CHAPTER 2

Unsupervised Learning

(Hierarchical clustering)

Content

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

HIERARCHICAL CLUSTERING

Hierarchical Clustering algorithms

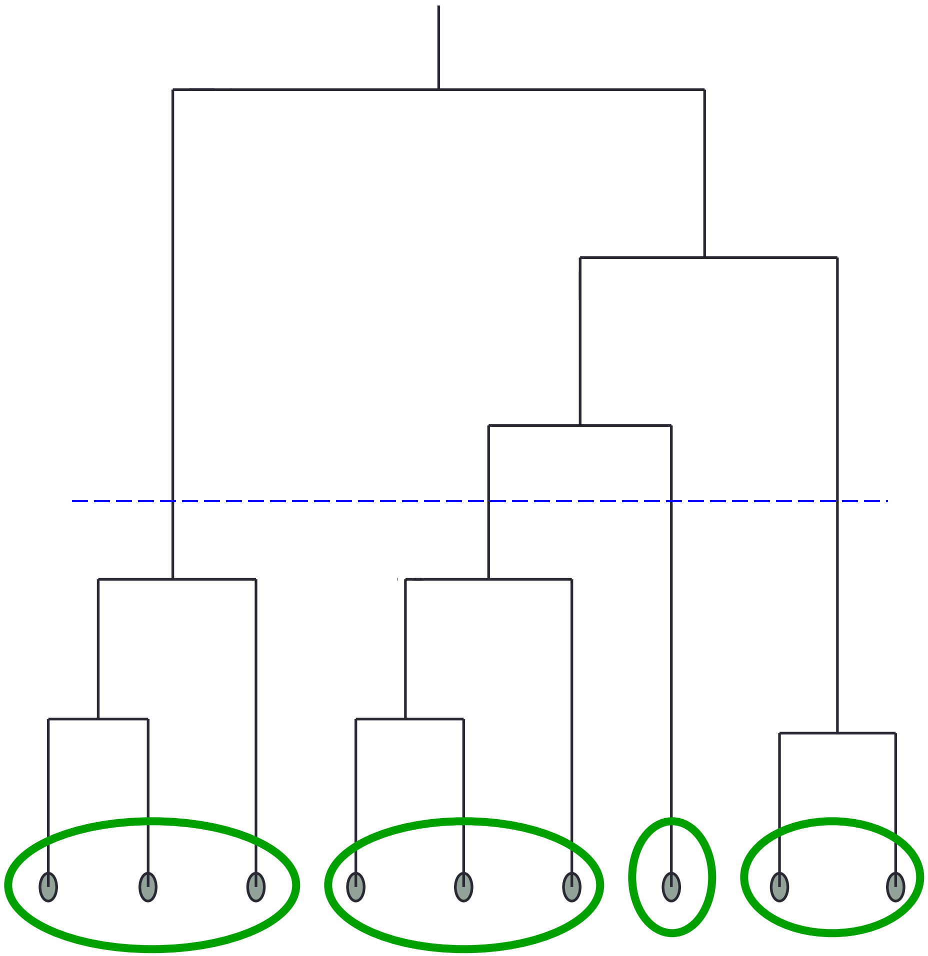
- **Agglomerative (bottom-up):**
  - Start with each document being a single cluster.
  - Eventually all documents belong to the same cluster.
- **Divisive (top-down):**
  - Start with all documents belong to the same cluster.
  - Eventually each node forms a cluster on its own.
  - Could be a recursive application of k-means like algorithms
- Does not require the number of clusters ***k*** in advance
- Needs a termination/readout condition

Hierarchical Agglomerative Clustering (HAC)

- Assumes a similarity function for determining the similarity of two instances.
- Starts with all instances in a separate cluster and then repeatedly joins the two clusters that are most similar until there is only one cluster.
- The history of merging forms a binary tree or hierarchy.

Dendogram: Hierarchical Clustering

- Clustering obtained by cutting the dendrogram at a desired level: each connected component forms a cluster.



Hierarchical Agglomerative Clustering (HAC)

- Starts with each doc in a separate cluster
  - then repeatedly joins the *closest pair* of clusters, until there is only one cluster.
- The history of merging forms a binary tree or hierarchy.

How to measure distance of clusters??

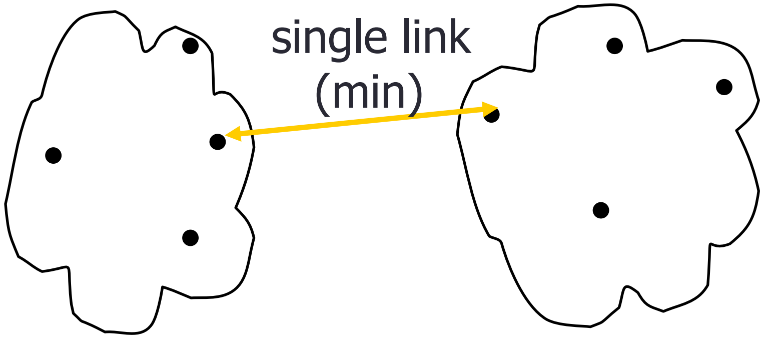
Closest pair of clusters

Many variants to defining closest pair of clusters

- **Single-link**
  - Distance of the *“closest” points* (single-link)
- **Complete-link**
  - Distance of the “furthest” points
- **Average-link**
  - Average distance between pairs of elements

Cluster Distance Measures

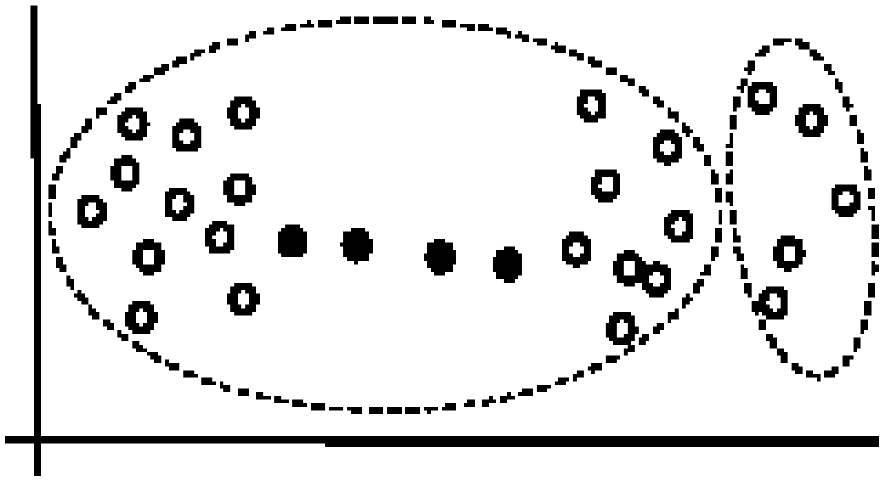
- Single link: smallest distance between an element in one cluster and an element in the other, i.e., $d(C_i, C_j) = \min\{d(x_{ip}, x_{jq})\}$



Obviously, d(C, C)=0

Single link method

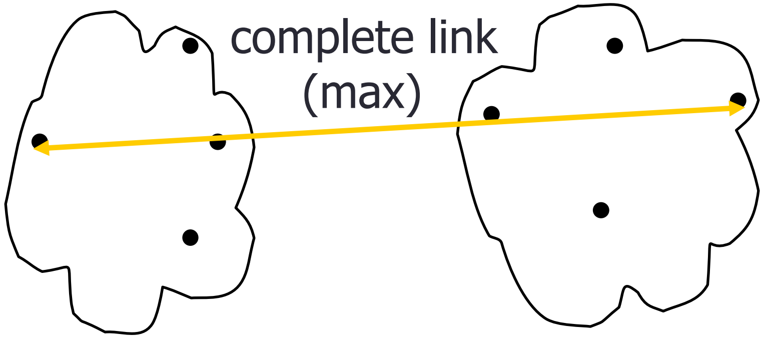
- The distance between two clusters is the distance between two closest data points in the two clusters, one data point from each cluster.
- It can find arbitrarily shaped clusters, but
  - It may cause the undesirable "chain effect" by noisy points



Two natural clusters are split into two

Cluster Distance Measures

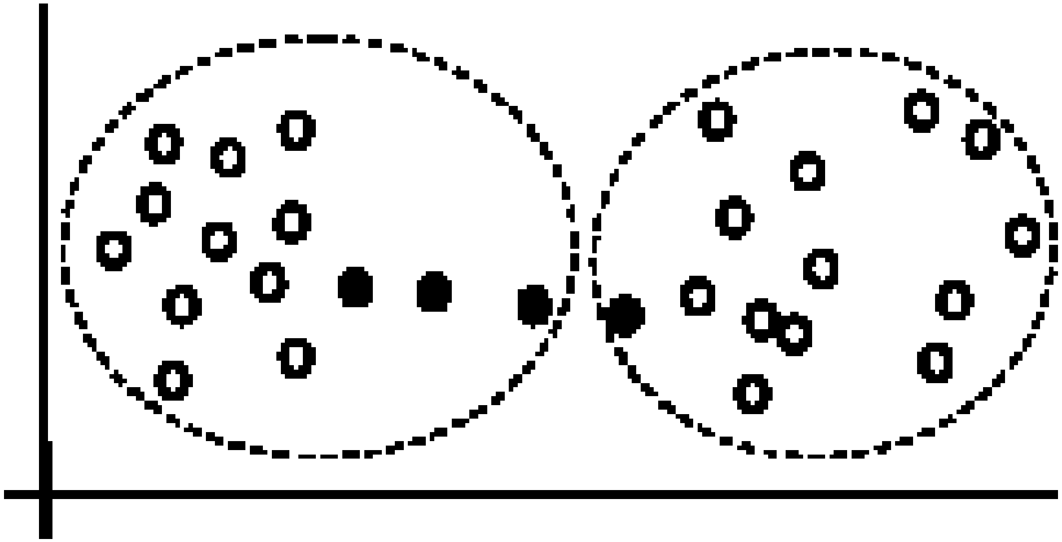
- **Complete link**: largest distance between an element in one cluster and an element in the other, i.e., $d(C_i, C_j) = \max\{d(x_{ip}, x_{jq})\}$



Obviously, d(C, C)=0

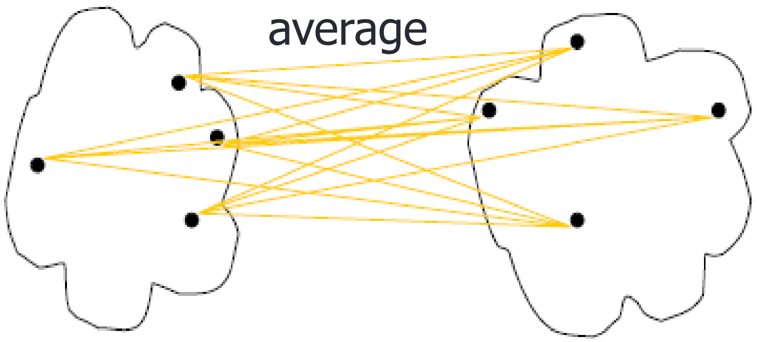
Complete link method

- The distance between two clusters is the distance of two furthest data points in the two clusters.



Cluster Distance Measures

- Average: avg distance between elements in one cluster and elements in the other, i.e., $d(C_i, C_j) = avg\{d(x_{ip}, x_{jq})\}$



Obviously, d(C, C)=0

Cluster Distance Measures

Example: Given a data set of five objects characterized by a single continuous feature, assume that there are two clusters: C1: {a, b} and C2: {c, d, e}.

	a	b	c	d	e
Feature	1	2	4	5	6

Calculate three cluster distances between C1 and C2 by

(i) Single link

(ii) Complete link

(iii) Average link

Cluster Distance Measures

Answer:

	a	b	c	d	e
Feature	1	2	4	5	6

Step 1. Calculate the distance matrix.

$$\begin{array}{c|c} & \begin{array}{ccccc} a & b & c & d & e \end{array} \\ \begin{array}{c} a \\ b \\ c \\ d \\ e \end{array} & \begin{bmatrix} 0 & 1 & 3 & 4 & 5 \\ 1 & 0 & 2 & 3 & 4 \\ 3 & 2 & 0 & 1 & 2 \\ 4 & 3 & 1 & 0 & 1 \\ 5 & 4 & 2 & 1 & 0 \end{bmatrix} \end{array}$$

Cluster Distance Measures

Step 2. Calculate three cluster distances between C1 and C2.

	a	b	c	d	e
a	0	1	3	4	5
b	1	0	2	3	4
c	3	2	0	1	2
d	4	3	1	0	1
e	5	4	2	1	0

Single link

$$\mathrm{dist}(\mathrm{C}_1,\mathrm{C}_2) = \min\{d(\mathrm{a,c}), d(\mathrm{a,d}), d(\mathrm{a,e}), d(\mathrm{b,c}), d(\mathrm{b,d}), d(\mathrm{b,e})\}$$
$$= \min\{3, 4, 5, 2, 3, 4\} = 2$$

Complete link

$$\mathrm{dist}(\mathrm{C}_1,\mathrm{C}_2) = \max\{d(\mathrm{a,c}), d(\mathrm{a,d}), d(\mathrm{a,e}), d(\mathrm{b,c}), d(\mathrm{b,d}), d(\mathrm{b,e})\}$$
$$= \max\{3, 4, 5, 2, 3, 4\} = 5$$

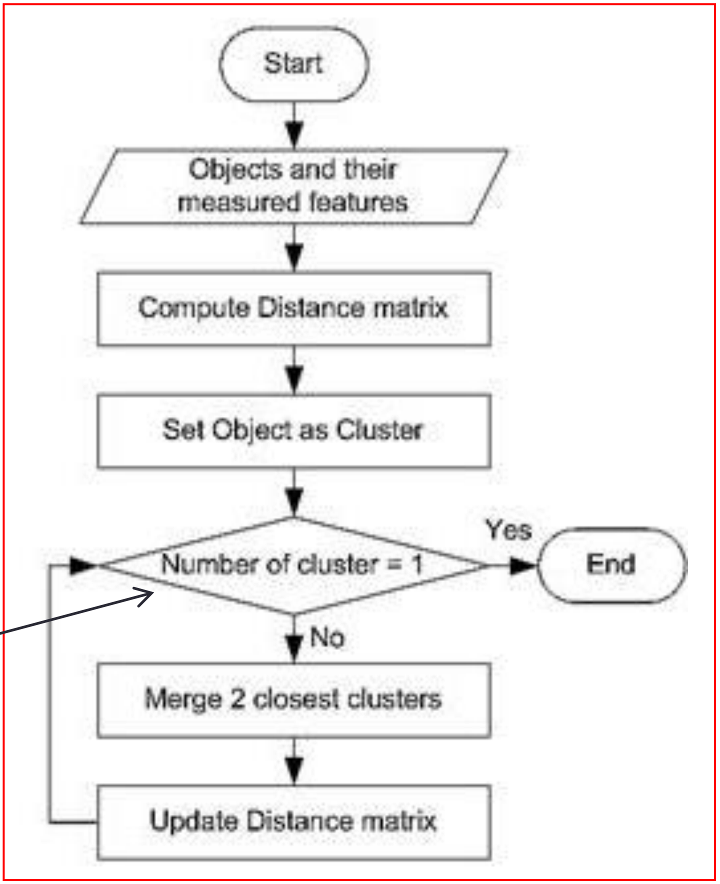
Average

$$\mathrm{dist}(\mathrm{C}_1,\mathrm{C}_2) = \frac{d(\mathrm{a,c}) + d(\mathrm{a,d}) + d(\mathrm{a,e}) + d(\mathrm{b,c}) + d(\mathrm{b,d}) + d(\mathrm{b,e})}{6}$$
$$= \frac{3+4+5+2+3+4}{6} = \frac{21}{6} = 3.5$$

Agglomerative Algorithm

- The *Agglomerative* algorithm is carried out in three steps:

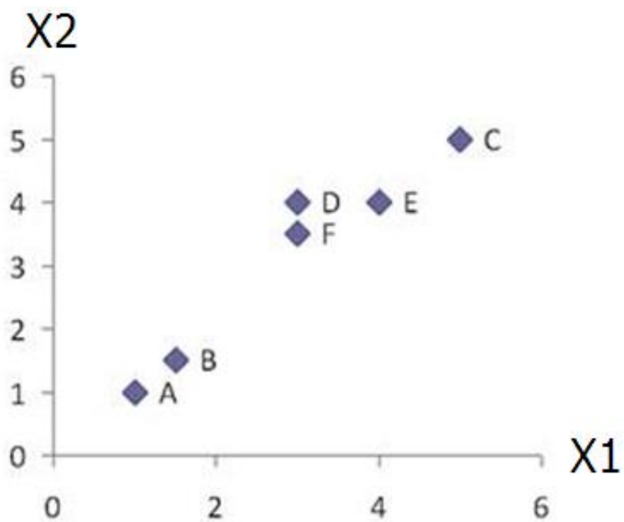
  1) Convert all object features into a distance matrix
  2) Set each object as a cluster (thus if we have *N* objects, we will have *N* clusters at the beginning)
  3) Repeat until number of cluster is one (or known # of clusters)
     - Merge two closest clusters
     - Update "distance matrix"



Example 1

- Given the following data, perform hierarchical clustering analysis with

1) Single linkage
2) Complete linkage
3) Average linkage



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

data matrix

Example 1

- **Answer**: First of all, find the distance matrix.

$$
\begin{array}{c} \\ A \\ B \\ C \\ D \\ E \\ F \end{array}
\begin{array}{c} \begin{array}{cc} X1 & X2 \end{array} \\ \left\{ \begin{array}{cc} 1 & 1 \\ 1.5 & 1.5 \\ 5 & 5 \\ 3 & 4 \\ 4 & 4 \\ 3 & 3.5 \end{array} \right\} \end{array}
$$

data matrix

$$d_{AB} = \left((1-1.5)^2 + (1-1.5)^2 \right)^{\frac{1}{2}} = \sqrt{\tfrac{1}{2}} = 0.7071$$

$$d_{DF} = \left((3-3)^2 + (4-3.5)^2 \right)^{\frac{1}{2}} = 0.5$$

Euclidean distance

	a	b	c	d	e	f
a	0	0.71	5.66	3.61	4.24	3.20
b	0.71	0	4.95	2.92	3.54	2.50
c	5.66	4.95	0	2.24	1.41	2.5
d	3.61	2.92	2.24	0	1	0.5
e	4.24	3.54	1.41	1	0	1.12
f	3.2	2.5	2.5	0.5	1.12	0

Example 1 – Single Linkage

- We have six clusters and they are a,b,c,d,e and f.
- Merge the two closest pair.
- The closest pair is (d,f).

	a	b	c	d	e	f
a	0	0.71	5.66	3.61	4.24	3.20
b	0.71	0	4.95	2.92	3.54	2.50
c	5.66	4.95	0	2.24	1.41	2.5
d	3.61	2.92	2.24	0	1	0.5
e	4.24	3.54	1.41	1	0	1.12
f	3.2	2.5	2.5	0.5	1.12	0

Example 1 – Single Linkage

- Highlight the first group in red

	a	b	c	d	e	f
a	0	0.71	5.66	3.61	4.24	3.20
b	0.71	0	4.95	2.92	3.54	2.50
c	5.66	4.95	0	2.24	1.41	2.5
d	3.61	2.92	2.24	0	1	0.5
e	4.24	3.54	1.41	1	0	1.12
f	3.2	2.5	2.5	0.5	1.12	0

Group	Pair	Distance
1	(d,f)	0.5

- There are C_2^5=6 combinations.
- We compute the minimum distance among these clusters.
- We pick the pair with min. distance.

	a	b	c	d	e	f
a	0	0.71	5.66	3.61	4.24	3.20
b	0.71	0	4.95	2.92	3.54	2.50
c	5.66	4.95	0	2.24	1.41	2.5
d	3.61	2.92	2.24	0	1	0.5
e	4.24	3.54	1.41	1	0	1.12
f	3.2	2.5	2.5	0.5	1.12	0

Pair	Distance
((d,f),a)	$d\big((d,f),a\big) = \min(d(d,a),d(f,a)) = 3.2$
((d,f),b)	$d\big((d,f),b\big) = \min(d(d,b),d(f,b)) = 2.5$
((d,f),c)	$d\big((d,f),c\big) = \min(d(d,c),d(f,c)) = 2.24$
((d,f),e)	$d\big((d,f),e\big) = \min(d(d,e),d(f,e)) = 1$
(a,b)	$d(a,b) = 0.71$
(a,c)	$d(a,c)$ =5.66
(a,e)	$d(a,e)$ =4.24
(b,c)	$d(b,c)$ =4.95
(b,e)	$d(b,e)$ =3.54
(c,e)	$d(c,e)$ =2.5

Example 1 – Single Linkage

- Highlight the second cluster in blue.
- We have four clusters and they are (a,b),c,(d,f) and e.

	a	b	c	d	e	f
a	0	0.71	5.66	3.61	4.24	3.20
b	0.71	0	4.95	2.92	3.54	2.50
c	5.66	4.95	0	2.24	1.41	2.5
d	3.61	2.92	2.24	0	1	0.5
e	4.24	3.54	1.41	1	0	1.12
f	3.2	2.5	2.5	0.5	1.12	0

Group	Pair	Distance
1	(d,f)	0.5
2	(a,b)	0.71

- There are C_2^4=6 combinations.
- We compute the minimum distance among these clusters.
- We pick the pair with min. distance.

	a	b	c	d	e	f
a	0	0.71	5.66	3.61	4.24	3.20
b	0.71	0	4.95	2.92	3.54	2.50
c	5.66	4.95	0	2.24	1.41	2.5
d	3.61	2.92	2.24	0	1	0.5
e	4.24	3.54	1.41	1	0	1.12
f	3.2	2.5	2.5	0.5	1.12	0

Pair	Distance
((a,b),c)	$d\big((a,b),c\big) = \min(d(a,c), d(b,c)) = 4.95$
((a,b),(d,f))	$d\big((a,b),(d,f)\big) = \min(d(a,d), d(a,f), d(b,d), d(b,f)) = 2.50$
((a,b),e)	$d\big((a,b),e\big) = \min(d(a,e), d(b,e)) = 3.54$
(c,(d,f))	$d\big(c,(d,f)\big) = \min\big(d(c,d), d(c,f)\big) = 2.24$
(c,e)	$d(c,e) = 1.41$
((d,f),e)	$d\big((d,f),e\big) = \min(d(d,e), d(f,e)) = 1$

Example 1 – Single Linkage

- Highlight the second cluster in gray.
- We have three clusters and they are (a,b),c and ((d,f),2).

	a	b	c	d	e	f
a	0	0.71	5.66	3.61	4.24	3.20
b	0.71	0	4.95	2.92	3.54	2.50
c	5.66	4.95	0	2.24	1.41	2.5
d	3.61	2.92	2.24	0	1	0.5
e	4.24	3.54	1.41	1	0	1.12
f	3.2	2.5	2.5	0.5	1.12	0

Group	Pair	Distance
1	(d,f)	0.5
2	(a,b)	0.71
3	((d,f),e)	1

- There are C_2^3=3 combinations.
- We compute the minimum distance among these clusters.
- We pick the pair with min. distance.

	a	b	c	d	e	f
a	0	0.71	5.66	3.61	4.24	3.20
b	0.71	0	4.95	2.92	3.54	2.50
c	5.66	4.95	0	2.24	1.41	2.5
d	3.61	2.92	2.24	0	1	0.5
e	4.24	3.54	1.41	1	0	1.12
f	3.2	2.5	2.5	0.5	1.12	0

Pair	Distance
((a,b),c)	$\min(d(a,c),d(b,c)) = 4.95$
((a,b),(e,(d,f)))	$d\left((a,b),\left(e,(d,f)\right)\right) = \min(d(a,e),d(a,d),d(a,f),d(b,e),d(b,d),d(b,f)) = 2.50$
(c,(e,(d,f)))	$d(c,(e,(d,f))) = \min\left(d(c,e),d(c,d),d(c,f)\right) = 1.41$

Example 1 – Single Linkage

- Highlight the second cluster in pale blue.
- We have two clusters and they are (a,b) and (c,((d,f),e)).

	a	b	c	d	e	f
a	0	0.71	5.66	3.61	4.24	3.20
b	0.71	0	4.95	2.92	3.54	2.50
c	5.66	4.95	0	2.24	1.41	2.5
d	3.61	2.92	2.24	0	1	0.5
e	4.24	3.54	1.41	1	0	1.12
f	3.2	2.5	2.5	0.5	1.12	0

Group	Pair	Distance
1	(d,f)	0.5
2	(a,b)	0.71
3	((d,f),e)	1
4	(c,((d,f),e))	1.41

- There is only one combination.

	a	b	c	d	e	f
a	0	0.71	5.66	3.61	4.24	3.20
b	0.71	0	4.95	2.92	3.54	2.50
c	5.66	4.95	0	2.24	1.41	2.5
d	3.61	2.92	2.24	0	1	0.5
e	4.24	3.54	1.41	1	0	1.12
f	3.2	2.5	2.5	0.5	1.12	0

Pair	Distance
((a,b),(c,(e,(d,f))))	$d\Big((a,b),\big(c,\big(e,(d,f)\big)\big)\Big) = \min\big(d(a,c),d(a,e),d(a,d),d(a,f),d(b,c),d(b,e),d(b,d),d(b,f)\big) = 2.50$

Example 1 – Single Linkage

- We only have one cluster and it is ((a,b), (c,((d,f),e))).

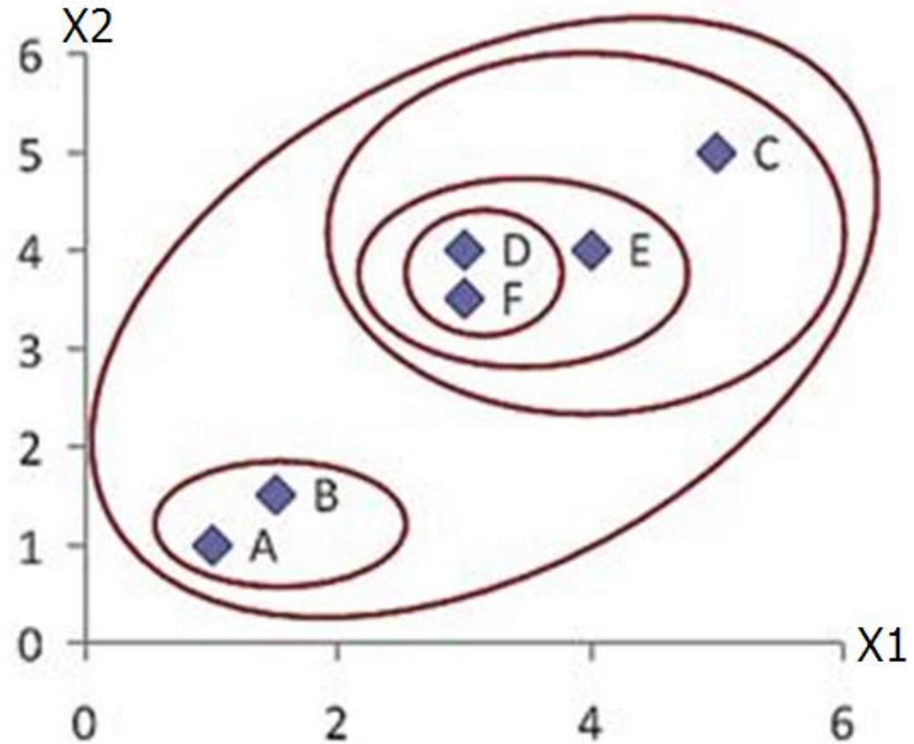
	a	b	c	d	e	f
a	0	0.71	5.66	3.61	4.24	3.20
b	0.71	0	4.95	2.92	3.54	2.50
c	5.66	4.95	0	2.24	1.41	2.5
d	3.61	2.92	2.24	0	1	0.5
e	4.24	3.54	1.41	1	0	1.12
f	3.2	2.5	2.5	0.5	1.12	0

Group	Pair	Distance
1	(d,f)	0.5
2	(a,b)	0.71
3	((d,f),e)	1
4	(c,((d,f),e))	1.41
5	((a,b), (c,((d,f),e)))	2.50

Example 1 – Single Linkage

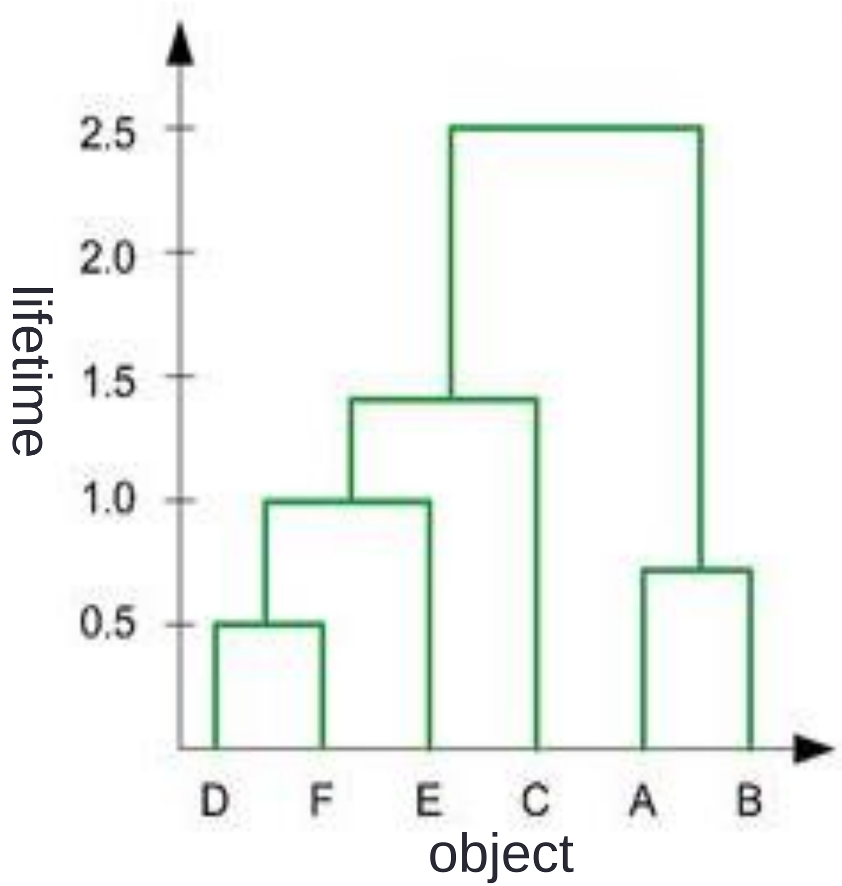
- Final result (meeting termination condition)

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



Example 1 – Single Linkage

- Dendrogram tree representation



1. In the beginning we have 6 clusters: A, B, C, D, E and F
2. We merge clusters D and F into cluster (D, F) at distance 0.50
3. We merge cluster A and cluster B into (A, B) at distance 0.71
4. We merge clusters E and (D, F) into ((D, F), E) at distance 1.00
5. We merge clusters ((D, F), E) and C into (((D, F), E), C) at distance 1.41
6. We merge clusters (((D, F), E), C) and (A, B) into ((((D, F), E), C), (A, B)) at distance 2.50
7. The last cluster contain all the objects, thus conclude the computation

Example 1

- This finished the hierarchical clustering with single linkage.
- Now, we apply the clustering algorithm with complete linkage.

Example 1 – Complete Linkage

- We have six clusters and they are a,b,c,d,e and f.
- Merge the two closest pair.
- The closest pair is (d,f).

	a	b	c	d	e	f
a	0	0.71	5.66	3.61	4.24	3.20
b	0.71	0	4.95	2.92	3.54	2.50
c	5.66	4.95	0	2.24	1.41	2.5
d	3.61	2.92	2.24	0	1	0.5
e	4.24	3.54	1.41	1	0	1.12
f	3.2	2.5	2.5	0.5	1.12	0

Example 1 – Complete Linkage

- Highlight the first group in red

	a	b	c	d	e	f
a	0	0.71	5.66	3.61	4.24	3.20
b	0.71	0	4.95	2.92	3.54	2.50
c	5.66	4.95	0	2.24	1.41	2.5
d	3.61	2.92	2.24	0	1	0.5
e	4.24	3.54	1.41	1	0	1.12
f	3.2	2.5	2.5	0.5	1.12	0

Group	Pair	Distance
1	(d,f)	0.5

- There are C_2^5=6 combinations.
- We compute the maximum distance among these clusters.
- We pick the pair with min. distance.

	a	b	c	d	e	f
a	0	0.71	5.66	3.61	4.24	3.20
b	0.71	0	4.95	2.92	3.54	2.50
c	5.66	4.95	0	2.24	1.41	2.5
d	3.61	2.92	2.24	0	1	0.5
e	4.24	3.54	1.41	1	0	1.12
f	3.2	2.5	2.5	0.5	1.12	0

Pair	Distance
((d,f),a)	$d\big((d,f),a\big) = \max(d(d,a), d(f,a)) = 3.61$
((d,f),b)	$d\big((d,f),b\big) = \max(d(d,b), d(f,b)) = 2.92$
((d,f),c)	$d\big((d,f),c\big) = \max(d(d,c), d(f,c)) = 2.5$
((d,f),e)	$d\big((d,f),e\big) = \max(d(d,e), d(f,e)) = 1.12$
(a,b)	$d(a,b) = 0.71$
(a,c)	$d(a,c)$ =5.66
(a,e)	$d(a,e)$ =4.24
(b,c)	$d(b,c)$ =4.95
(b,e)	$d(b,e)$ =3.54
(c,e)	$d(c,e)$ =2.5

Example 1 – Complete Linkage

- Highlight the second cluster in blue.
- We have four clusters and they are (a,b),c,(d,f) and e.

	a	b	c	d	e	f
a	0	0.71	5.66	3.61	4.24	3.20
b	0.71	0	4.95	2.92	3.54	2.50
c	5.66	4.95	0	2.24	1.41	2.5
d	3.61	2.92	2.24	0	1	0.5
e	4.24	3.54	1.41	1	0	1.12
f	3.2	2.5	2.5	0.5	1.12	0

Group	Pair	Distance
1	(d,f)	0.5
2	(a,b)	0.71

- There are C_2^4=6 combinations.
- We compute the maximum distance among these clusters.
- We pick the pair with min. distance.

	a	b	c	d	e	f
a	0	0.71	5.66	3.61	4.24	3.20
b	0.71	0	4.95	2.92	3.54	2.50
c	5.66	4.95	0	2.24	1.41	2.5
d	3.61	2.92	2.24	0	1	0.5
e	4.24	3.54	1.41	1	0	1.12
f	3.2	2.5	2.5	0.5	1.12	0

Pair	Distance
((a,b),c)	$d\big((a,b),c\big) = max(d(a,c),d(b,c)) = 5.66$
((a,b),(d,f))	$d\big((a,b),(d,f)\big) = max(d(a,d),d(a,f),d(b,d),d(b,f)) = 3.61$
((a,b),e)	$d\big((a,b),e\big) = max(d(a,e),d(b,e)) = 4.24$
(c,(d,f))	$d\big(c,(d,f)\big) = max\big(d(c,d),d(c,f)\big) = 2.50$
(c,e)	$d(c,e) = 1.41$
((d,f),e)	$d\big((d,f),e\big) = \max(d(d,e),d(f,e)) = 1.12$

Example 1 – Complete Linkage

- Highlight the second cluster in gray.
- We have three clusters and they are (a,b),c and ((d,f),2).

	a	b	c	d	e	f
a	0	0.71	5.66	3.61	4.24	3.20
b	0.71	0	4.95	2.92	3.54	2.50
c	5.66	4.95	0	2.24	1.41	2.5
d	3.61	2.92	2.24	0	1	0.5
e	4.24	3.54	1.41	1	0	1.12
f	3.2	2.5	2.5	0.5	1.12	0

Group	Pair	Distance
1	(d,f)	0.5
2	(a,b)	0.71
3	((d,f),e)	1.12

- There are C_2^3=3 combinations.
- We compute the minimum distance among these clusters.
- We pick the pair with min. distance.

	a	b	c	d	e	f
a	0	0.71	5.66	3.61	4.24	3.20
b	0.71	0	4.95	2.92	3.54	2.50
c	5.66	4.95	0	2.24	1.41	2.5
d	3.61	2.92	2.24	0	1	0.5
e	4.24	3.54	1.41	1	0	1.12
f	3.2	2.5	2.5	0.5	1.12	0

Pair	Distance
((a,b),c)	$\max(d(a,c),d(b,c)) = 5.66$
((a,b),(e,(d,f)))	$d\Big((a,b),\big(e,(d,f)\big)\Big) = \max(d(a,e),d(a,d),d(a,f),d(b,e),d(b,d),d(b,f)) = 4.24$
(c,(e,(d,f)))	$d(c,(e,(d,f))) = \max\big(d(c,e),d(c,d),d(c,f)\big) = 2.5$

Example 1 – Complete Linkage

- Highlight the second cluster in pale blue.
- We have two clusters and they are (a,b) and (c,((d,f),e)).

	a	b	c	d	e	f
a	0	0.71	5.66	3.61	4.24	3.20
b	0.71	0	4.95	2.92	3.54	2.50
c	5.66	4.95	0	2.24	1.41	2.5
d	3.61	2.92	2.24	0	1	0.5
e	4.24	3.54	1.41	1	0	1.12
f	3.2	2.5	2.5	0.5	1.12	0

Group	Pair	Distance
1	(d,f)	0.5
2	(a,b)	0.71
3	((d,f),e)	1.12
4	(c,((d,f),e))	2.5

- There is only one combination.

	a	b	c	d	e	f
a	0	0.71	5.66	3.61	4.24	3.20
b	0.71	0	4.95	2.92	3.54	2.50
c	5.66	4.95	0	2.24	1.41	2.5
d	3.61	2.92	2.24	0	1	0.5
e	4.24	3.54	1.41	1	0	1.12
f	3.2	2.5	2.5	0.5	1.12	0

Pair	Distance
((a,b),(c,(e,(d,f))))	$d\Big((a,b),(c,\big(e,(d,f)\big))\Big)$ $= \max\big(d(a,c), d(a,e), d(a,d), d(a,f), d(b,c), d(b,e), d(b,d), d(b,f)\big)$ $= 5.66$

Example 1 – Complete Linkage

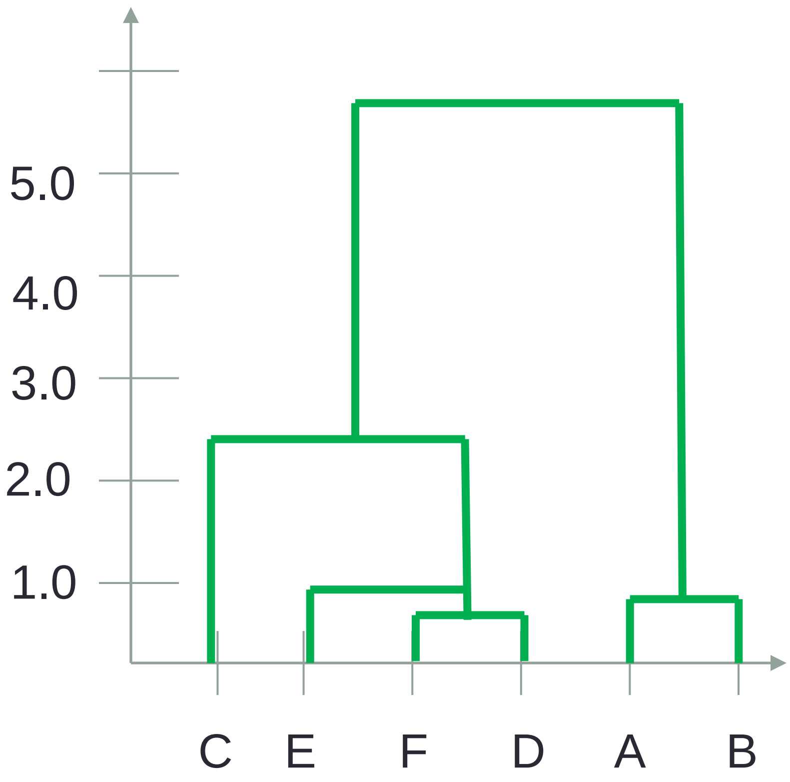
- We only have one cluster and it is ((a,b), (c,((d,f),e))).

	a	b	c	d	e	f
a	0	0.71	5.66	3.61	4.24	3.20
b	0.71	0	4.95	2.92	3.54	2.50
c	5.66	4.95	0	2.24	1.41	2.5
d	3.61	2.92	2.24	0	1	0.5
e	4.24	3.54	1.41	1	0	1.12
f	3.2	2.5	2.5	0.5	1.12	0

Group	Pair	Distance
1	(d,f)	0.5
2	(a,b)	0.71
3	((d,f),e)	1.12
4	(c,((d,f),e))	2.5
5	((a,b), (c,((d,f),e)))	5.66

Example 1 – Complete Linkage

- Dendrogram tree representation



1. In the beginning we have 6 clusters: A, B, C, D, E and F
2. We merge clusters D and F into cluster (D, F) at distance 0.50
3. We merge cluster A and cluster B into (A, B) at distance 0.71
4. We merge clusters E and (D, F) into ((D, F), E) at distance 1.12
5. We merge clusters ((D, F), E) and C into (((D, F), E), C) at distance 2.50
6. We merge clusters (((D, F), E), C) and (A, B) into ((((D, F), E), C), (A, B)) at distance 5.66
7. The last cluster contain all the objects, thus conclude the computation

Example 1

- This finished the hierarchical clustering with single and complete linkages.
- Now, we apply the clustering algorithm with average linkage.

Example 1 – Average Linkage

- We have six clusters and they are a,b,c,d,e and f.
- Merge the two closest pair.
- The closest pair is (d,f).

	a	**b**	**c**	**d**	**e**	**f**
a	0	0.71	5.66	3.61	4.24	3.20
b	0.71	0	4.95	2.92	3.54	2.50
c	5.66	4.95	0	2.24	1.41	2.5
d	3.61	2.92	2.24	0	1	0.5
e	4.24	3.54	1.41	1	0	1.12
f	3.2	2.5	2.5	0.5	1.12	0

Example 1 – Average Linkage

- Highlight the first group in red

	a	b	c	d	e	f
a	0	0.71	5.66	3.61	4.24	3.20
b	0.71	0	4.95	2.92	3.54	2.50
c	5.66	4.95	0	2.24	1.41	2.5
d	3.61	2.92	2.24	0	1	0.5
e	4.24	3.54	1.41	1	0	1.12
f	3.2	2.5	2.5	0.5	1.12	0

Group	Pair	Distance
1	(d,f)	0.5

- There are C_2^5=6 combinations.
- We compute the average distance among these clusters.
- We pick the pair with min. distance.

	a	b	c	d	e	f
a	0	0.71	5.66	3.61	4.24	3.20
b	0.71	0	4.95	2.92	3.54	2.50
c	5.66	4.95	0	2.24	1.41	2.5
d	3.61	2.92	2.24	0	1	0.5
e	4.24	3.54	1.41	1	0	1.12
f	3.2	2.5	2.5	0.5	1.12	0

Pair	Distance
((d,f),a)	$\frac{d(d,a)+d(f,a)}{2} = 3.4050$
((d,f),b)	$\frac{\left(d(d,b)+d(f,b)\right)}{2} = 2.71$
((d,f),c)	$\frac{\left(d(d,c)+d(f,c)\right)}{2} = 2.368$
((d,f),e)	$\frac{d(d,e)+d(f,e)}{2} = 1.059$
(a,b)	$d(a,b) = 0.71$
(a,c)	$d(a,c)$ =5.66
(a,e)	$d(a,e)$ =4.24
(b,c)	$d(b,c)$ =4.95
(b,e)	$d(b,e)$ =3.54
(c,e)	$d(c,e)$ =2.5

Example 1 – Average Linkage

- Highlight the second cluster in blue.
- We have four clusters and they are (a,b),c,(d,f) and e.

	a	b	c	d	e	f
a	0	0.71	5.66	3.61	4.24	3.20
b	0.71	0	4.95	2.92	3.54	2.50
c	5.66	4.95	0	2.24	1.41	2.5
d	3.61	2.92	2.24	0	1	0.5
e	4.24	3.54	1.41	1	0	1.12
f	3.2	2.5	2.5	0.5	1.12	0

Group	Pair	Distance
1	(d,f)	0.5
2	(a,b)	0.71

- There are C_2^4=6 combinations.
- We compute the average distance among these clusters.
- We pick the pair with min. distance.

	a	b	c	d	e	f
a	0	0.71	5.66	3.61	4.24	3.20
b	0.71	0	4.95	2.92	3.54	2.50
c	5.66	4.95	0	2.24	1.41	2.5
d	3.61	2.92	2.24	0	1	0.5
e	4.24	3.54	1.41	1	0	1.12
f	3.2	2.5	2.5	0.5	1.12	0

Pair	Distance
((a,b),c)	$\frac{\left(d(a,c)+d(b,c)\right)}{2} = 5.305$
((a,b),(d,f))	$\frac{\left(d(a,d)+d(a,f)+d(b,d)+d(b,f)\right)}{4} = 3.0556$
((a,b),e)	$\frac{d(a,e)+d(b,e)}{2} = 3.89$
(c,(d,f))	$\frac{\left(d(c,d)+d(c,f)\right)}{2} = 2.37$
(c,e)	$d(c,e) = 1.41$
((d,f),e)	$\frac{(d(d,e)+d(f,e))}{2}$ =1.06

Example 1 – Average Linkage

- Highlight the second cluster in gray.
- We have three clusters and they are (a,b),c and ((d,f),e).

	a	b	c	d	e	f
a	0	0.71	5.66	3.61	4.24	3.20
b	0.71	0	4.95	2.92	3.54	2.50
c	5.66	4.95	0	2.24	1.41	2.5
d	3.61	2.92	2.24	0	1	0.5
e	4.24	3.54	1.41	1	0	1.12
f	3.2	2.5	2.5	0.5	1.12	0

Group	Pair	Distance
1	(d,f)	0.5
2	(a,b)	0.71
3	((d,f),e)	1.06

- There are C_2^3=3 combinations.
- We compute the average distance among these clusters.
- We pick the pair with min. distance.

	a	b	c	d	e	f
a	0	0.71	5.66	3.61	4.24	3.20
b	0.71	0	4.95	2.92	3.54	2.50
c	5.66	4.95	0	2.24	1.41	2.5
d	3.61	2.92	2.24	0	1	0.5
e	4.24	3.54	1.41	1	0	1.12
f	3.2	2.5	2.5	0.5	1.12	0

Pair	Distance
((a,b),c)	$\frac{\left(d(a,c)+d(b,c)\right)}{2}=5.3050$
((a,b),(e,(d,f)))	$\frac{\left(d(a,e)+d(a,d)+d(a,f)+d(b,e)+\ d(b,d)+d(b,f)\right)}{6}=3.335$
(c,(e,(d,f)))	$\frac{\left(d(c,e)+d(c,d)+d(c,f)\right)}{3}=2.05$

Example 1 – Average Linkage

- Highlight the second cluster in pale blue.
- We have two clusters and they are (a,b) and (c,((d,f),e)).

	a	b	c	d	e	f
a	0	0.71	5.66	3.61	4.24	3.20
b	0.71	0	4.95	2.92	3.54	2.50
c	5.66	4.95	0	2.24	1.41	2.5
d	3.61	2.92	2.24	0	1	0.5
e	4.24	3.54	1.41	1	0	1.12
f	3.2	2.5	2.5	0.5	1.12	0

Group	Pair	Distance
1	(d,f)	0.5
2	(a,b)	0.71
3	((d,f),e)	1.06
4	(c,((d,f),e))	2.05

- There is only one combination.

	a	b	c	d	e	f
a	0	0.71	5.66	3.61	4.24	3.20
b	0.71	0	4.95	2.92	3.54	2.50
c	5.66	4.95	0	2.24	1.41	2.5
d	3.61	2.92	2.24	0	1	0.5
e	4.24	3.54	1.41	1	0	1.12
f	3.2	2.5	2.5	0.5	1.12	0

Pair	Distance
((a,b),(c,(e,(d,f))))	$\frac{\left(d(a,c)+d(a,e)+d(a,d)+d(a,f)+d(b,c)+d(b,e)+d(b,d)+d(b,f)\right)}{8} = 3.8275$

Example 1 – Average Linkage

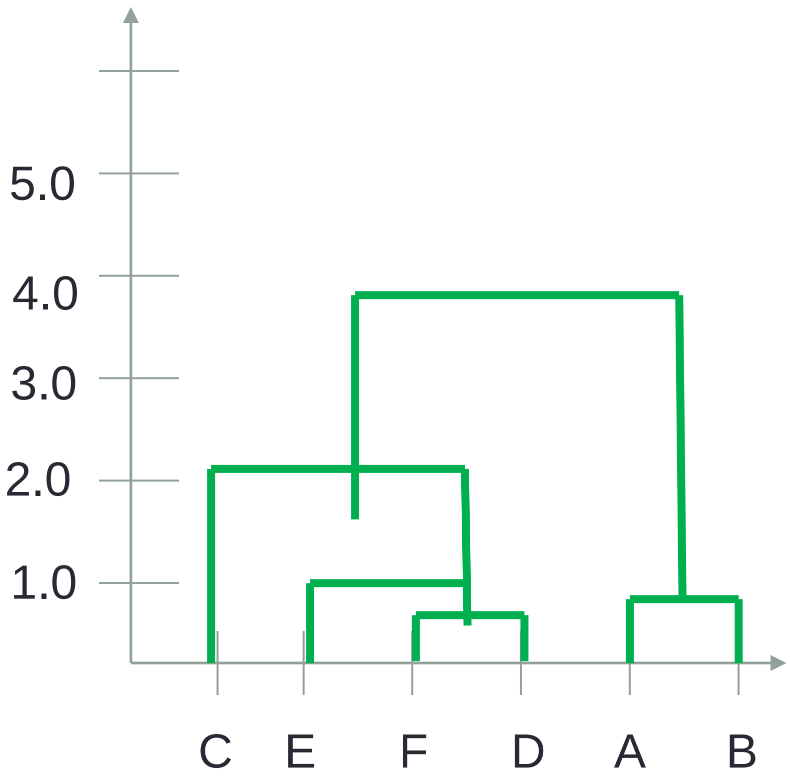
- We only have one cluster and it is ((a,b), (c,((d,f),e))).

	a	b	c	d	e	f
a	0	0.71	5.66	3.61	4.24	3.20
b	0.71	0	4.95	2.92	3.54	2.50
c	5.66	4.95	0	2.24	1.41	2.5
d	3.61	2.92	2.24	0	1	0.5
e	4.24	3.54	1.41	1	0	1.12
f	3.2	2.5	2.5	0.5	1.12	0

Group	Pair	Distance
1	(d,f)	0.5
2	(a,b)	0.71
3	((d,f),e)	1.06
4	(c,((d,f),e))	2.05
5	((a,b), (c,((d,f),e)))	3.8275

Example 1 – Average Linkage

- Dendrogram tree representation



1. In the beginning we have 6 clusters: A, B, C, D, E and F
2. We merge clusters D and F into cluster (D, F) at distance 0.50
3. We merge cluster A and cluster B into (A, B) at distance 0.71
4. We merge clusters E and (D, F) into ((D, F), E) at distance 1.06
5. We merge clusters ((D, F), E) and C into (((D, F), E), C) at distance 2.05
6. We merge clusters (((D, F), E), C) and (A, B) into ((((D, F), E), C), (A, B)) at distance 3.8275
7. The last cluster contain all the objects, thus conclude the computation

Example 2

Given a data set of five objects characterised by a single continuous feature:

	a	b	C	d	e
Feature	1	2	4	5	6

Apply the agglomerative algorithm with single-link, complete-link and averaging cluster distance measures to produce three dendrogram trees, respectively.

Example 2

- **Answer**: First of all, find the distance matrix.

	a	b	C	d	e
Feature	1	2	4	5	6

$$d(a, b) = |1 - 2| = 1$$
$$d(d, e) = |5 - 6| = 1$$

$$\begin{array}{c|ccccc} & \mathbf{a} & \mathbf{b} & \mathbf{c} & \mathbf{d} & \mathbf{e} \\ \hline \mathbf{a} & 0 & 1 & 3 & 4 & 5 \\ \mathbf{b} & 1 & 0 & 2 & 3 & 4 \\ \mathbf{c} & 3 & 2 & 0 & 1 & 2 \\ \mathbf{d} & 4 & 3 & 1 & 0 & 1 \\ \mathbf{e} & 5 & 4 & 2 & 1 & 0 \end{array}$$

Example 2 – Single Linkage

- Merge two closest clusters
- Obviously, the shortest distance is 1.

	a	b	c	d	e
a	0	1	3	4	5
b	1	0	2	3	4
c	3	2	0	1	2
d	4	3	1	0	1
e	5	4	2	1	0

Example 2 – Single Linkage

- There are many pairs with this value.
- We can randomly pick any one of them, say $a \leftrightarrow b$.
- At the end , we may get different clustering results.

	a	**b**	**c**	**d**	**e**
a	0	1	3	4	5
b	1	0	2	3	4
c	3	2	0	1	2
d	4	3	1	0	1
e	5	4	2	1	0

Example 2 – Single Linkage

- Now, we have four clusters. They are (a,b), c, d and e.

	a	b	c	d	e
a	0	1	3	4	5
b	1	0	2	3	4
c	3	2	0	1	2
d	4	3	1	0	1
e	5	4	2	1	0

Group	Pair	Distance
1	(a,b)	1

Example 2 – Single Linkage

- There are C_2^4=6 combinations
- We compute the minimum distance among these clusters.
- We pick the pair with min. distance, which is (c,d) (in fact, (d,e) is also fine).

	a	b	c	d	e
a	0	1	3	4	5
b	1	0	2	3	4
c	3	2	0	1	2
d	4	3	1	0	1
e	5	4	2	1	0

Pair	Distance
((a,b),c)	$d\big((a,b),c\big) = \min(d(a,c),d(b,c)) = 2$
((a,b),d)	$d\big((a,b),d\big) = \min\big(d(a,d),d(b,d)\big) = 3$
((a,b),e)	$d\big((a,b),e\big) = \min\big(d(a,e),d(b,e)\big) = 4$
(c,d)	$d(c,d)$ =1
(c,e)	$d(c,e) = 2$
(d,e)	$d(d,e)$ =1

Example 2 – Single Linkage

- Now, we have three clusters. They are (a,b), (c,d) and e.

	a	b	c	d	e
a	0	1	3	4	5
b	1	0	2	3	4
c	3	2	0	1	2
d	4	3	1	0	1
e	5	4	2	1	0

Group	Pair	Distance
1	(a,b)	1
2	(c,d)	1

Example 2 – Single Linkage

- There are C_2^3=3 combinations
- We compute the minimum distance among these clusters.
- We pick the pair with min. distance, which is ((c,d),e).

	a	b	c	d	e
a	0	1	3	4	5
b	1	0	2	3	4
c	3	2	0	1	2
d	4	3	1	0	1
e	5	4	2	1	0

Pair	Distance
((a,b),(c,d))	$d\big((a,b),(c,d)\big)$ $= \min(d(a,c),d(a,d),d(b,c),d(b,d)) = 2$
((a,b),e)	$d\big((a,b),e\big) = \min\big(d(a,e),d(b,e)\big) = 4$
((c,d),e)	$d\big((c,d),e\big) = \min\big(d(c,e),d(d,e)\big) = 1$

Example 2 – Single Linkage

- Now, we have three clusters. They are (a,b), (c,d) and e.

	a	b	c	d	e
a	0	1	3	4	5
b	1	0	2	3	4
c	3	2	0	1	2
d	4	3	1	0	1
e	5	4	2	1	0

Group	Pair	Distance
1	(a,b)	1
2	(c,d)	1
3	((c,d),e)	1

Example 2 – Single Linkage

- There are two clusters.
- We compute the minimum distance among these clusters.

	a	b	c	d	e
a	0	1	3	4	5
b	1	0	2	3	4
c	3	2	0	1	2
d	4	3	1	0	1
e	5	4	2	1	0

Pair	Distance
((a,b),(e,(c,d)))	$d\big((a,b),(e,(c,d))\big)$ $= \min(d(a,e),d(a,c),d(a,d),d(b,e),d(b,c),d(b,d)) = 2$

Example 2 – Single Linkage

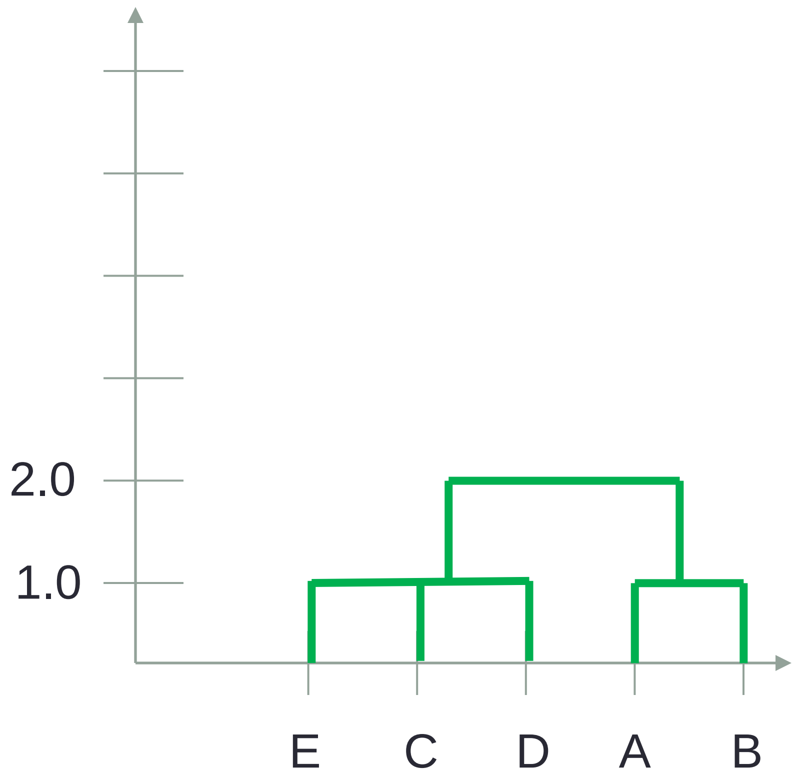
- Now, we have three clusters. They are (a,b), (c,d) and e.

	a	b	c	d	e
a	0	1	3	4	5
b	1	0	2	3	4
c	3	2	0	1	2
d	4	3	1	0	1
e	5	4	2	1	0

Group	Pair	Distance
1	(a,b)	1
2	(c,d)	1
3	((c,d),e)	1
4	((a,b),((c,d),e)	2

Example 2 – Single Linkage

- Dendrogram tree representation



1. In the beginning we have 5 clusters: A, B, C, D and E
2. We merge clusters A and B into cluster (A, B) at distance 1.0
3. We merge cluster C and cluster D into (C, D) at distance 1.0
4. We merge clusters E and (C, D) at distance 1.00
5. We merge clusters (E,(C,D)) and (A,B) at distance 2.0

6. The last cluster contain all the objects, thus conclude the computation

Example 2

- The analysis using single linkage has been finished.
- Now, we use the same technique but with complete linkage.

Example 2 – Complete Linkage

- Merge two closest clusters
- Obviously, the shortest distance is 1.

	a	b	c	d	e
a	0	1	3	4	5
b	1	0	2	3	4
c	3	2	0	1	2
d	4	3	1	0	1
e	5	4	2	1	0

Example 2 – Complete Linkage

- There are many pairs with this value.
- We can randomly pick any one of them, say $a \leftrightarrow b$.
- At the end , we may get different clustering results.

	a	b	c	d	e
a	0	1	3	4	5
b	1	0	2	3	4
c	3	2	0	1	2
d	4	3	1	0	1
e	5	4	2	1	0

Example 2 – Complete Linkage

- Now, we have four clusters. They are (a,b), c, d and e.

	a	b	c	d	e
a	0	1	3	4	5
b	1	0	2	3	4
c	3	2	0	1	2
d	4	3	1	0	1
e	5	4	2	1	0

Group	Pair	Distance
1	(a,b)	1

Example 2 – Complete Linkage

- There are C_2^4=6 combinations
- We compute the maximum distance among these clusters.
- We pick the pair with max. distance, which is (c,d) (in fact, (d,e) is also fine).

	a	b	c	d	e
a	0	1	3	4	5
b	1	0	2	3	4
c	3	2	0	1	2
d	4	3	1	0	1
e	5	4	2	1	0

Pair	Distance
((a,b),c)	$d\big((a,b),c\big) = \max(d(a,c),d(b,c)) = 3$
((a,b),d)	$d\big((a,b),d\big) = \max\big(d(a,d),d(b,d)\big) = 4$
((a,b),e)	$d\big((a,b),e\big) = \max\big(d(a,e),d(b,e)\big) = 5$
(c,d)	$d(c,d)$ =1
(c,e)	$d(c,e) = 2$
(d,e)	$d(d,e)$ =1

Example 2 – Complete Linkage

- Now, we have three clusters. They are (a,b), (c,d) and e.

	a	b	c	d	e
a	0	1	3	4	5
b	1	0	2	3	4
c	3	2	0	1	2
d	4	3	1	0	1
e	5	4	2	1	0

Group	Pair	Distance
1	(a,b)	1
2	(c,d)	1

Example 2 – Complete Linkage

- There are C_2^3=3 combinations
- We compute the maximum distance among these clusters.
- We pick the pair with mx. distance, which is ((c,d),e).

	a	b	c	d	e
a	0	1	3	4	5
b	1	0	2	3	4
c	3	2	0	1	2
d	4	3	1	0	1
e	5	4	2	1	0

Pair	Distance
((a,b),(c,d))	$d\big((a,b),(c,d)\big) = \max(d(a,c),d(a,d),d(b,c),d(b,d)) = 4$
((a,b),e)	$d\big((a,b),e\big) = max\big(d(a,e),d(b,e)\big) = 5$
((c,d),e)	$d\big((c,d),e\big) = max\big(d(c,e),d(d,e)\big) = 2$

Example 2 – Complete Linkage

- Now, we have three clusters. They are (a,b), (c,d) and e.

	a	b	c	d	e
a	0	1	3	4	5
b	1	0	2	3	4
c	3	2	0	1	2
d	4	3	1	0	1
e	5	4	2	1	0

Group	Pair	Distance
1	(a,b)	1
2	(c,d)	1
3	((c,d),e)	2

Example 2 – Complete Linkage

- There are two clusters.
- We compute the minimum distance among these clusters.

	a	b	c	d	e
a	0	1	3	4	5
b	1	0	2	3	4
c	3	2	0	1	2
d	4	3	1	0	1
e	5	4	2	1	0

Pair	Distance
((a,b),(e,(c,d)))	$d\big((a,b),(e,(c,d))\big) = \max(d(a,e),d(a,c),d(a,d),d(b,e),d(b,c),d(b,d)) = 5$

Example 2 – Complete Linkage

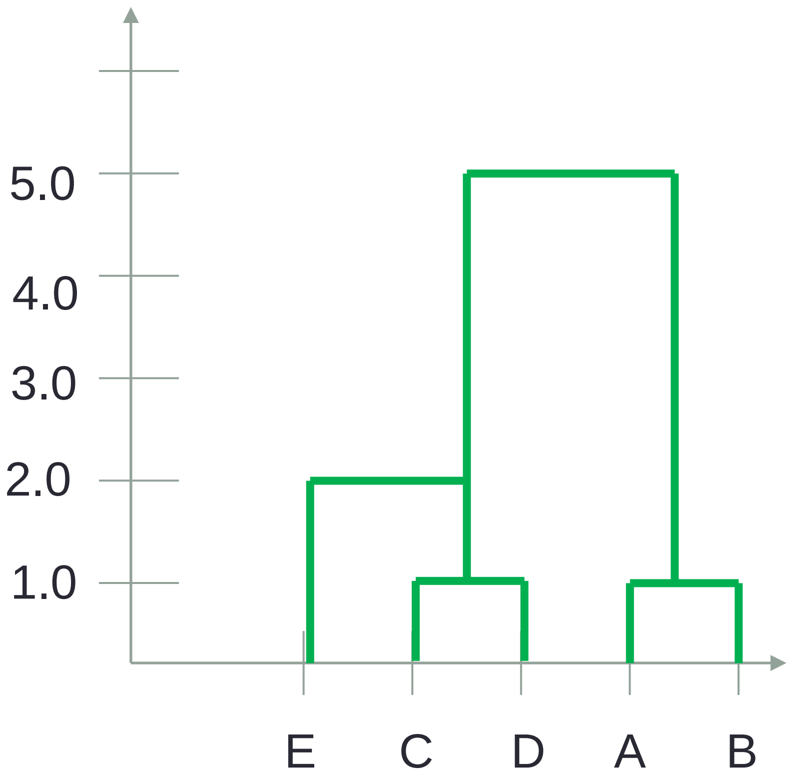
- Now, we have three clusters. They are (a,b), (c,d) and e.

	a	b	c	d	e
a	0	1	3	4	5
b	1	0	2	3	4
c	3	2	0	1	2
d	4	3	1	0	1
e	5	4	2	1	0

Group	Pair	Distance
1	(a,b)	1
2	(c,d)	1
3	((c,d),e)	2
4	((a,b),((c,d),e)	5

Example 2 – Complete Linkage

- Dendrogram tree representation



1. In the beginning we have 5 clusters: A, B, C, D and E
2. We merge clusters A and B into cluster (A, B) at distance 1.0
3. We merge cluster C and cluster D into (C, D) at distance 1.0
4. We merge clusters E and (C, D) at distance 2.00
5. We merge clusters (E,(C,D)) and (A,B) at distance 5.0
6. The last cluster contain all the objects, thus conclude the computation

Example 2

- The analysis using single linkage and complete linkage have been finished.
- Now, we use the same technique but with average linkage.

Example 2 – Average Linkage

- Merge two closest clusters
- Obviously, the shortest distance is 1.

	a	b	c	d	e
a	0	1	3	4	5
b	1	0	2	3	4
c	3	2	0	1	2
d	4	3	1	0	1
e	5	4	2	1	0

Example 2 – Average Linkage

- There are many pairs with this value.
- We can randomly pick any one of them, say $a \leftrightarrow b$.
- At the end , we may get different clustering results.

	a	b	c	d	e
a	0	1	3	4	5
b	1	0	2	3	4
c	3	2	0	1	2
d	4	3	1	0	1
e	5	4	2	1	0

Group	Pair	Distance
1	(a,b)	1

Example 2 – Average Linkage

- Now, we have four clusters and they are (a,b), c, d and e.
- There are C_2^4=6 combinations
- We compute the average distance among these clusters.
- We pick the pair with min. distance, which is (d,e) (in fact, (c,d) is also fine).

	a	**b**	**c**	**d**	**e**
a	0	1	3	4	5
b	1	0	2	3	4
c	3	2	0	1	2
d	4	3	1	0	1
e	5	4	2	1	0

Pair	Distance
((a,b),c)	$d\big((a,b),c\big) = \dfrac{d(a,c)+d(b,c)}{2} = 2.5$
((a,b),d)	$d\big((a,b),d\big) = \dfrac{d(a,d)+d(b,d)}{2} = 3.5$
((a,b),e)	$d\big((a,b),e\big) = \dfrac{d(a,e)+d(b,e)}{2} = 4.5$
(c,d)	$d(c,d)$ =1
(c,e)	$d(c,e) = 2$
(d,e)	$d(d,e)$ =1

Example 2 – Average Linkage

- Now, we have three clusters and they are (a,b), (c,d) and e.

	a	b	c	d	e
a	0	1	3	4	5
b	1	0	2	3	4
c	3	2	0	1	2
d	4	3	1	0	1
e	5	4	2	1	0

Group	Pair	Distance
1	(a,b)	1
2	(c,d)	1

Example 2 – Average Linkage

- Now, we have three clusters and they are (a,b), (c,d) and e.
- There are C_2^3=3 combinations
- We compute the average distance among these clusters.
- We pick the pair with min. distance, which is (c,(d,e)).

	a	b	c	d	e
a	0	1	3	4	5
b	1	0	2	3	4
c	3	2	0	1	2
d	4	3	1	0	1
e	5	4	2	1	0

Pair	Distance
((a,b),(c,d))	$d\big((a,b),(c,d)\big) = \frac{d(a,c)+d(b,c)+d(a,d)+d(b,d)}{4} = 3$
((a,b),e)	$d\big((a,b),e\big) = \frac{d(a,e)+d(b,e)}{2} = 4.5$
((c,d),e)	$d\text{((c,d),e)} = \frac{d(c,e)+d(d,e)}{2} = 1.5$

Example 2 – Average Linkage

- Now, we have two clusters and they are (a,b) and (c,(d,e)).

	a	b	c	d	e
a	0	1	3	4	5
b	1	0	2	3	4
c	3	2	0	1	2
d	4	3	1	0	1
e	5	4	2	1	0

Group	Pair	Distance
1	(a,b)	1
2	(c,d)	1
3	((c,d),e))	1.5

Example 2 – Average Linkage

- We compute the average distance among these clusters.

	a	b	c	d	e
a	0	1	3	4	5
b	1	0	2	3	4
c	3	2	0	1	2
d	4	3	1	0	1
e	5	4	2	1	0

Pair	Distance
((a,b),(c,d),e)))	$d\big((a,b),(c,(d,e))\big) = 3.5$

Example 2 – Average Linkage

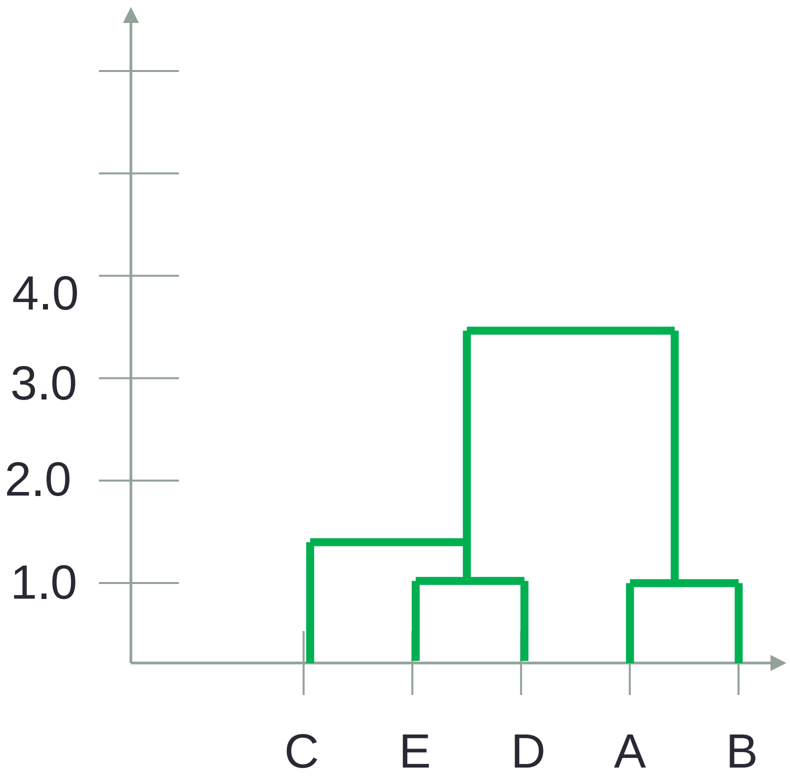
- Now, we only have one cluster.

	a	b	c	d	e
a	0	1	3	4	5
b	1	0	2	3	4
c	3	2	0	1	2
d	4	3	1	0	1
e	5	4	2	1	0

Group	Pair	Distance
1	(a,b)	1
2	(c,d)	1
3	((c,d),e))	1.5
4	((a,b),(c,d),e))	3.5

Example 2 – Average Linkage

- Dendrogram tree representation



1. In the beginning we have 5 clusters: A, B, C, D and E
2. We merge clusters A and B into cluster (A, B) at distance 1.0
3. We merge cluster C and cluster D into (C, D) at distance 1.0
4. We merge clusters E and (C, D) at distance 1.50
5. We merge clusters (E,(C,D)) and (A,B) at distance 3.5

6. The last cluster contain all the objects, thus conclude the computation