Scientific Computing for Biologists

Lecture 9: Clustering

Instructor: Paul M. Magwene

29 October 2013

Outline of Lecture

- Distance and dissimilarity measures
  - Quantitative data
  - Dichotomous data
  - Qualitative data
- Hierarchical clustering
- K-means clustering

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 vise 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

- Euclidean distance

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

- Scaled Euclidean distance

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

  where w_k are suitable weight for the k-th variable, e.g. $\sigma_{x_k}^{-1}$ or $(max(x_k) - min(x_k))^{-1}$

- Manhattan (taxi cab, city block) distance

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

Dissimilarity Measures for Quantitative Data, cont.

- Manhattan (taxi cab, city block) distance

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

- Chebychev distance

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

- 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.

More distance measures

- Canberra distance (weighted Manhattan distance)

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

- Cosine distance

$$d_{ij} = \frac{x_{i\cdot} \cdot x_{j\cdot}}{|x_{i\cdot}||x_{j\cdot}|}$$

  where $x_{j\cdot}$ and $x_{j\cdot}$ indicate the row vectors, representing objects i and j

- Hamming distance

$$d_{ij} = \frac{\text{count}(x_{ik} = x_{jk})}{p}$$

Metric Distance Functions

A non-negative function, $g(x, y)$, is metric if:

1. $g(x, y)$ satisifies the triangle inequality:

$$g(x, y) \leq g(x, z) + g(y, z)$$

2. symmetric:

$$g(x, y) = g(y, x)$$

3. $g(x, y) = 0$ only if $x = y$

Dissimilarity Measures for Dichotomous Data

For each pair of objects (samples) of interest form a 2×2 contigency table:

	1	0
1	a	b
0	c	d

- Simple matching coefficient:

$$d_{ij} = 1 - \frac{a+d}{p} = \frac{b+c}{p}$$

- Jaccard's coefficient (ignores joint absence):

$$d_{ij} = \frac{b+c}{a+b+c}$$

- Czenkanowski coefficient:

$$d_{ij} = \frac{b+c}{2a+b+c}$$

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:

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

Introduction to Clustering

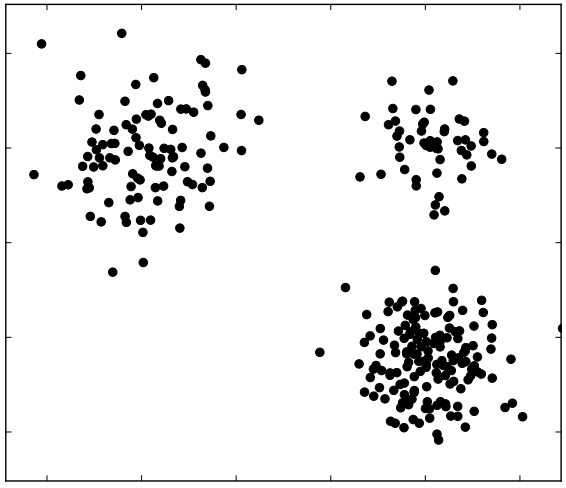
Goal of Clustering

Goal

Find “natural groups” in data

What’s a “natural group”?

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

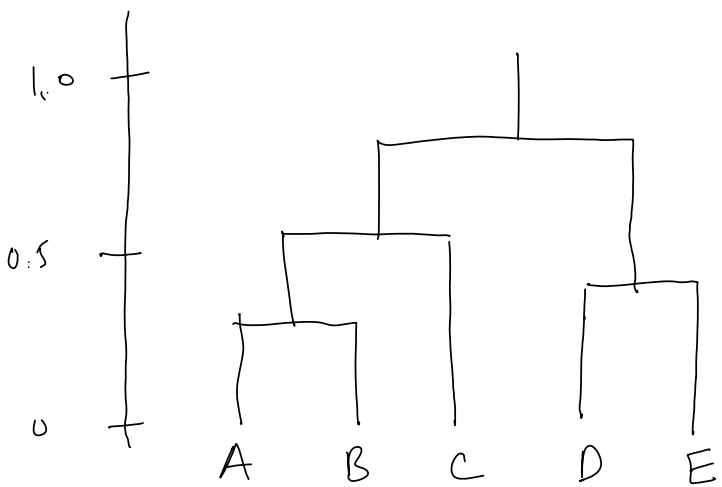


Hierarchical Clustering

Agglomerative / Divisive methods

- In practice almost always agglomerative

For n data points define a set of n-1 joins that represent groupings of objects @ different levels of similarity



Generic Algorithm for Agglomerative 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, A worked Example

	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

Single Linkage

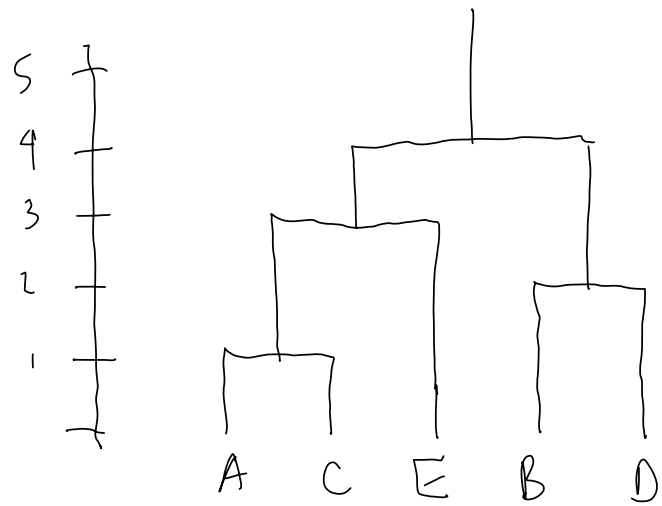
	(A,C)	B	D	E
(A,C)				
B	4			
D	4	2	0	
E	3	5	4	0

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

Worked Example, cont.

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

→ Only one Choice ((A, C), E), (B, D))



Single Linkage Clustering

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

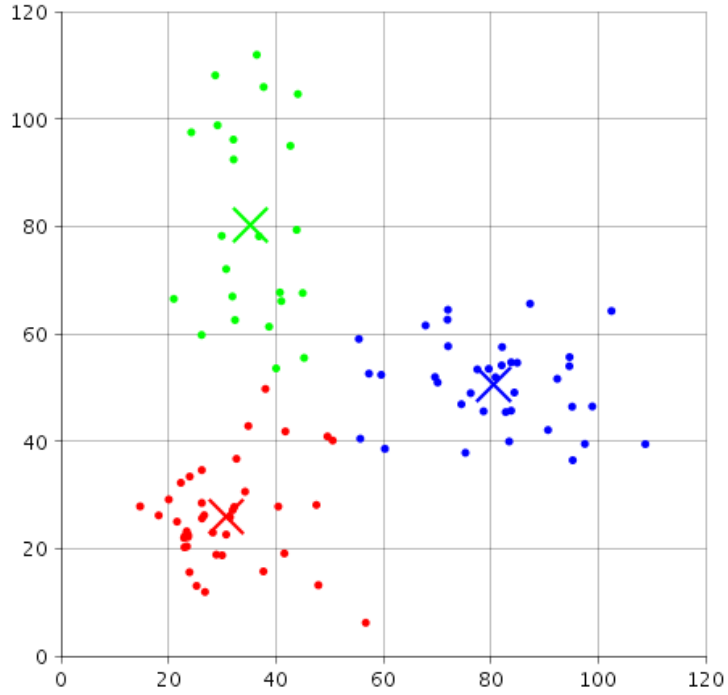
K-means Clustering

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.

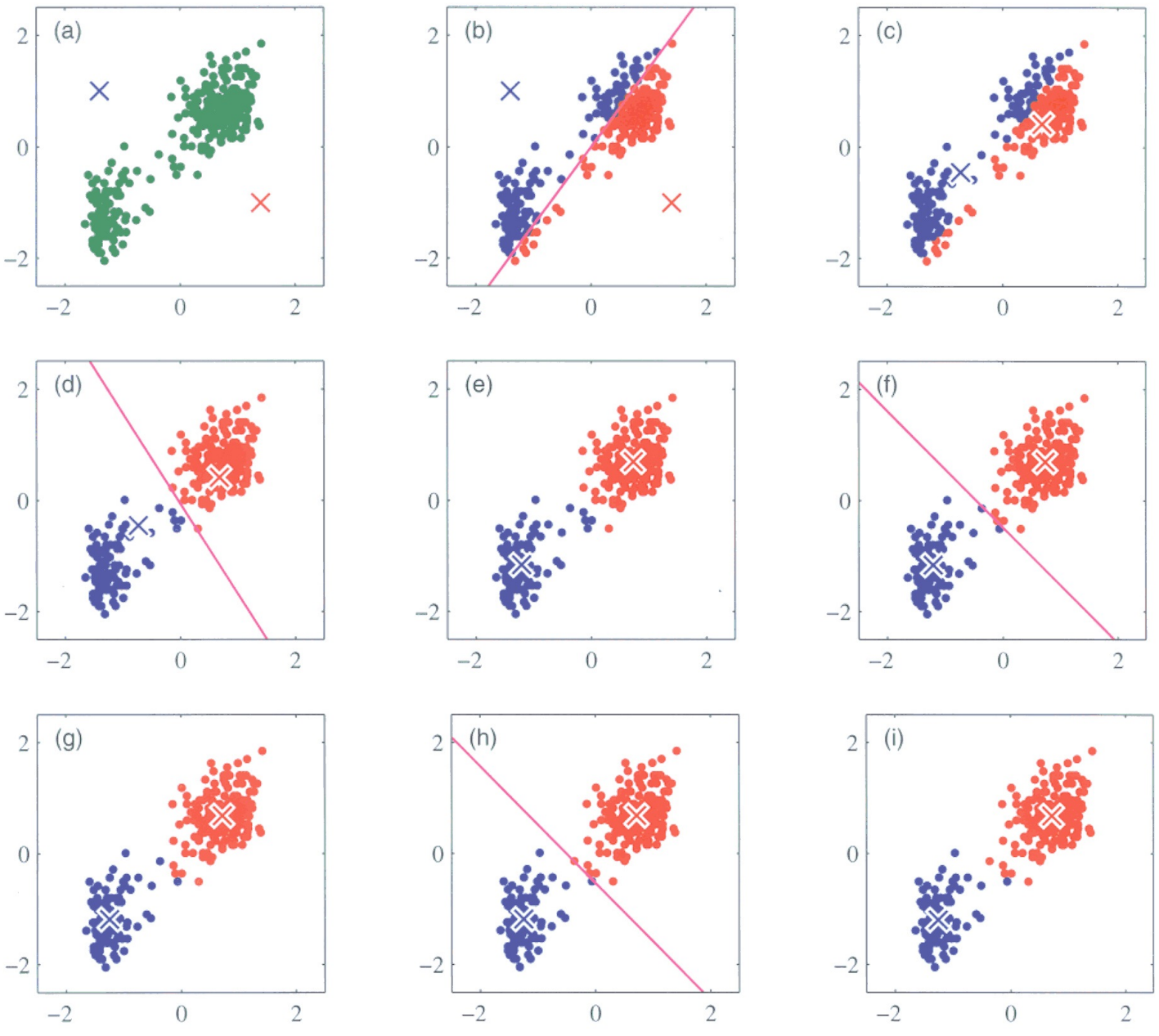
$$V = \sum_{i=1}^{k} \sum_{j \in g_i} |x_j - \mu_i|^2$$



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.