Scientific Computing for Biologists

Lecture 8: Clustering

Instructor: Paul M. Magwene

22 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:

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

- 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 $(max(x_k) - min(x_k))^{-1}$

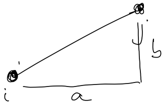
- Manhattan (taxi cab, city block) distance

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

Dissimilarity Measures for Quantitative Data

- **Euclidean Distance**

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



$$d_{ij} = \sqrt{a^2 + b^2}$$

- **Manhattan (taxi-cab) distance**

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

$$d_{ij} = a + b$$

- **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 weights

e.g. $\left(\text{std. dev of variable } k \right)^{-1}$ or $\left(\text{range of } k\text{th variable} \right)^{-1}$

Metric vs. Non-metric

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

i) Satisfies the triangle inequality:

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

ii) Symmetric:

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

iii) $g(x,y) = 0$ only if $x = y$

Euclidean Dist. is a metric function
(as is Manhattan distance)

Other Quantitative Measures of Dissimilarity

- Minkowski Metric

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

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

- Canberra Metric

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

[Accts for distance btw. points & relationship to origin]

→ only for non-negative values

- Czekanowski Coefficient

$$d_{ij} = 1 - \frac{2 \sum_{k=1}^{p} \min(x_{ik}, x_{jk})}{\sum_{k=1}^{p} (x_{ik} + x_{jk})}$$

[% dissimilarity over all variables]

Quantitative Dissimilarity for Variables

Correlation provides a suitable measure of similarity

$d_{kl} = 1 - r_{kl}$ if $r_{kl} = -1$ is taken to indicate maximum disagreement

$d_{kl} = 1 - r_{kl}^2$ is appropriate if $r_{kl} = 1$ and $r_{kl} = -1$ are treated equivalently (predictive power)

$$d_{kl} = 1 - \frac{\sum_{i=1}^{n} x_{ik} x_{il}}{\left(\sum_{i=1}^{n} x_{ik}^2 \sum_{i=1}^{n} x_{il}^2\right)} \leftarrow \text{uncentered correlation}$$

Dissimilarity for Dichotomous Data

For each pair of objects of interest form a 2×2 contingency table

		Obj 2	
		+	−
obj 1	+	a	b
	−	c	d

$a+b+c+d=p$

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

Jaccard Coefficient: $d_{ij} = \frac{b+c}{a+b+c}$ (joint absence does not contribute)

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

Introduction to Clustering

Goal of Clustering

- Find "natural groups" in data

→ one definition:

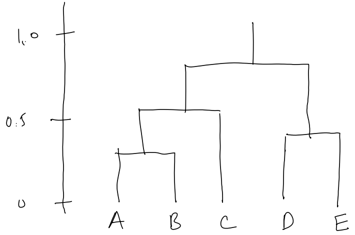
patches of high dimensity surrounded by patches of lower density in the p-dimensional space defined by the variates

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



Simple Algorithm for Hierarchical Clustering

1) Calculate a dissimilarity matrix for the n items

2) Join the two nearest items, i & j

3) Delete the i^{th} & j^{th} row and column of the dissimilarity matrix; add a new row/column
* that represents dissimilarity of new group $(i;j)$ to all other items

4) Repeat from step 2 until there is a single group

Methods of Hierarchical Clustering

The different methods are determined by the function used to determine the distance between groups

Some Common Group Distance Criteria

Single linkage (nearest neighbor)

Complete linkage (furthest neighbor)

Group average

Centroid

Single Linkage Clustering

n_i, n_j are # of objects in groups i & j

☆ D_{ij} is the <u>smallest</u> of the $n_i n_j$ dissimilarities between each element of i & each element of j

→ Invariant under monotonic transformation of the d_{ij}

→ Unaffected by ties

→ Provably nice assymptotiz properties

→ Susceptible to "chaining"

Complete Linkage

D_{ij} is the maximum of the $n_i n_j$ dissimilarities between the two groups

→ also invariant under monotonic transformation

Group average

D_{ij} is the average of the $n_i n_j$ dissimilarities between the two groups (UPGMA, WPGMA)

Centroid method

D_{ij} is the squared euclidean distance between the centroids of groups i & j

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

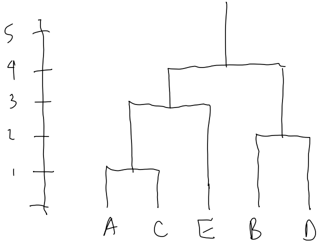
	(A,C)	(B,D)	E
(A,C)	0		
(B,D)	4	0	
E	3	4	0

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

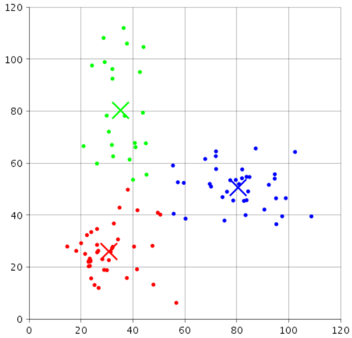
K-mean Clustering

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.

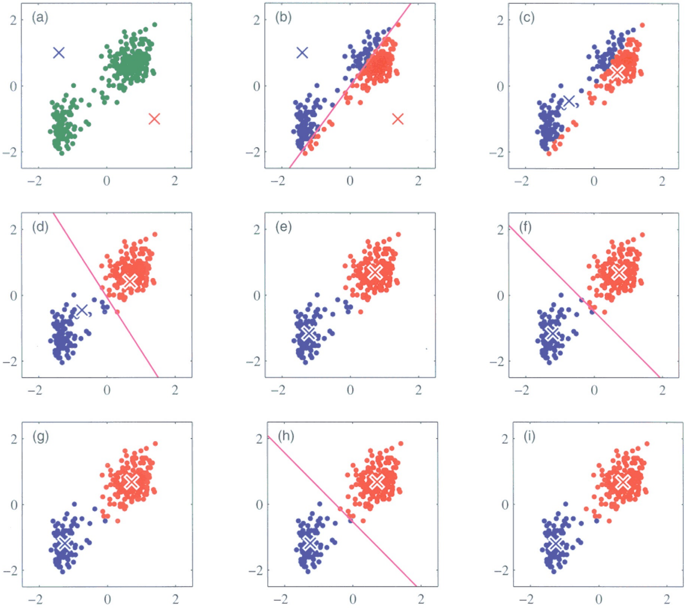
$$V = \sum_{i=1}^{k} \sum_{j \in g_i} |x_j - \mu_i|^2$$



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.