterol biosynthesis,
(B) the cell cycle,
(C) the immediate–early response,
(D) signaling and angiogenesis,
(E) wound healing and tissue remodeling

Conclusions

- Hierarchical clustering effectively grouped genes into several functional groups
- Enabled functional characterization of genes for which little information was available
- Viewing the genes as functional groups helps impart biological significance to the broad patterns seen in expression data
- Allows for greater dissection of the timing of cellular processes and functional groups in response to growth stimulation
- Lastly, describes a method that can be applied to other similar studies to survey large-scale features of complex datasets and then focus in on the interesting details

The End

Any Questions?
Thank you!