Clustering

Clustering Outline

- ***Goal:*** Provide an overview of the clustering problem and introduce some of the basic algorithms
- Clustering Problem Overview
- Clustering Techniques
  - Hierarchical Algorithms
  - Partitional Algorithms
  - Genetic Algorithm
  - Clustering Large Databases

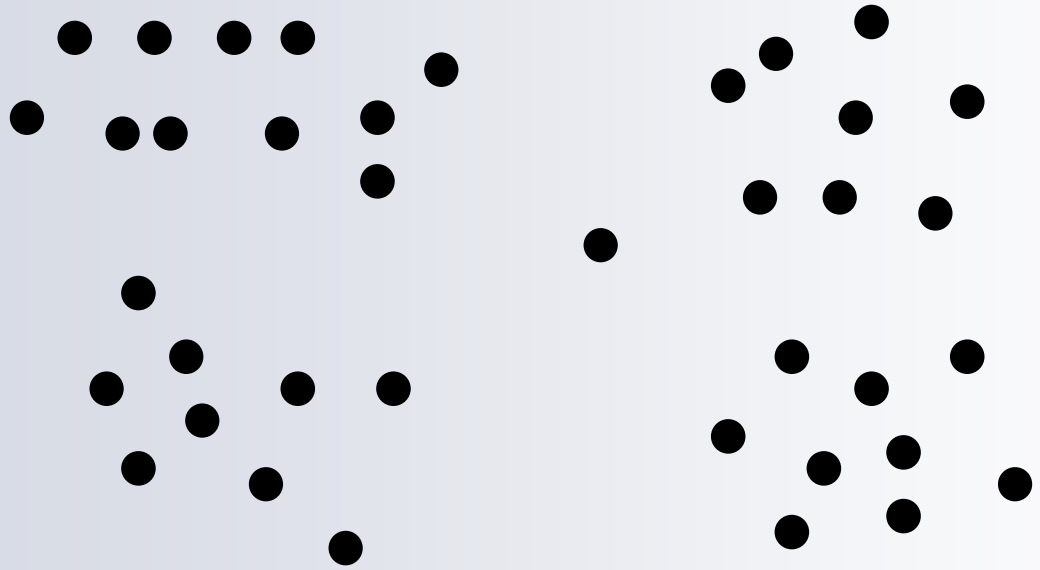
Clustering Examples

- ***Segment*** customer database based on similar buying patterns.
- Group houses in a town into neighborhoods based on similar features.
- Identify new plant species
- Identify similar Web usage patterns

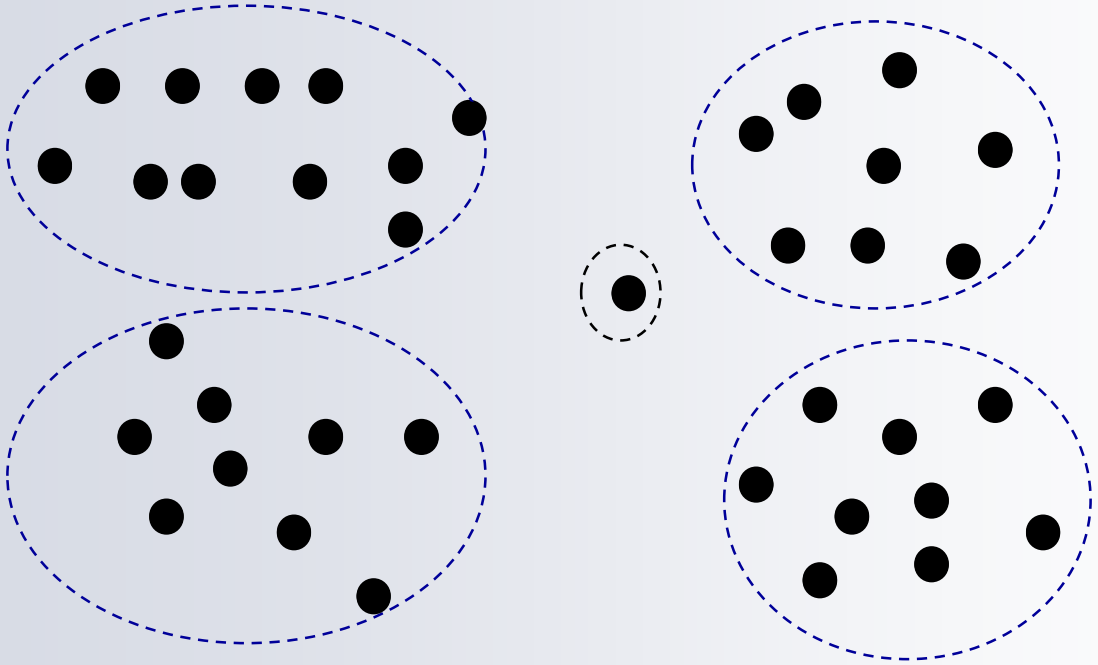
Clustering Example

Income	Age	Children	Marital Status	Education
$25,000	35	3	Single	High School
$15,000	25	1	Married	High School
$20,000	40	0	Single	High School
$30,000	20	0	Divorced	High School
$20,000	25	3	Divorced	College
$70,000	60	0	Married	College
$90,000	30	0	Married	Graduate School
$200,000	45	5	Married	Graduate School
$100,000	50	2	Divorced	College

Clustering Houses

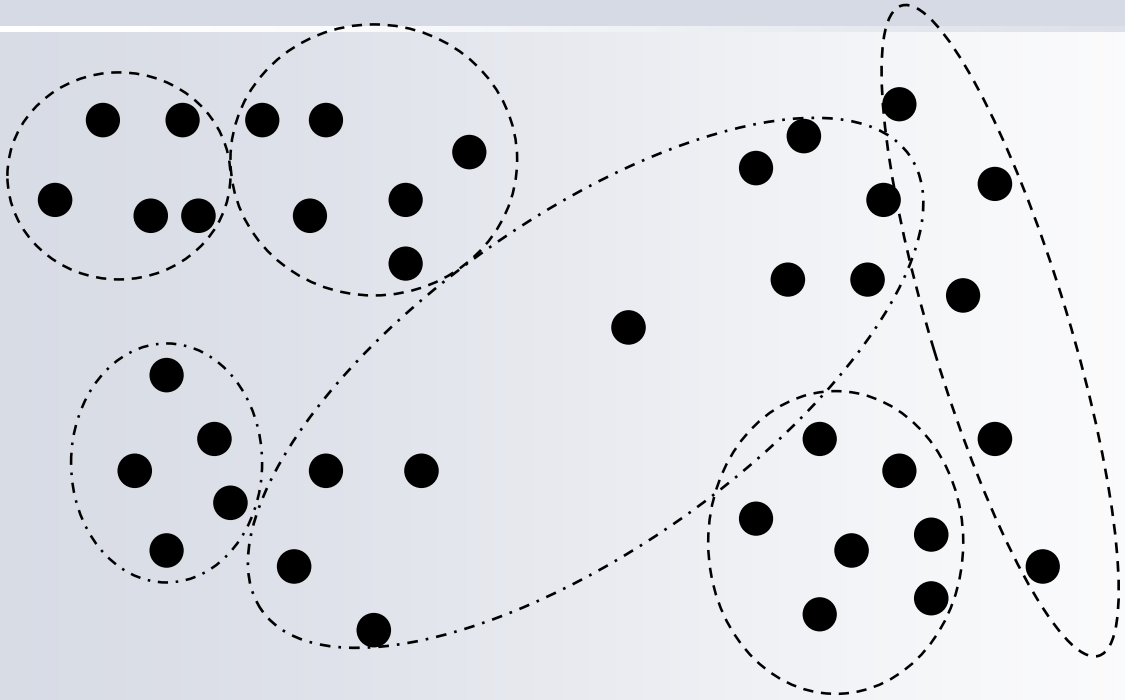


Clustering Houses



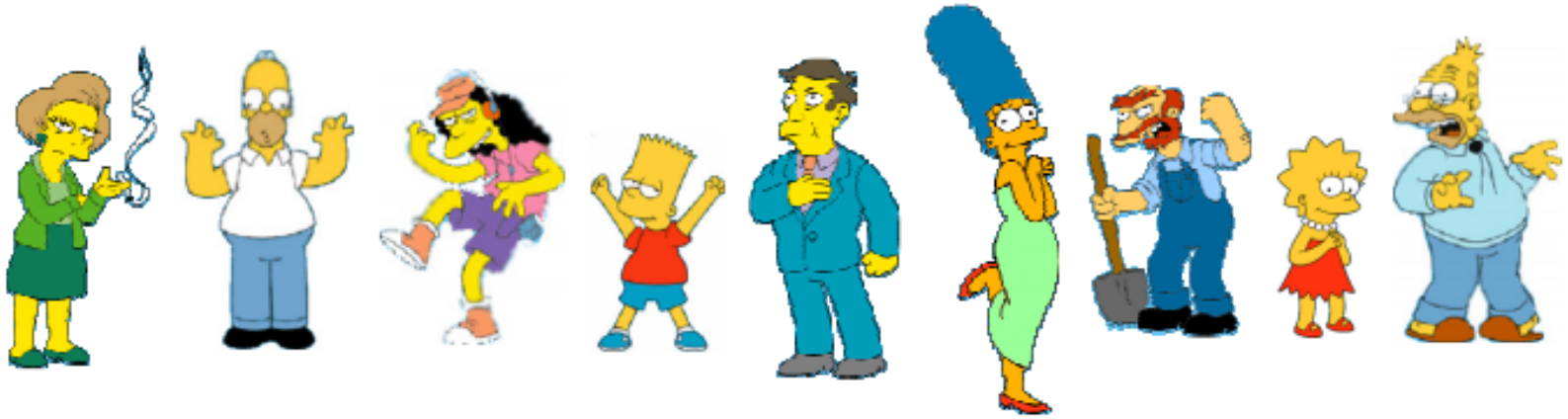
Geographic Distance Based

Clustering Houses



Size Based

What is a natural grouping among these objects?



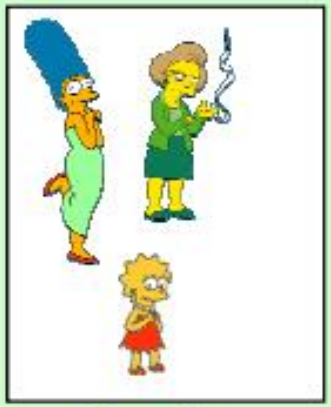
http://www.cs.cmu.edu/afs/andrew/course/15/381-f08/www/lectures/clustering.pdf

What is a natural grouping among these objects?

Clustering is subjective

Simpson's Family

School Employees



Females

Males

http://www.cs.cmu.edu/afs/andrew/course/15/381-f08/www/lectures/clustering.pdf

What is Similarity?

The quality or state of being similar; likeness; resemblance; as, a similarity of features.
Webster's Dictionary

Similarity is hard to define, but…
"*We know it when we see it*"

The real meaning of similarity is a philosophical question. We will take a more pragmatic approach.

http://www.cs.cmu.edu/afs/andrew/course/15/381-f08/www/lectures/clustering.pdf

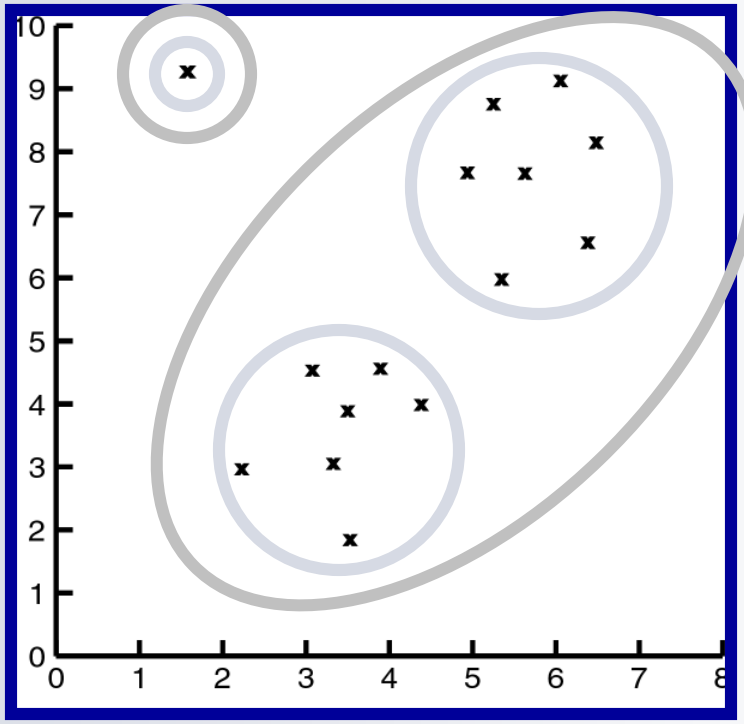
Clustering vs. Classification

- No prior knowledge
  - Number of clusters
  - Meaning of clusters
- Unsupervised learning

Clustering Issues

- Outlier handling
- Dynamic data
- Interpreting results
- Evaluating results
- Number of clusters
- Data to be used
- Scalability

Impact of Outliers on Clustering



Clustering Problem

- Given a database $D=\{t_1,t_2,\ldots,t_n\}$ of tuples and an integer value k, the ***Clustering Problem*** is to define a mapping $f:D\rightarrow\{1,\ldots,k\}$ where each t_i is assigned to one cluster K_j, $1<=j<=k$.
- A ***Cluster***, K_j, contains precisely those tuples mapped to it.
- Unlike classification problem, clusters are not known a priori.

Cluster Parameters

$$centroid = C_m = \frac{\sum_{i=1}^{N}(t_{mi})}{N}$$

$$radius = R_m = \sqrt{\frac{\sum_{i=1}^{N}(t_{mi} - C_m)^2}{N}}$$

$$diameter = D_m = \sqrt{\frac{\sum_{i=1}^{N}\sum_{j=1}^{N}(t_{mi} - t_{mj})^2}{(N)(N-1)}}$$

Distance

- **Euclidean:** $$d(x,y)=\sqrt{\sum_{k=1}^{n}(x_k-y_k)^2}$$

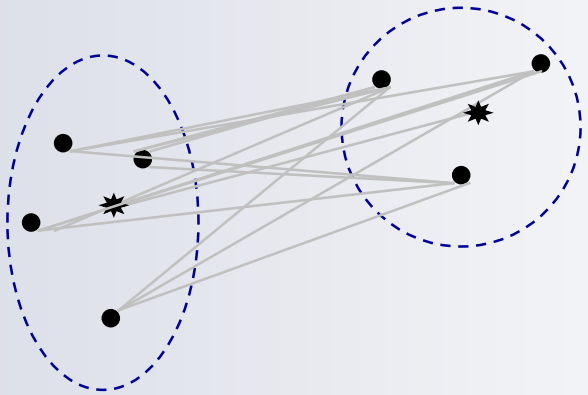
X(1,2)
Y(3,5)

- **Manhattan** $$d(x,y)=\sum_{k=1}^{n}(x_k-y_k)$$

- **Minkowski:** $$d(x,y)=(\sum_{k=1}^{n}(x_k-y_k)^p)^{(1/p)}$$

Linkages Between Clusters

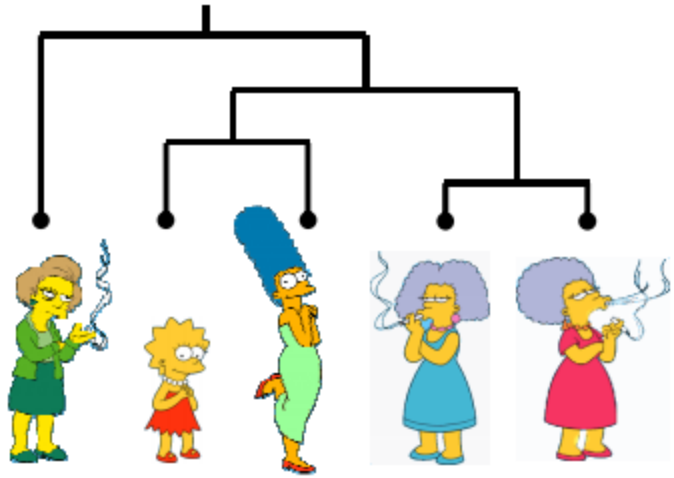
- ***Single Link***: smallest distance between points
- ***Complete Link:*** largest distance between points
- ***Average Link:*** average distance between points
- ***Centroid:*** distance between centroids



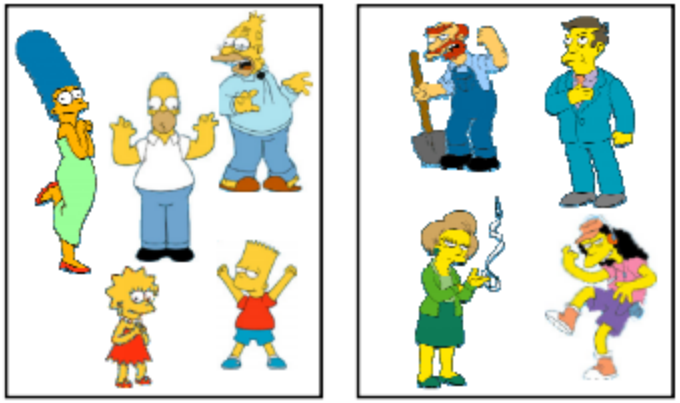
Types of Clustering

- ***Hierarchical*** – clusters recursively constructed and evaluated, top-down or bottom-up.
- ***Partitional*** – instances moved one cluster to another, starting from an initial cluster.

Hierarchical

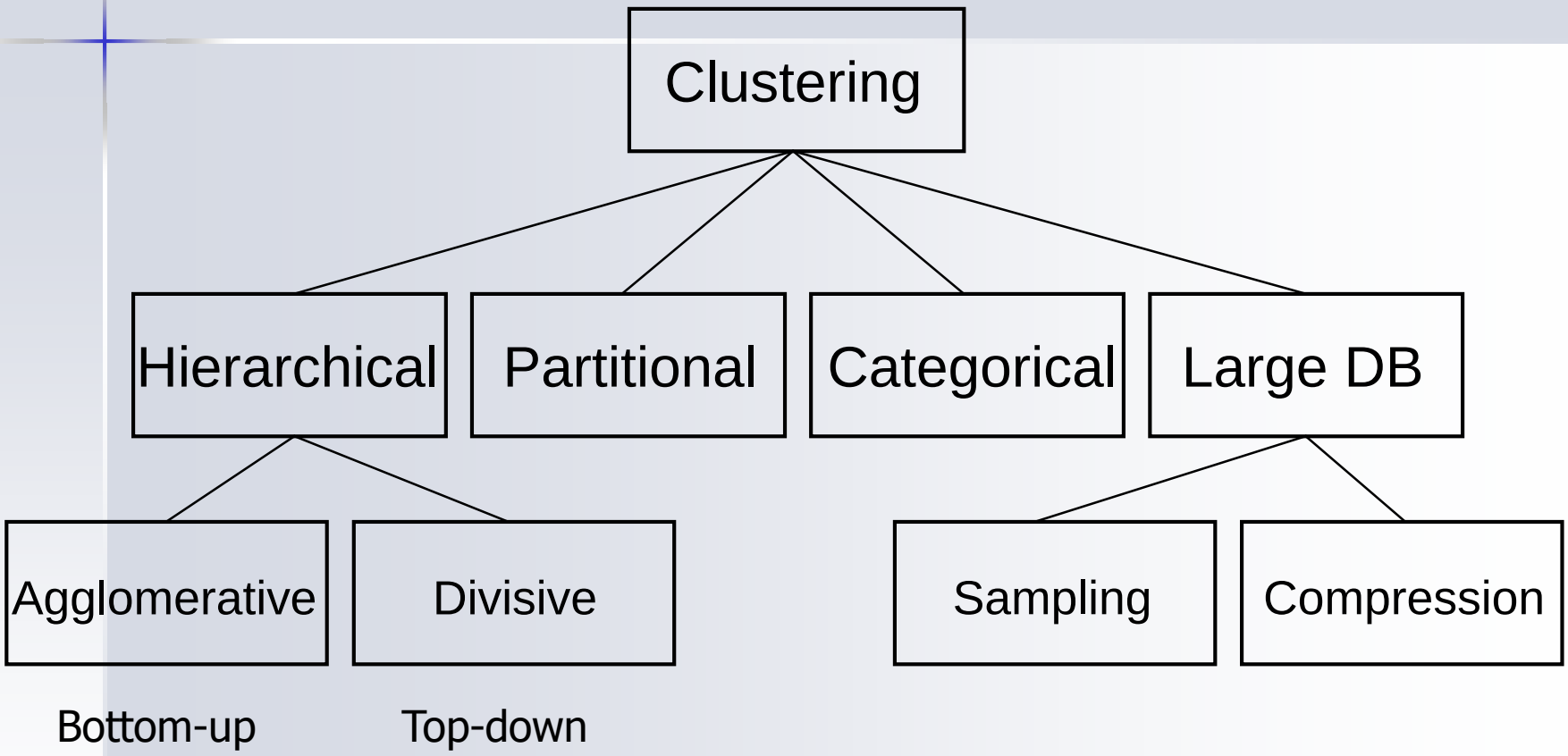


Partitional



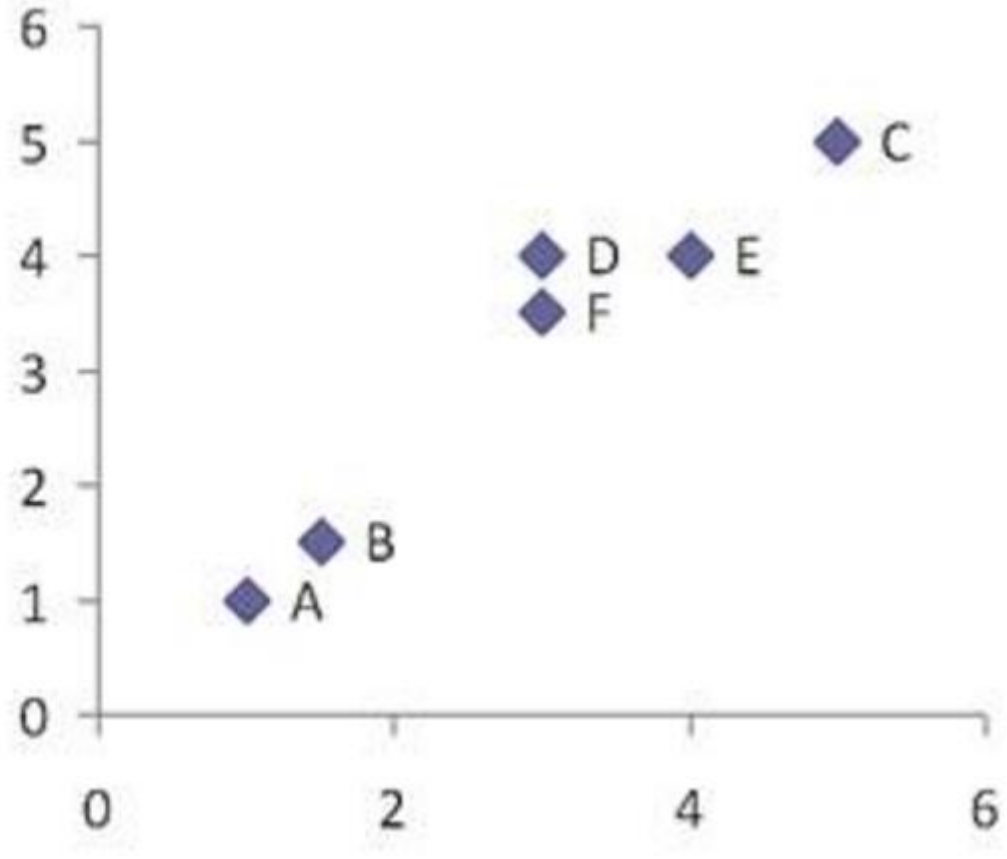
http://www.cs.cmu.edu/afs/andrew/course/15/381-f08/www/lectures/clustering.pdf

Clustering Approaches



Example: Agglomerative

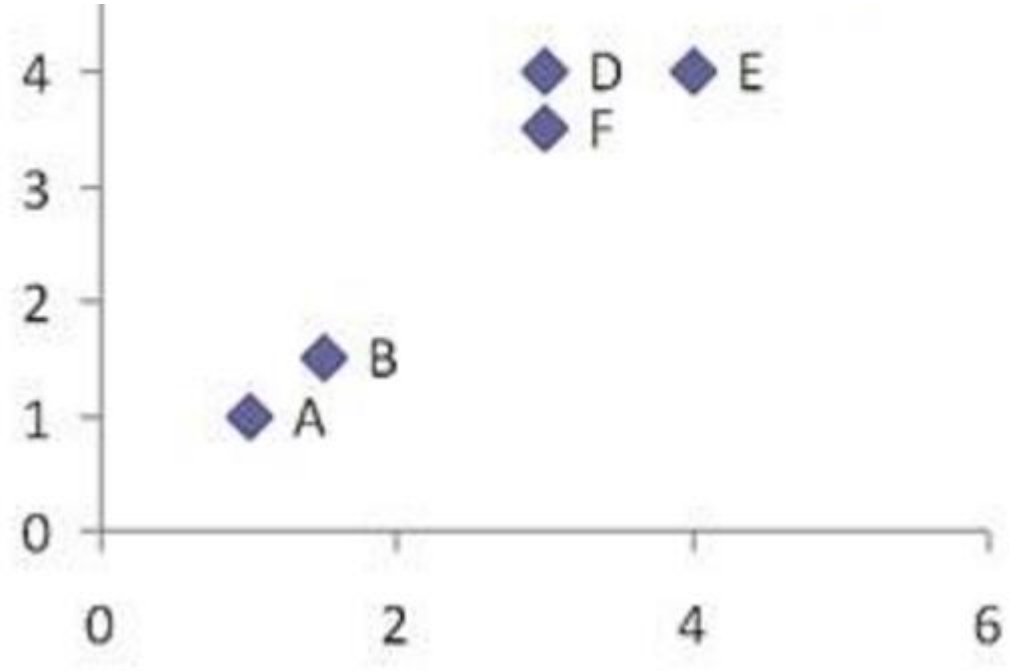
	X1	X2
A	1	1
B	1.5	1.5
C	5	5
D	3	4
E	4	4
F	3	3.5



http://people.revoledu.com/kardi/tutorial/Clustering/Distance%20Matrix.htm

Dist	A	B	C	D	E	F
A	0.00	0.71	5.66	3.61	4.24	3.20
B	0.71	0.00	4.95	2.92	3.54	2.50
C	5.66	4.95	0.00	2.24	1.41	2.50
D	3.61	2.92	2.24	0.00	1.00	0.50
E	4.24	3.54	1.41	1.00	0.00	1.12
F	3.20	2.50	2.50	0.50	1.12	0.00

	X1	X2
A	1	1
B	1.5	1.5
C	5	5
D	3	4
E	4	4
F	3	3.5



http://people.revoledu.com/kardi/tutorial/Clustering/Distance%20Matrix.htm

Dist	A	B	C	D	E	F
A	0.00	0.71	5.66	3.61	4.24	3.20
B	0.71	0.00	4.95	2.92	3.54	2.50
C	5.66	4.95	0.00	2.24	1.41	2.50
D	3.61	2.92	2.24	0.00	1.00	0.50
E	4.24	3.54	1.41	1.00	0.00	1.12
F	3.20	2.50	2.50	0.50	1.12	0.00

Dist	A	B	C	D, F	E
A	0.00	0.71	5.66	?	4.24
B	0.71	0.00	4.95	?	3.54
C	5.66	4.95	0.00	?	1.41
D, F	?	?	?	0.00	?
E	4.24	3.54	1.41	?	0.00

Dist	A	B	C	D, F	E
A	0.00	0.71	5.66	3.20	4.24
B	0.71	0.00	4.95	2.50	3.54
C	5.66	4.95	0.00	2.24	1.41
D, F	**3.20**	**2.50**	**2.24**	0.00	1.00
E	4.24	3.54	1.41	**1.00**	0.00

Dist	A,B	C	(D, F)	E
A,B	0	?	?	?
C	?	0	2.24	1.41
(D, F)	?	2.24	0	1.00
E	?	1.41	1.00	0

Dist	A,B	C	(D, F)	E
A,B	0	4.95	2.50	3.54
C	**4.95**	0	2.24	1.41
(D, F)	**2.50**	2.24	0	1.00
E	**3.54**	1.41	1.00	0

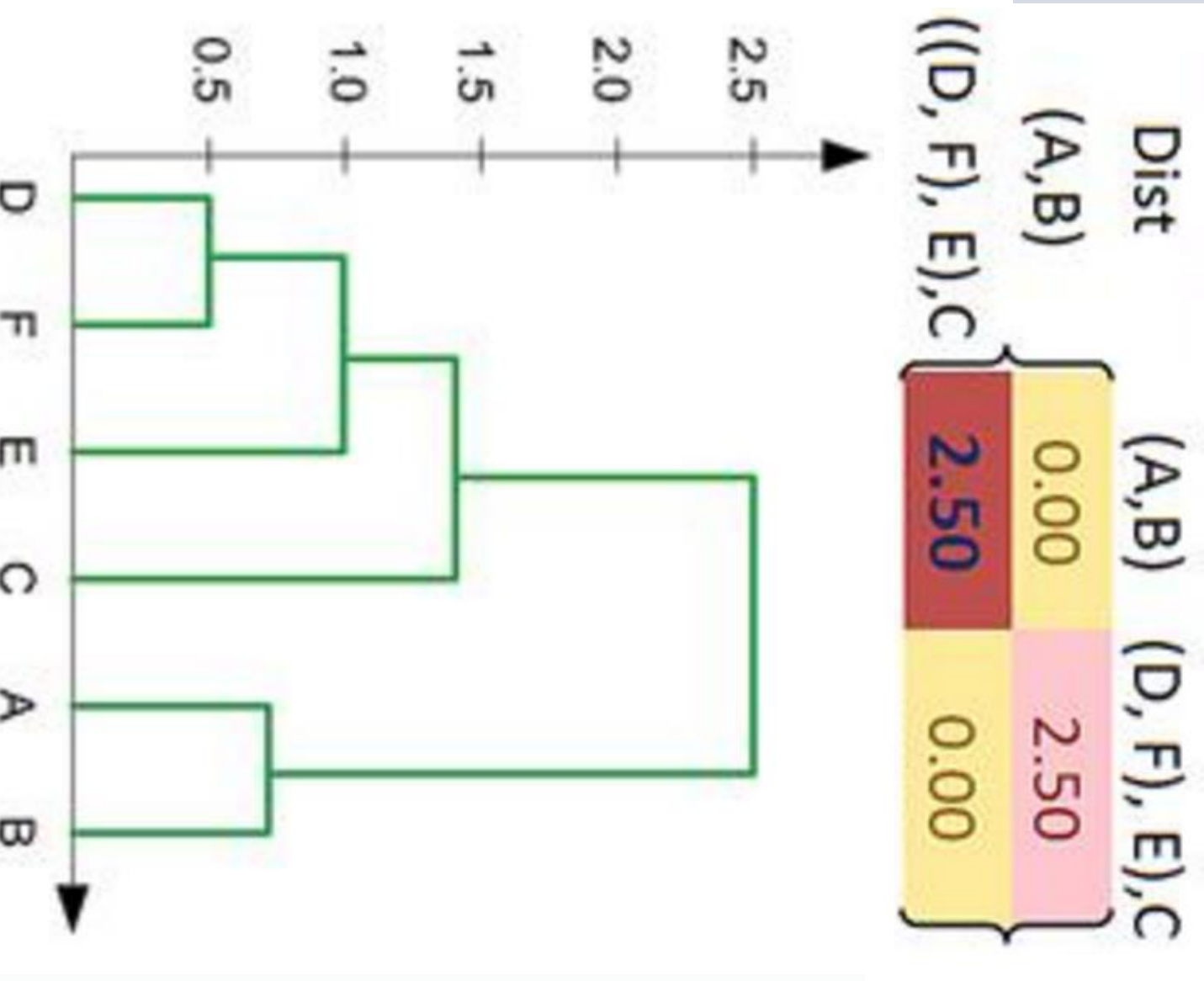
Dist	(A,B)	C	(D, F), E
(A,B)	0.00	4.95	2.50
C	4.95	0.00	1.41
(D, F), E	**2.50**	**1.41**	0.00

den·dro·gram

/ˈdendrəˌgram/

noun

a tree diagram, especially one showing taxonomic relationships.



K-Means

- Initial set of clusters randomly chosen.
- Iteratively, items are moved among sets of clusters until the desired set is reached.
- High degree of similarity among elements in a cluster is obtained.
- Given a cluster $K_i=\{t_{i1},t_{i2},\ldots,t_{im}\}$, the ***cluster mean*** is $m_i = (1/m)(t_{i1} + \ldots + t_{im})$

K-Means

1. Select the number of cluster, K
2. Select the initial cluster center/centroid
3. Calculate the distance between instances and cluster centroid
4. Calculate the new centroid centre
5. If the new cluster centroid is the same
   stop iteration
   else
   repeat steps 3-5.

K-Means Example

- Given: {2,4,10,12,3,20,30,11,25}, k=2
- Randomly assign means: $m_1=3, m_2=4$
- $K_1=\{2,3\}$, $K_2=\{4,10,12,20,30,11,25\}$, $m_1=2.5, m_2=16$
- $K_1=\{2,3,4\}, K_2=\{10,12,20,30,11,25\}$,
- $m_1=3, m_2=18$
- $K_1=\{2,3,4,10\}, K_2=\{12,20,30,11,25\}$, $m_1=4.75, m_2=19.6$
- $K_1=\{2,3,4,10,11,12\}, K_2=\{20,30,25\}$,
- $m_1=7, m_2=25$
- Stop as the clusters with these means are the same.

K-Means Example

Object	X: Weight Index	Y: pH
Sample A	1	1
Sample B	2	1
Sample C	4	3
Sample D	5	4

K-Means Algorithm

Input:

$D = \{t_1, t_2, ..., t_n\}$ // Set of elements

A // Adjacency matrix showing distance between elements.

k // Number of desired clusters.

Output:

K // Set of clusters.

K-Means Algorithm:

assign initial values for means $m_1, m_2, ..., m_k$;

repeat

assign each item t_i to the cluster which has the closest mean ;

calculate new mean for each cluster;

until convergence criteria is met;

Nearest Neighbour (kNN)

- Items are iteratively merged into the existing clusters that are closest.
- Incremental
- Threshold, t, used to determine if items are added to existing clusters or a new cluster is created.

- *To classify a new object based on the attributes and training samples. Parameter k is the number of nearest neighbours*

Nearest Neighbour Algorithm

Input:

$D = \{t_1, t_2, ..., t_n\}$ // Set of elements

A // Adjacency matrix showing distance between elements.

Output:

K // Set of clusters.

Nearest Neighbor Algorithm:

$K_1 = \{t_1\}$;

$K = \{K_1\}$;

$k = 1$;

for $i = 1$ *to* n do

find the t_m in some cluster K_m in K such that $dis(t_i, t_m)$ is the smallest;

if $dis(t_i, t_m) \leq t$ then

$K_m = K_m \cup t_i$

else

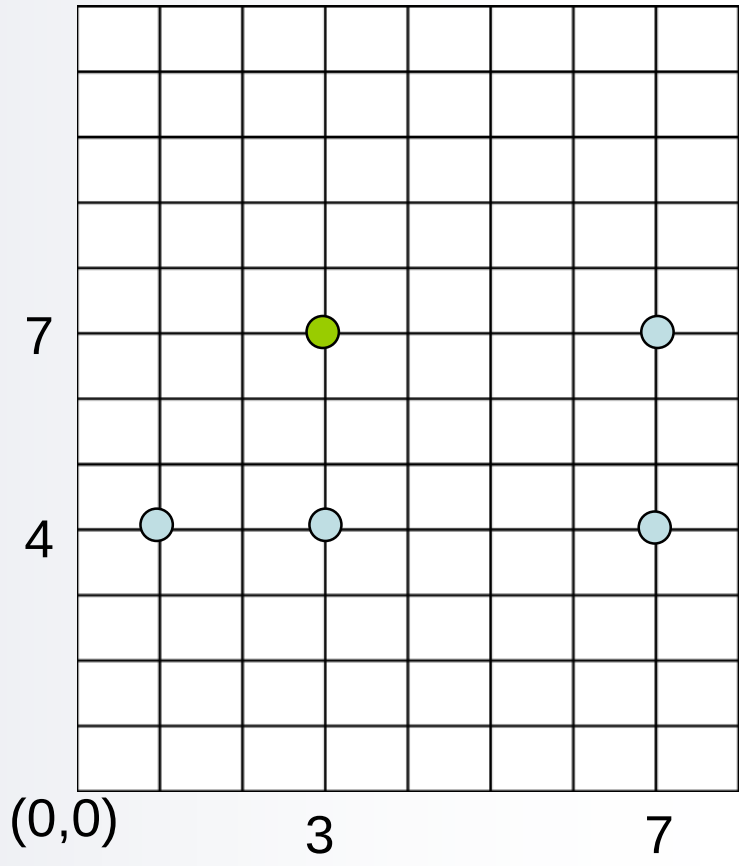
$k = k + 1$;

$K_k = \{t_i\}$;

kNN Example (1)

Acid durability, second	Strength, kg/m^2	Classification
7	7	Bad
7	4	Bad
3	4	Good
1	4	Good

New data (3,7)
Let k = 3
What'll be the classification?



http://people.revoledu.com/kardi/tutorial/KNN/KNN_Numerical-example.html

kNN Example (1)

Acid durability, second	Strength, kg/m^2	Squared distance to query instance (3,7)	Rank	
7	7	16	3	Bad
7	4	25	4	
3	4	9	1	Good
1	4	13	2	Good

1. Rank the squared distance and select the ‘k’ nearest neighbours
2. Use simple majority of the category as the prediction value.

http://people.revoledu.com/kardi/tutorial/KNN/KNN_Numerical-example.html

More Examples

age	income	student	credit_rating	buys_computer
<=30	high	no	fair	no
<=30	high	no	excellent	no
31…40	high	no	fair	yes
>40	medium	no	fair	yes
>40	low	yes	fair	yes
>40	low	yes	excellent	no
31…40	low	yes	excellent	yes
<=30	medium	no	fair	no
<=30	low	yes	fair	yes
>40	medium	yes	fair	yes
<=30	medium	yes	excellent	yes
31…40	medium	no	excellent	yes
31…40	high	yes	fair	yes
>40	medium	no	excellent	no

Let k=5

Age<=30, income=medium, student=yes, credit-rating=fair

Use similarity

<=30	medium	yes	fair	???

age	income	student	credit_rating	buys_computer
<=30	high	no	fair	no
<=30	high	no	excellent	no
31…40	high	no	fair	yes
>40	medium	no	fair	yes
>40	low	yes	fair	yes
>40	low	yes	excellent	no
31…40	low	yes	excellent	yes
<=30	medium	no	fair	no
<=30	low	yes	fair	yes
>40	medium	yes	fair	yes
<=30	medium	yes	excellent	yes
31…40	medium	no	excellent	yes
31…40	high	yes	fair	yes
>40	medium	no	excellent	no

unsorted

<=30	medium	yes	fair	Σ	
1	0	0	1	2	no
1	0	0	0	1	no
0	0	0	1	1	yes
0	1	0	1	2	yes
0	0	1	1	2	yes
0	0	1	0	1	no
0	0	1	0	1	yes
1	1	0	1	3	no
1	0	1	1	3	yes
0	1	1	1	3	yes
1	1	1	0	3	yes
0	1	0	0	1	yes
0	0	1	1	2	yes
0	1	0	0	1	no

sorted

Σ	
3	no
3	yes
3	yes
3	yes
2	no
2	yes
2	yes
2	yes
1	no
1	yes
1	no
1	yes
1	yes
1	no

More Examples

Scenario	Outlook	Temperature	Humidity	Wind	PlayTennis
Day 1	Sunny	Hot	High	Weak	**No**
Day 2	Sunny	Hot	High	Strong	**No**
Day 3	Overcast	Hot	High	Weak	**Yes**
Day 4	Rain	Mild	High	Weak	**Yes**
Day 5	Rain	Cool	Normal	Weak	**Yes**
Day 6	Rain	Cool	Normal	Strong	**No**
Day 7	Overcast	Cool	Normal	Strong	**Yes**
Day 8	Sunny	Mild	High	Weak	**No**
Day 9	Sunny	Cool	Normal	Weak	**Yes**
Day 10	Rain	Mild	Normal	Weak	**Yes**
Day 11	Sunny	Mild	Normal	Strong	**Yes**
Day 12	Overcast	Mild	High	Strong	**Yes**
Day 13	Overcast	Hot	Normal	Weak	**Yes**
Day 14	Rain	Mild	High	Strong	**No**

$$q_1 = \{Outlook = \text{Sunny}, Temperature = \text{Cool}, Humidity = \text{High}, Wind = \text{Strong}\}$$

paul.luminos.nl/download/document/**knn**.pdf

sunny	cool	high	strong	???

outlook	temp	humidity	wind	play_tennis
sunny	hot	high	weak	no
sunny	hot	high	strong	no
overcast	hot	high	weak	yes
rain	mild	high	weak	yes
rain	cool	normal	weak	yes
rain	cool	normal	strong	no
overcast	cool	normal	strong	yes
sunny	mild	high	weak	no
sunny	cool	normal	weak	yes
rain	mild	normal	weak	yes
sunny	mild	normal	strong	yes
overcast	mild	high	strong	yes
overcast	hot	normal	weak	yes
rain	mild	high	strong	no

unsorted

sunny	cool	high	strong	Σ	
1	0	1	0	2	no
1	0	1	1	3	no
0	0	1	0	1	yes
0	0	1	0	1	yes
0	1	0	0	1	yes
0	1	0	1	2	no
0	1	0	1	2	yes
1	0	1	0	2	no
1	1	0	0	2	yes
0	0	0	0	0	yes
1	0	0	1	2	yes
0	0	1	1	2	yes
0	0	0	0	0	yes
0	0	1	1	2	no

sorted

Σ	
3	no
2	no
2	no
2	yes
2	no
2	yes
2	yes
2	yes
2	no
1	yes
1	yes
1	yes
0	yes
0	yes