Cluster Analysis

Statistics 407, ISU

Definition

The aim of cluster analysis is to ________ cases (objects) according to their _________ on the variables. It is also often called _______________ classification, meaning that classification is the ultimate goal, but the classes (groups) are not known ahead of time.

Hence the first task in cluster analysis is to construct the class information. To determine closeness we start with measuring the _________ distances.

Distance Measures

Let $\mathbf{X} = (X_1\ X_2\ \ldots X_p)'$ and $\mathbf{Y} = (Y_1\ Y_2\ \ldots Y_p)'$ be two points in p-space (two rows of a data matrix).

Euclidean Distance:

$$d(\mathbf{X},\mathbf{Y}) = \sqrt{(\mathbf{X}-\mathbf{Y})'(\mathbf{X}-\mathbf{Y})} = \sqrt{(X_1-Y_1)^2 + \ldots + (X_p-Y_p)^2}$$

Statistical Distance:

$$d(\mathbf{X},\mathbf{Y}) = \sqrt{(\mathbf{X}-\mathbf{Y})'S^{-1}(\mathbf{X}-\mathbf{Y})}$$

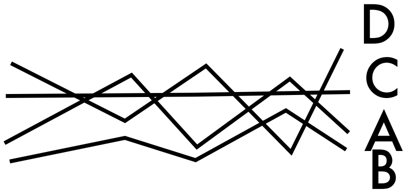
Both of these distance measures benefit from ___________ the variables first.

Distance Metrics

- Kendall tau distance: _____ each variable. For all ______ of elements of the two points, count 1 for each pair which the ranks are in the same relationship (___, ___; ___, ____) and 0 otherwise.
- Measures the _______ between two points, eg height value is often as highly ranked as weight value says that the two variables are positively correlated. Not as affected by outliers as raw data values.

Distance Metrics

- Pearson correlation: d=____
  - d=_ when r=1
- Pearson square correlation: d=_____
  - d=_ when r=0, d=_ when r=1 or -1
- Measures the similarity in _____, rather than global closeness



- Also can be considered to be ______ distance

Hierarchical Clustering

- Hierarchical algorithms sequentially ____ (or ____) cases to make clusters.
- Process can be viewed using a __________.
- The vertical heights of the dendrogram are used to decide __________________.

Linkage

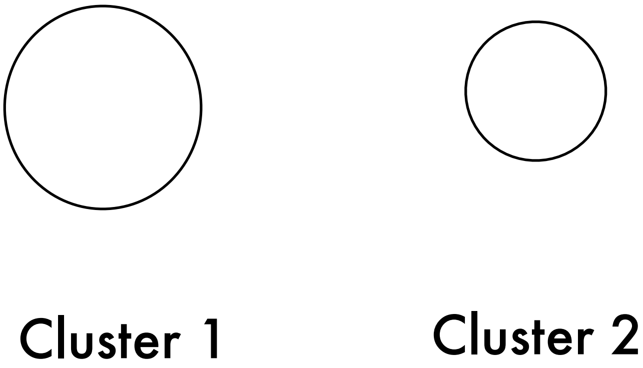
When a cluster is formed, containing two or more cases, there are now multiple ways to define the _______ from the _______ to other _______ or cases. For example, we could define the distance from one cluster to another as the minimum interpoint distance, or the maximum interpoint distance or the average interpoint distance. These are called ________ __________. Each method changes the results of the cluster analysis.

Common linkage methods

The intercluster distance is described by:

- **Single:** the distance between the two ______ points.
- **Complete:** the distance between the two ______ points.
- **Average:** the _______ of all the interpoint distance.
- **Centroid:** the distance between the two ______.
- **Wards:** the smallest increase in the ____________ ________ after fusing two clusters, like ANOVA.

Single Linkage

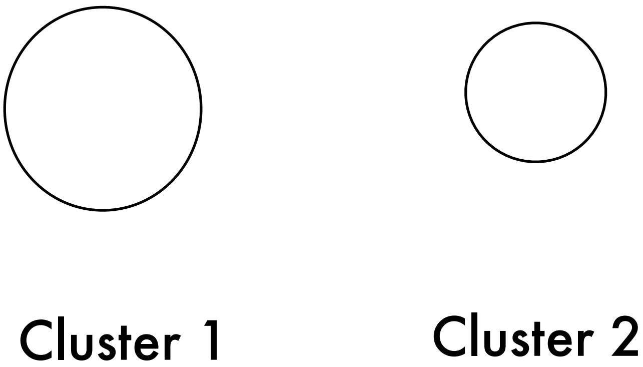


Cluster 1

Cluster 2

Closest points define the intercluster distance

Complete Linkage

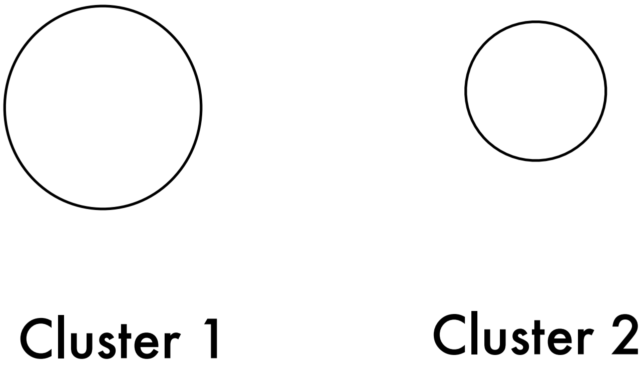


Cluster 1

Cluster 2

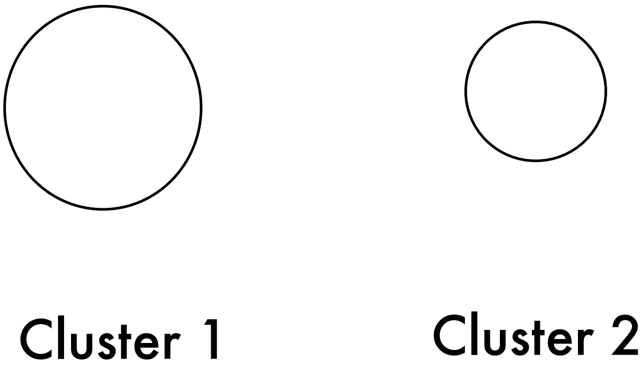
Farthest points define the intercluster distance

Average Linkage



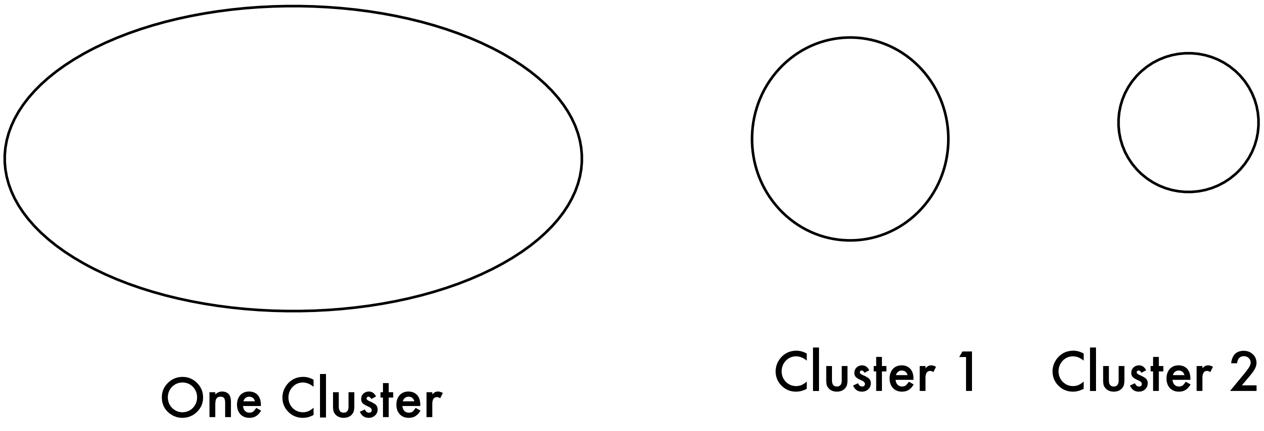
Average of all of the distances defines the intercluster distance

Centroid Linkage



Distance between the cluster means defines the intercluster distance

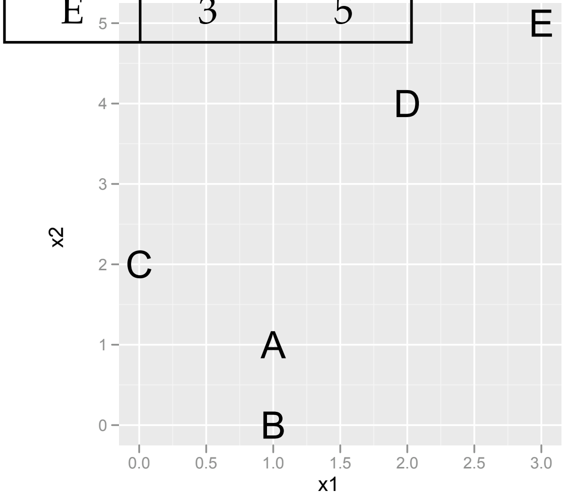
Ward Linkage



Ratio of sum of squared distance from means, between one cluster, and the two clusters defines the intercluster distance

Example

i	X_1	X_2
A	1	1
B	1	0
C	0	2
D	2	4
E	3	5



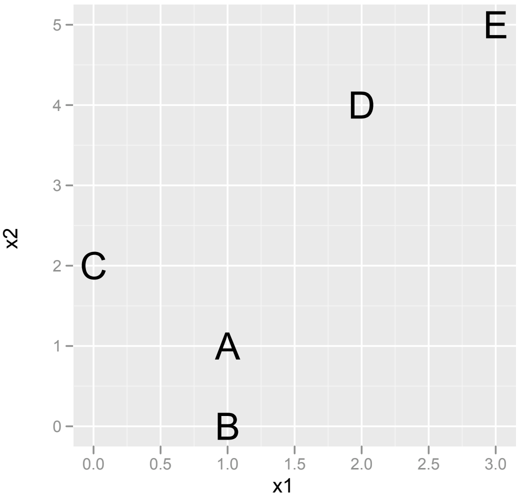
Euclidean distances

	A	B	C	D	E
A	0	1	1.4	3.2	4.5
B	1	0	2.2	4.1	5.4
C			0		4.2
D	3.2	4.1	2.8	0	1.4
E	4.5		4.2	1.4	0

Step 1.1

Join the two closest points into a cluster.

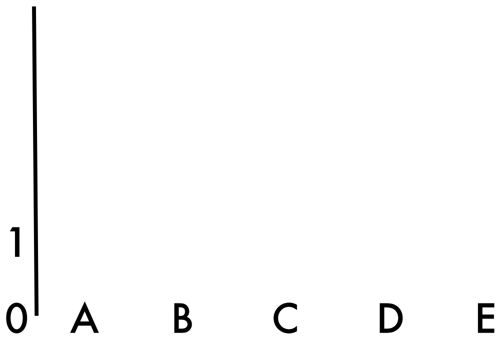
	A	B	C	D	E
A	0	1	1.4	3.2	4.5
B	1	0	2.2	4.1	5.4
C	1.4	2.2	0	2.8	4.2
D	3.2	4.1	2.8	0	1.4
E	4.5	5.4	4.2	1.4	0



Step 1.2

Reduce the distance matrix, using the linkage methods. Draw the dendrogram.

	AB	C	D	E
AB	0	1.8		
C	1.8	0	2.8	4.2
D	3.6		0	1.4
E		4.2	1.4	0

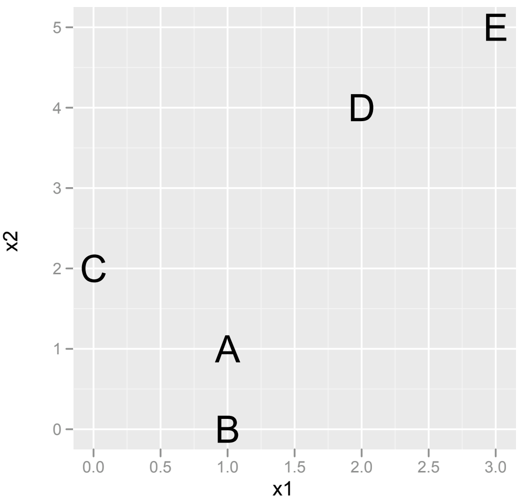


Average linkage used.

Step 2.1

Join the two closest points into a cluster.

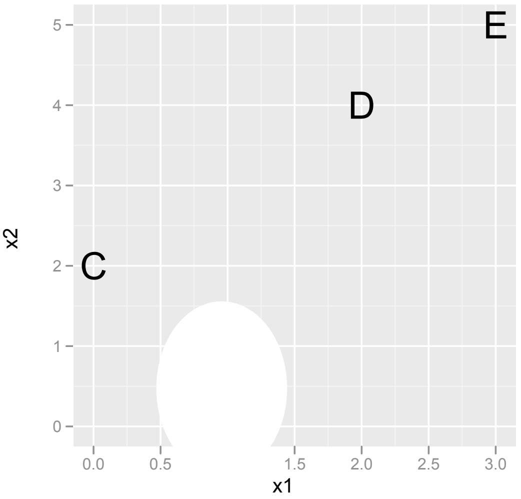
	AB	C	D	E
AB	0	1.8	3.6	4.9
C	1.8	0	2.8	4.2
D	3.6	2.8	0	1.4
E	4.9	4.2	1.4	0



Step 2.1

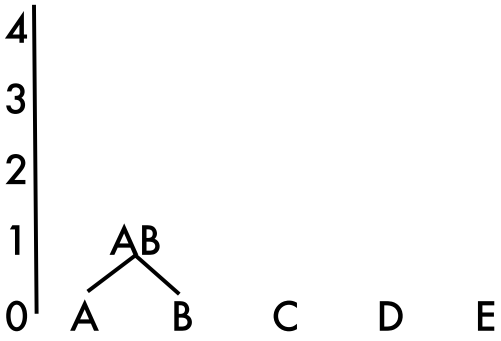
Join the two closest points into a cluster.

	AB	C	D	E
AB	0	1.8	3.6	4.9
C	1.8	0	2.8	4.2
D	3.6	2.8	0	1.4
E	4.9	4.2	1.4	0



Step 2.2

Reduce the distance matrix, using the linkage methods. Draw the dendrogram.



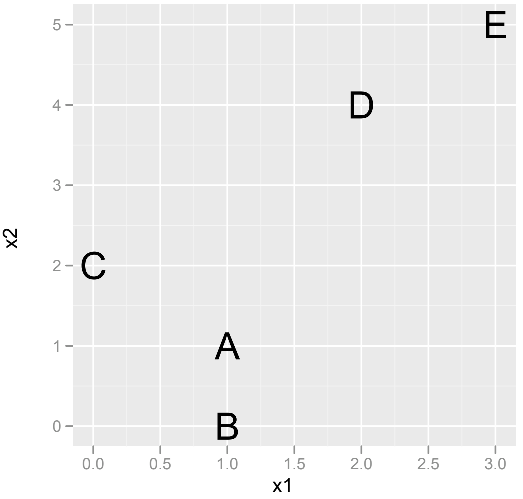
	AB	C	DE
AB	0	1.8	
C		0	
DE	4.3	3.5	0

Average linkage used.

Step 3.1

Join the two closest points into a cluster.

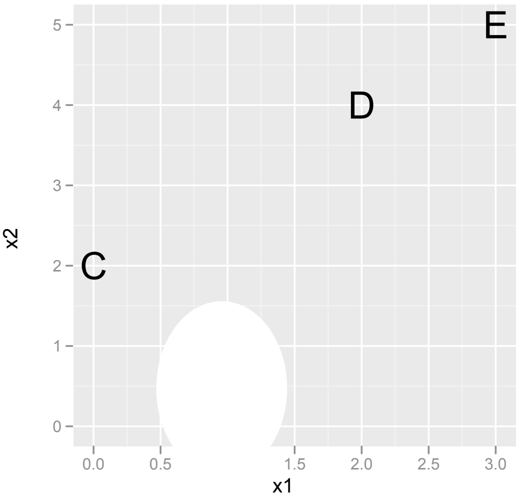
	AB	C	DE
AB	0	1.8	4.3
C	1.8	0	3.5
DE	4.3	3.5	0



Step 3.1

Join the two closest points into a cluster.

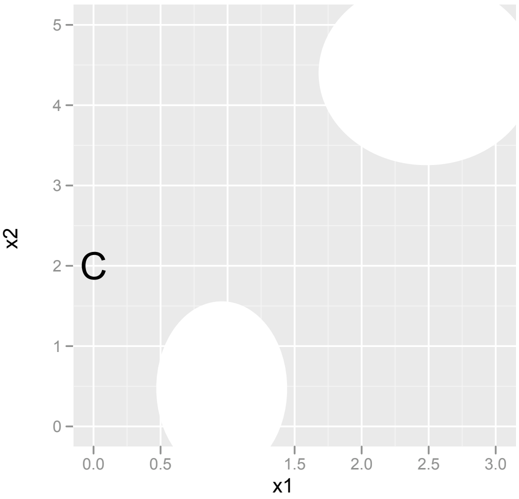
	AB	C	DE
AB	0	1.8	4.3
C	1.8	0	3.5
DE	4.3	3.5	0



Step 3.1

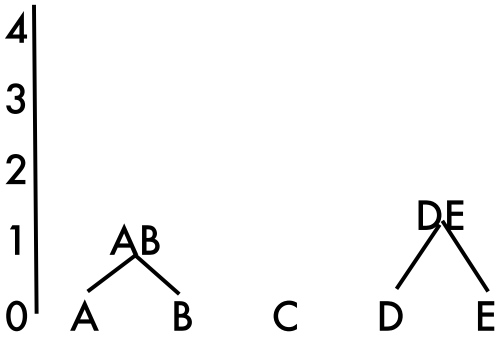
Join the two closest points into a cluster.

	AB	C	DE
AB	0	1.8	4.3
C	1.8	0	3.5
DE	4.3	3.5	0



Step 2.2

Reduce the distance matrix, using the linkage methods. Draw the dendrogram.

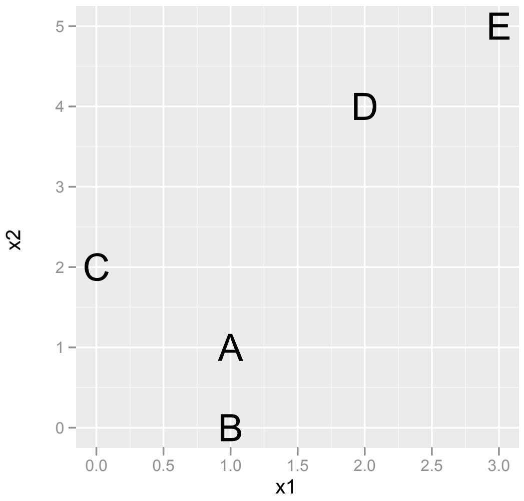


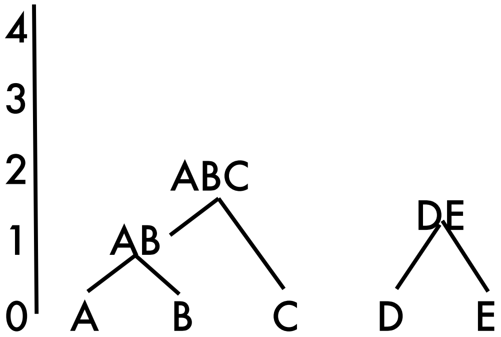
	ABC	DE
ABC	0	
DE	4.0	0

Average linkage used.

Step 3

Join last two clusters


E
D
C
A
B
5
4
3
2
1
0
x2
0.0
0.5
1.0
1.5
2.0
2.5
3.0
x1


4
3
2
1
0
ABC
AB
DE
A
B
C
D
E