

# Bio 311: Introduction to Clustering

Paul M. Magwene

# What is Clustering?

“Clustering” is a broad term for algorithms in statistics and machine learning that try to discover “natural groups” in data.

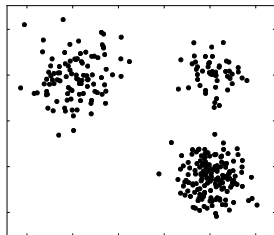
What’s a “natural group”?

- Common sense definition: Groups of objects (or variables) where similarity between objects is higher within groups than between groups

# Natural Groups: Geometric Perspective

What's a “natural group”?

- Geometric definition: Patches of high density points surrounded by patches of lower density in the  $p$ -dimensional space defined by the variates.



# Similarity/Dissimilarity

## Intuition

Similarity is a measure of “likeness” between two entities of interest. Dissimilarity is the complement of similarity.

- Dissimilarities may be converted to similarities (and vice versa) by taking any monotonically decreasing function. For example:

$$s = 1 - d_{ij} \text{ (for } 0 \leq d_{ij} \leq 1)$$

- Dissimilarities are usually in range  $0 \leq d_{ij} \leq C$  where  $C$  is the maximum dissimilarity
- Distances are one measure of dissimilarity but distances are unbounded to the right

$$d_{ij} \in [0, \infty]$$

# Dissimilarity Measures for Quantitative Data

This simplest measure of dissimilarity is Euclidean distance.

$$d_{ij} = \left\{ \sum_{k=1}^p (x_{ik} - x_{jk})^2 \right\}^{1/2}$$

## Dissimilarity Measures for Quantitative Data, cont.

- Manhattan (taxi cab, city block) distance

$$d_{ij} = \sum_{k=1}^p |x_{ik} - x_{jk}|$$

- Chebychev distance

$$d_{ij} = \max_k \{|x_{ik} - x_{jk}|\}$$

- Minkowski Metric

$$d_{ij} = \left\{ \sum_{k=1}^p |x_{ik} - x_{jk}|^\lambda \right\}^{1/\lambda}$$

$\lambda = 1$  is Manhattan distance,  $\lambda = 2$  is Euclidean distance,  
 $\lambda = \infty$  is Chebychev distance.

# Dissimilarity Measures for Variables

Correlation provides a suitable measure of *similarity*. Common *dissimilarity* measures based on correlation include:

- $d_{kl} = 1 - r_{kl}$  if  $r_{kl} = -1$  is taken to indicate maximum disagreement
- $d_{kl} = 1 - r_{kl}^2$  if  $r_{kl} = 1$  and  $r_{kl} = -1$  are treated equivalently (predictive power)
- Based on uncentered correlation:

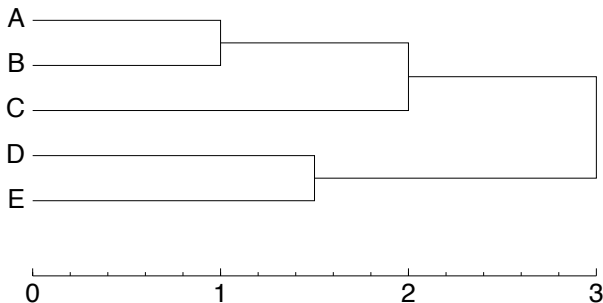
$$d_{kl} = 1 - \frac{\sum_{i=1}^n x_{ik} x_{il}}{\sum_{i=1}^n x_{ik}^2 \sum_{i=1}^n x_{il}^2}$$

# Hierarchical Clustering



# Clustering Method: Hierarchical Clustering

For  $n$  data points define a set of  $n - 1$  “joins” that represent the groupings of objects at different levels of similarity. Represent the series of joins as a “tree” graph.



# Generic Algorithm for Hierarchical Clustering

- 1 Calculate a dissimilarity matrix for the  $n$  items
- 2 Join the two nearest items,  $i$  and  $j$
- 3 Delete the  $i$ -th and  $j$ -th rows and columns of the dissimilarity matrix; and a new row/column that represents the dissimilarity of a new group  $(i,j)$  to all other items
- 4 Repeat from step 2 until there is a single group

## Key Point

The different hierarchical clustering methods are determined by the function used to calculate the distance between groups in step 3.

# Single Linkage Clustering

## Group Distance Measure

Let  $i$  and  $j$  be groups, and  $n_i$  and  $n_j$  be the number of objects in the respective groups.

$D_{ij}$  is the *smallest* of the  $n_i n_j$  dissimilarities between each element of  $i$  and each element of  $j$

## Properties of Single Linkage Clustering

- Invariant under monotonic transformation of the  $d_{ij}$
- Unaffected by ties
- Provably nice asymptotic properties
- Disadvantage: susceptible to chaining

# Hierarchical Clustering, Single Linkage Example

Step 1: Calculate Distance Matrix

Step 2: Find closest elements

	A	B	C	D	E
A	0				
B	4	0			
C	1	4	0		
D	4	2	4	0	
E	5	5	3	4	0

Step 3: Update distance matrix

	(A,C)	B	D	E
(A,C)	0			
B	4	0		
D	4	2	0	
E	3	5	4	0

## Worked Example, cont.

Repeat from Step 2

	(A,C)	B	D	E
(A,C)	0			
B	4	0		
D	4	2	0	
E	3	5	4	0

	(A,C)	(B,D)	E
(A,C)	0		
(B,D)	4	0	
E	3	4	0

Repeat from Step 2

	(A,C)	(B,D)	E
(A,C)	0		
(B,D)	4	0	
E	3	4	0

	((A,C),E)	(B,D)
((A,C),E)	0	
(B,D)	4	0

## Worked Example, cont.

Repeat from Step 2

	((A,C),E)	(B,D)	Final Join
((A,C),E)	0		(((A,C),E),(B,D))
(B,D)	4	0	

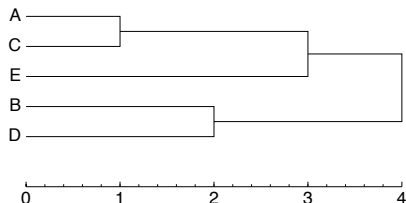


Figure: Final dendrogram for worked example

# More Hierarchical Clustering Functions

**Complete Linkage** –  $D_{ij}$  is the maximum of the  $n_i n_j$  dissimilarities between the two groups.

**Group Average Methods** –  $D_{ij}$  is the average of the  $n_i n_j$  dissimilarities between the two group (UPGMA, WPGMA)

**Centroid Method** –  $D_{ij}$  is the squared Euclidean distance between the centroids of groups  $i$  and  $j$

# K-means Clustering



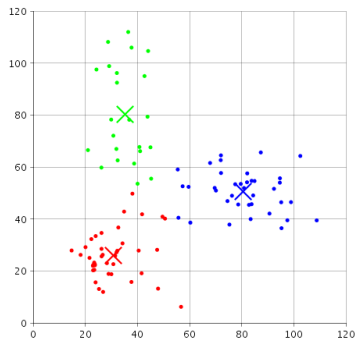
# K-mean Clustering

## General idea

Assign the  $n$  data points (or  $p$  variables) to one of  $K$  clusters to as to optimize some criterion of interest.

- The most common criterion to minimize is the sum-of-squares from the group centroids.

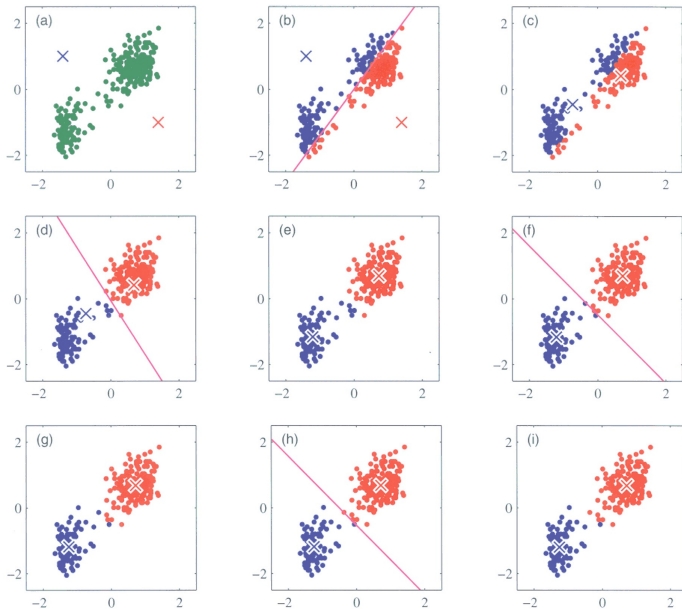
$$V = \sum_{i=1}^k \sum_{j \in g_i} |x_j - \mu_i|^2$$



# Simple algorithm for K-means clustering

- 1 Decide on  $k$ , the number of groups
- 2 Randomly pick  $k$  of the objects to act as the initial centers
- 3 Assign each object to the group whose center it is closest to
- 4 Recalculate the  $k$  centers as the centroids of the objects assigned to them
- 5 Repeat from step 3 until centroids no longer move (convergence)

# Illustration of K-means algorithm



## Things to note re: K-means clustering

- The algorithm described above does not necessarily find the global optimum
- The algorithm is sensitive to choice of initial cluster center; k-means is often run multiple-time with different initial centers to insure inferred clusters are robust.