Tree Models of Similarity and Association

Clustering and Classification

Lecture 5

Today's Class

- Tree models.
- Hierarchical clustering methods.
- Fun with ultrametrics.

Preliminaries

- Today's lecture is based on the monograph by Corter (1996).
  - *Tree Models of Similarity and Association*

- I am using this monograph because I believe it treats hierarchical clustering at a level of abstraction that is general enough for our purposes.

Introduction

- Tree models use matrices of similarities (or association) to produce representations of data.
- The monograph focuses on the use of trees as models for proximity data.
  - Proximity data = similarities/associations of objects.
- The monograph also makes a distinction between ultrametric and additive trees.

Types of Data Used in Scaling

- Tree models can be applied to three general types of data:

1. Two-way one-mode data:
   - two-way = two dimensions to data matrix
   - one-mode = rows and columns represent a single set of objects.
   - Examples – ratings of similarity for a set of stimuli; correlations of items on a test.

TABLE 1.1
An Example Proximity Matrix

	Bio	*Phy*	*Law*	*Pol*	*Ins*	*Car*	*Cab*	*Con*
Biologist	—							
Physician	812	—						
Lawyer	727	714	—					
Police officer	156	175	195	—				
Insurance agent	123	156	162	221	—			
Carpenter	39	65	45	110	227	—		
Cab driver	52	78	39	149	188	396	—	
Construction worker	32	52	19	84	136	610	364	—

NOTE: Data are similarities among 8 occupations, derived from a sorting task (Kraus, 1976).

Types of Data Used in Scaling

2. Rectangular matrices of preference or choice data.
   - N rows
   - M columns
   - Entries represent preference of i^{th} person for j^{th} stimulus.
   - Called two-way two-mode data.

TABLE 1.2
An Example Rectangular Proximity Data Matrix: Preference Rankings of 12 Described Jobs by 8 Subjects

	Described Job											
Subject	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*
1	4	1	12	7	2	3	8	6	5	11	9	10
2	9	12	4	8	3	10	1	11	7	2	5	6
3	11	2	1	7	6	8	3	12	9	5	4	10
4	10	5	1	12	3	6	8	11	9	4	2	7
5	7	4	2	1	8	12	5	10	6	9	3	11
6	3	8	5	6	4	11	9	7	2	12	1	10
7	7	4	6	11	5	8	1	10	3	2	12	9
8	7	3	2	9	4	11	8	6	1	10	5	12

Types of Data Used in Scaling

3. Multivariate profile data:
   - N cases/observations.
   - M variables.
- Examples:
   - Questionnaire data
   - Frequencies of purchase of products
- Most applications here begin with the analyst taking the data and converting them into similarity or association coefficient.
   - Correlation
   - Euclidean Distance
- Resulting application to symmetric matrix (either N x N for subjects or M x M for variables).

Types of Models of Proximity Relations

- Spatial/Dimensional Models
  - Principal Components
  - Multidimensional Scaling
  - Factor Analysis
- Cluster Models
  - Ultrametric trees
  - Additive Trees
- Set-theoretic Models
  - Mathematical models of similarities
- Graph-theoretic Models
  - Minimal spanning trees

Two Types of Tree Models

Two Types of Tree Models

- Ultrametric (from wikipedia)

Formally, an ultrametric space is a set of points M with an associated distance function (also called a metric)

$d : M \times M \rightarrow \mathbf{R}$

where **R** is the set of real numbers), such that for all x, y, z in M, one has:

$d(x, y) \geq 0$

$d(x, y) = 0$ iff $x=y$

$d(x, y) = d(y, x)$ (**symmetry**)

$d(x, z) \leq \max\{d(x, y), d(y, z)\}$ (**strong triangle** or **ultrametric inequality**).

- Additive

A tree in which the distance between any two points is the sum of the lengths of the branches along the path connecting two points.

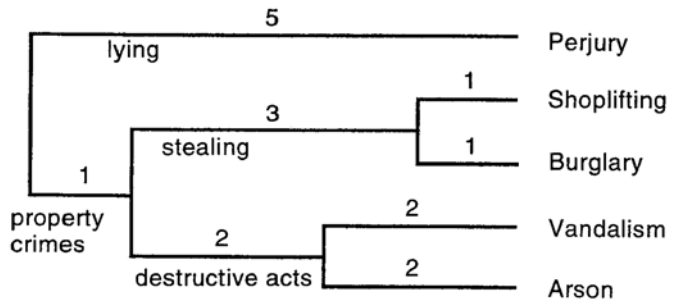
Ultrametric Trees

- Hypothetical example ultrametric shown in image.
  - Dissimilarity matrix between crimes.
  - Resulting tree diagram is output from a hierarchical clustering program.

a. Dissimilarities

	Ar	Bu	Pe	Sh	Va
Arson	--				
Burglary	8	--			
Perjury	10	10	--		
Shoplifting	8	2	10	--	
Vandalism	4	8	10	8	--

b. Ultrametric Tree



Ultrametric Trees

- Path-length distances between objects in an ultrametric tree satisfy the mathematical relationship called the ultrametric inqeuality:

$$\hat{d}(a,b) \leq \max\left[\hat{d}(a,c), \hat{d}(b,c)\right]$$

- The inequality holds under all possible re-labelings of the three points, so we can re-express the inequality as:

$$\hat{d}(x,y) \leq \hat{d}(x,z) = \hat{d}(y,z)$$

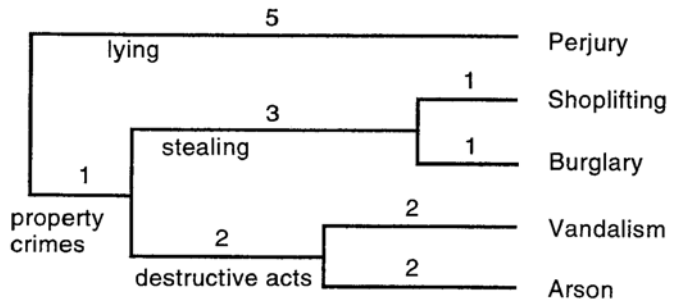
What Does This Mean?

- To demonstrate, consider the crimes burglary, shoplifting, and arson.
  - The distance from burglary to shoplifting is 2.
  - The distance from burglary to arson is 8.
  - The distance from shoplifting to arson is 8.
- These distances satisfy the ultrametric inequality.

a. Dissimilarities

	Ar	Bu	Pe	Sh	Va
Arson	--				
Burglary	8	--			
Perjury	10	10	--		
Shoplifting	8	2	10	--	
Vandalism	4	8	10	8	--

b. Ultrametric Tree



Great…But How About Real Data?

- The example we just saw was nice…but fabricated.
- Most real data will not satisfy the ultrametric inequality for all triples of points.
  - One possible reason is measurement error.
  - Minor violations can be ignored.
    - Use clustering procedure to find best fitting solution.
- If inequality is violated in certain systematic ways, an additive tree may be more appropriate.

Ultrametric Trees

- The nodes of the tree can be relabeled to represent what the clustered objects seem to represent.
- An ultrametric tree can be interpreted as a graphical representation of an additive common-features model of proximity.
- Methods for ultrametric trees will be discussed following our brief description of additive trees.

Additive Trees

- An additive tree is a less restrictive model than the ultrametric tree.

- The mathematical relationship characterizing distances in an additive tree is called the additive or tree inequality:

$$\hat{d}(a,b) + \hat{d}(c,d) \leq \max\left[\hat{d}(a,c) + \hat{d}(b,d), \hat{d}(b,c) + \hat{d}(a,d)\right]$$

- Again, this inequality can be re-expressed:

$$\hat{d}(x,y) + \hat{d}(u,v) \leq \hat{d}(x,u) + \hat{d}(y,v) = \hat{d}(x,v)\,\hat{d}(y,u)$$

Additive Tree Example

- Unlike ultrametric trees, the lengths of the leaf arcs in additive trees are free to be of any nonnegative length.
- In an ultrametric tree, all leaf nodes are equally distant from the root of the tree.

a. Dissimiliarities

	A	B	C	D	E
Worker A	--				
Worker B	15	--			
Worker C	20	25	--		
Worker D	18	23	6	--	
Worker E	20	25	20	18	--

b. Additive Tree

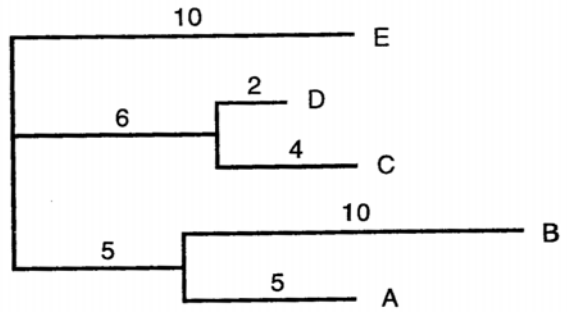
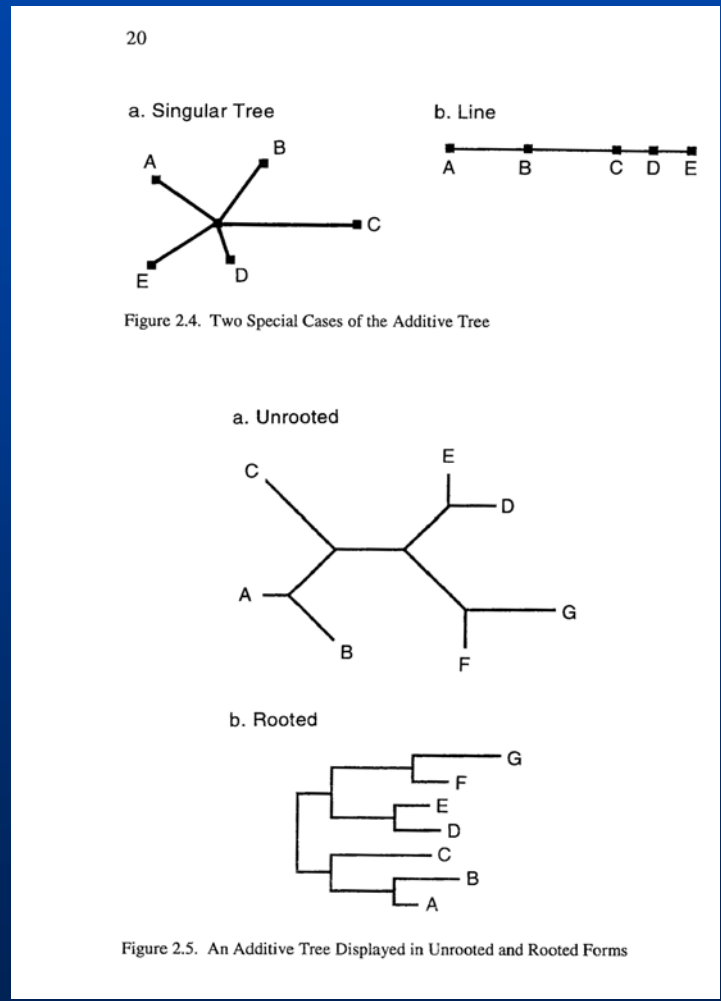


Figure 2.2. Additive Tree of Hypothetical Worker Communication Patterns

Roots of Additive Trees

- Additive trees can be displayed in either rooted or unrooted form.
  - Path length distances are the same.
- Algorithms for fitting additive trees are more difficult to construct…
  - Unlike ultrametric trees, which have routines in many stat packages.



Hierarchical Clustering Methods Using Ultrametric Trees

Hierarchical Clustering Methods

Hierarchical clustering techniques proceed by taking a set of objects and grouping a set at a time.

- Two types of hierarchical clustering methods exist:
  - Agglomerative hierarchical methods.
  - Divisive hierarchical methods.

Agglomerative Clustering Methods

- Agglomerative clustering methods start first with the individual objects.
- Initially, each object is it's own cluster.
- The ***most similar*** objects are then grouped together into a single cluster (with two objects).

We will find that what we mean by *similar* will change depending on the method

Agglomerative Clustering Methods

- The remaining steps involve merging the clusters according the similarity or dissimilarity of the objects within the cluster to those outside of the cluster.

- The method concludes when all objects are part of a single cluster.

Divisive Clustering Methods

- Divisive hierarchical methods works in the opposite direction – beginning with a single, *n*-object sized cluster.

- The large cluster is then divided into two subgroups where the objects in opposing groups are relatively distant from each other.

Divisive Clustering Methods

- The process continues similarly until there are as many clusters as there are objects.

- While it is briefly discussed in the chapter, we will not discuss divisive clustering methods in detail here in this class.

To Summarize

- So you can see that we have this idea of steps
  - At each step two clusters combine to form one (Agglomerative)

  OR…

  - At each step a cluster is divided into two new clusters (Divisive)

Methods for Viewing Clusters

- As you could imagine, when we consider the methods for hierarchical clustering, there are a large number of clusters that are formed sequentially.
- One of the most frequently used tools to view the clusters (and level at which they were formed) is the dendrogram.
- A dendrogram is a graph that describes the differing hierarchical clusters, and the distance at which each is formed.

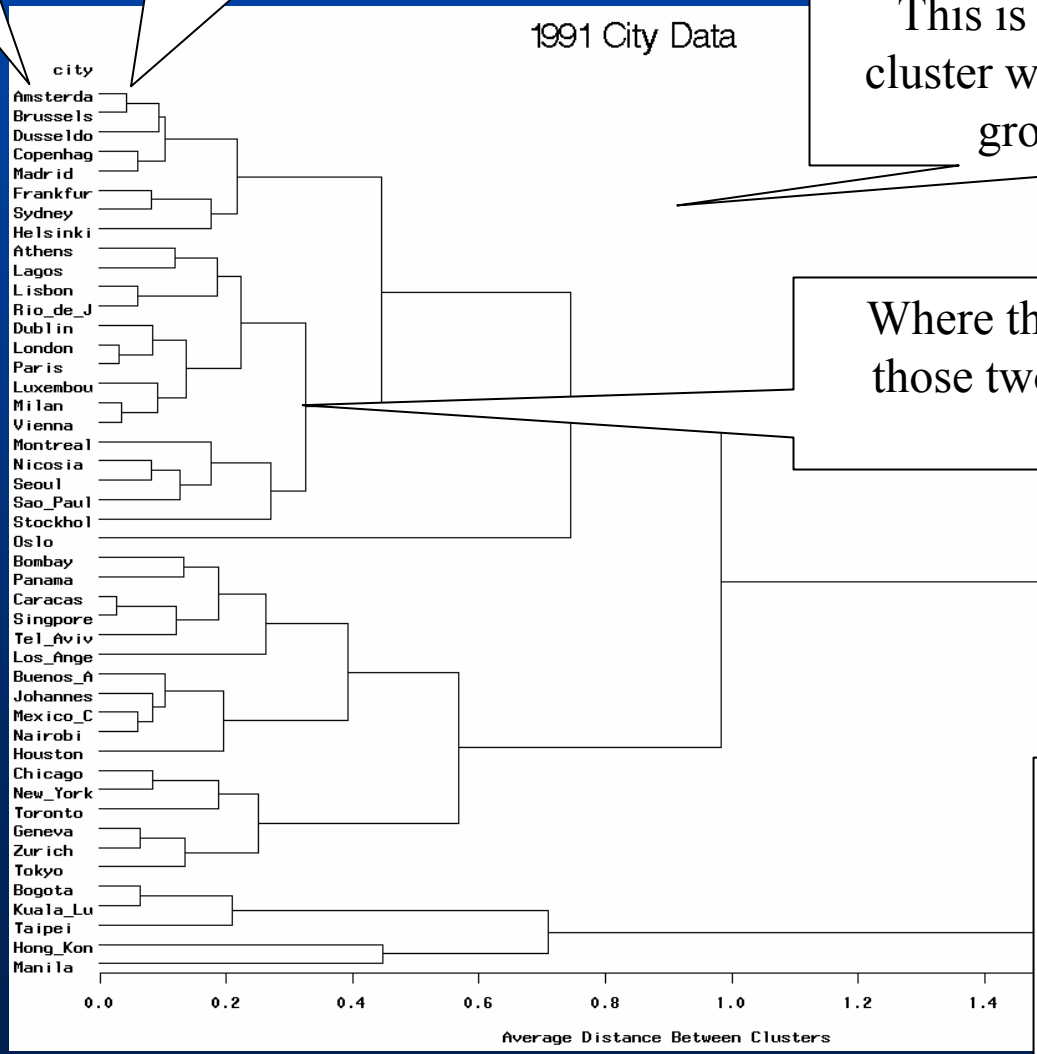
Example Data Set #1

- To demonstrate several of the hierarchical clustering methods, an example data set is used.
- Data come from a 1991 study by the economic research department of the union bank of Switzerland representing economic conditions of 48 cities around the world.
- Three variables were collected:
  - Average working hours for 12 occupations.
  - Price of 112 goods and services excluding rent.
  - Index of net hourly earnings in 12 occupations.

Dendogram

The Cities

For example, Here Amsterdam and Brussels were combined to form a group



This is an example of an agglomerate cluster where cities start off in there own group and then are combined

Where the lines connect is when those two previous groups were joined

Notice that we do not specify groups, but if we know how many we want…we simply go to the step where there are that many groups

Similarity?

- So, we mentioned that:
  - The ***most similar*** objects are then grouped together into a single cluster (with two objects).

- So the next question is how do we measure similarity between clusters.
  - More specifically, how do we redefine it when a cluster contains a combination of old clusters.

- We find that there are several ways to define similar and each way defines a new method of clustering.

Agglomerative Methods

- Next we discuss several different way to complete Agglomerative hierarchical clustering:
  - Single Linkage
  - Complete Linkage
  - Average Linkage
  - Centroid
  - Median
  - Ward Method

Agglomerative Methods

- In describing these…
  - I will give the definition of how we define similar when clusters are combined.
  - And for the first two I will give detailed examples.

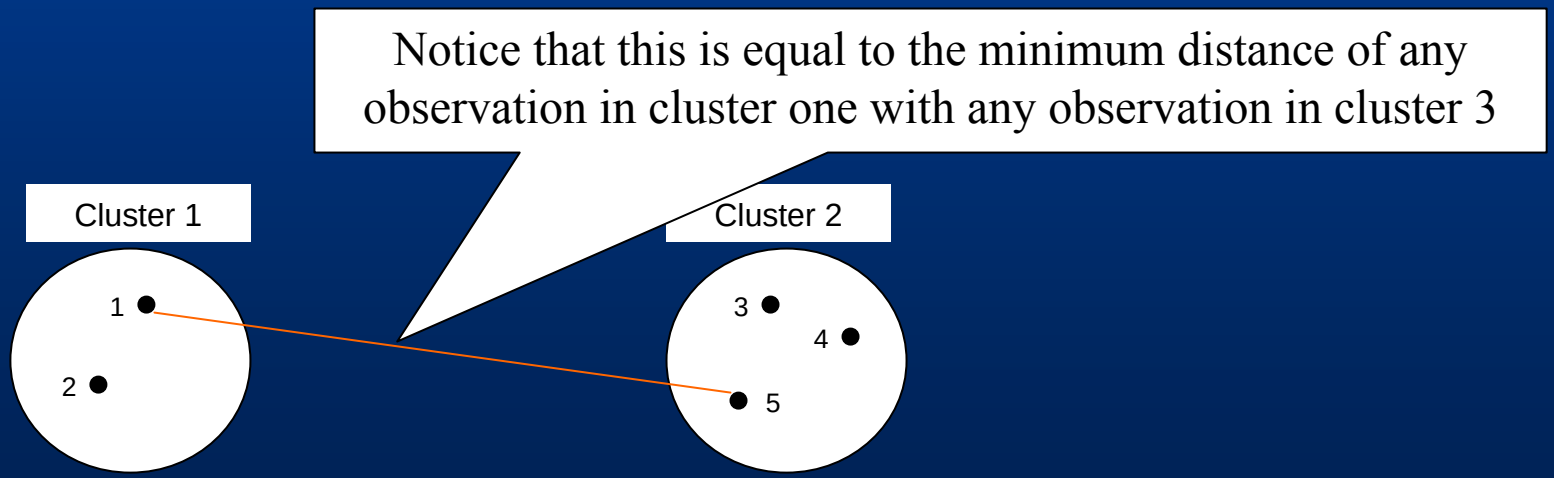
Example Distance Matrix

- The example will be based on the distance matrix below

	1	2	3	4	5
1	0	9	3	6	11
2	9	0	7	5	10
3	3	7	0	9	2
4	6	5	9	0	8
5	11	10	2	8	0

Single Linkage

- The single linkage method of clustering involves combining clusters by finding the "**nearest neighbor**" – the cluster closest to any given observation within the current cluster.



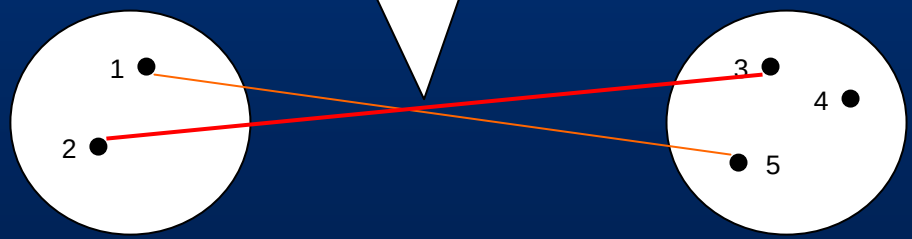
Single Linkage

- So the distance between any two clusters is:

$$D(A,B) = \min\{d(\mathbf{y}_i, \mathbf{y}_j)\}$$

For all y_i in A and y_j in B

Notice any other distance is longer



So how would we do this using our distance matrix?

Single Linkage Example

- The first step in the process is to determine the two elements with the smallest distance, and combine them into a single cluster.

- Here, the two objects that are most similar are objects 3 and 5…we will now combine these into a new cluster, and compute the distance from that cluster to the remaining clusters (objects) via the single linkage rule.

	1	2	3	4	5
1	0	9	3	6	11
2	9	0	7	5	10
3	3	7	0	9	2
4	6	5	9	0	8
5	11	10	2	8	0

Single Linkage Example

- The shaded rows/columns are the portions of the table w
  o
  5
- T
  cluster to the remain
  objects is given below:

These are the distances of 3 and 5 with 2. Our rule says that the distance of our new cluster with 2 is equal to the minimum of theses two values…7

	1	2	3	4	5
	0	9	3	6	11
	9	0	7	5	10
3	3	7	0	9	2
4	6	5	9	0	8
5	11	10	2	8	0

$d_{(35),1} = \min\{d_{31},d_{51}\} = \min\{3,11\} = 3$
$d_{(35),2} = \min\{d_{32},d_{52}\} = \min\{7,10\} = 7$
$d_{(35),4} = \min\{d_{34},d_{54}\} = \min\{9,8\} = 8$

In equation form our new distances are

Single Linkage Example

- Using the distance values, we now consolidate our table so that (35) is now a single row/column.
- The distance from the (35) cluster to the remaining objects is given below:

$d_{(35)1} = \min\{d_{31}, d_{51}\} = \min\{3, 11\} = 3$
$d_{(35)2} = \min\{d_{32}, d_{52}\} = \min\{7, 10\} = 7$
$d_{(35)4} = \min\{d_{34}, d_{54}\} = \min\{9, 8\} = 8$

	(35)	1	2	4
(35)	0			
1	3	0		
2	7	9	0	
4	8	6	5	0

Single Linkage Example

- We now repeat the process, by finding the smallest distance between within the set of remaining clusters.
- The smallest distance is between object 1 and cluster (35).
- Therefore, object 1 joins cluster (35), creating cluster (135).

	(35)	1	2	4
(35)	0	3	7	8
1	3	0	9	6
2	7	9	0	5
4	8	6	5	0

The distance from cluster (135) to the other clusters is then computed:

$$d(135)2 = \min\{d(35)2, d12\} = \min\{7,9\} = 7$$

$$d(135)4 = \min\{d(35)4, d14\} = \min\{8,6\} = 6$$

Single Linkage Example

- Using the distance values, we now consolidate our table so that (135) is now a single row/column.
- The distance from the (135) cluster to the remaining objects is given below:

	(135)	2	4
(135)	0		
2	7	0	
4	6	5	0

$$d(135)2 = \min\{d(35)2, d12\} = \min\{7,9\} = 7$$
$$d(135)4 = \min\{d(35)4, d14\} = \min\{8,6\} = 6$$

Single Linkage Example

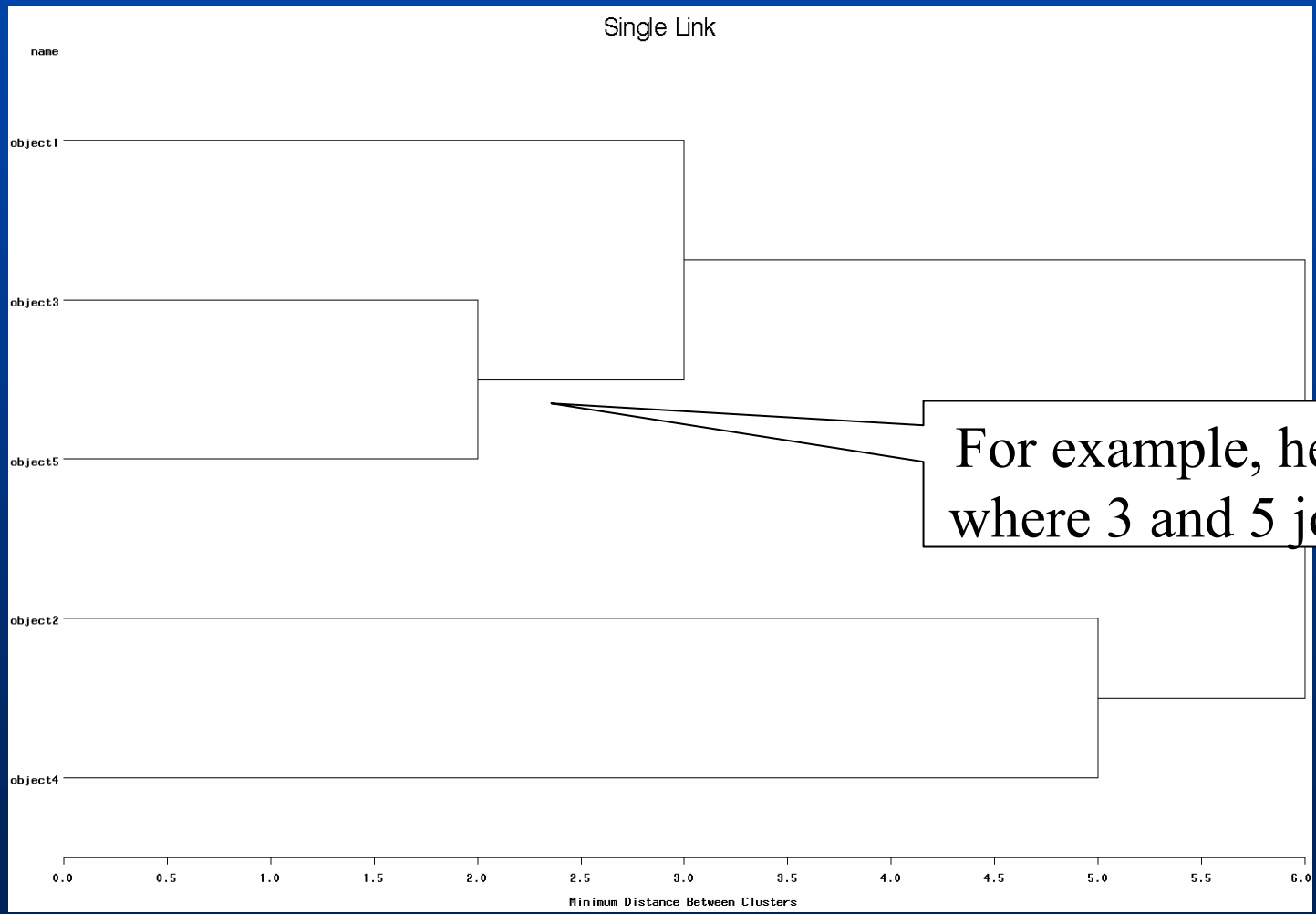
- We now repeat the process, by finding the smallest distance between within the set of remaining clusters.
- The smallest distance is between object 2 and object 4.
- These two objects will be joined to from cluster (24).
- The distance from (24) to (135) is then computed.

$d_{(135)(24)} = \min\{d_{(135)2}, d_{(135)4}\} = \min\{7,6\} = 6$

- The final cluster is formed (12345) with a distance of 6.

	(135)	2	4
(135)	0		
2	7	0	
4	6	5	0

The Dendogram



For example, here is where 3 and 5 joined

Complete Linkage

- The complete linkage method of clustering involves combining clusters by finding the “farthest neighbor” – the cluster farthest to any given observation within the current cluster.
- This ensures that all objects in a cluster are within some maximum distance of each other.

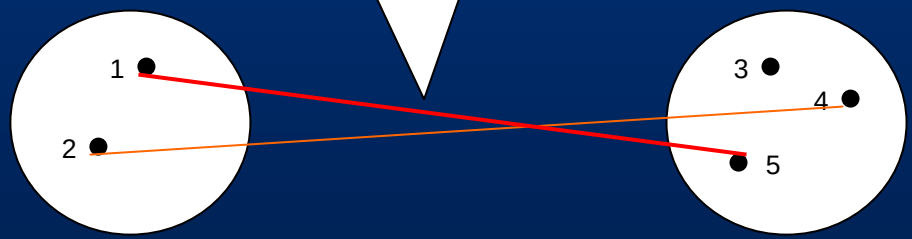
Complete Linkage

- So the distance between any two clusters is:

$$D(A,B) = \max\left\{d(\mathbf{y}_i, \mathbf{y}_j)\right\}$$

For all y_i in A and y_j in B

Notice any other distance is shorter



So how would we do this using our distance matrix?

Complete Linkage Example

- To demonstrate complete linkage in action, consider the five-object distance matrix.
- The first step in the process is to determine the two elements with the smallest distance, and combine them into a single cluster.
- Here, the two objects that are most similar are objects 3 and 5.
- We will now combine these into a new cluster, and compute the distance from that cluster to the remaining clusters (objects) via the complete linkage rule.

	1	2	3	4	5
1	0				
2	9	0			
3	3	7	0		
4	6	5	9	0	
5	11	10	2	8	0

Complete Linkage Example

- The shaded rows/columns are the portions of the table with the distances from an object to object 3 or object 5.

- The distance from the (35) cluster to the remaining objects is given below:

$$d_{(35)1} = \max\{d_{31},d_{51}\} = \max\{3,11\} = 11$$
$$d_{(35)2} = \max\{d_{32},d_{52}\} = \max\{7,10\} = 10$$
$$d_{(35)4} = \max\{d_{34},d_{54}\} = \max\{9,8\} = 9$$

	1	2	3	4	5
1	0	9	3	6	11
2	9	0	7	5	10
3	3	7	0	9	2
4	6	5	9	0	8
5	11	10	2	8	0

Complete Linkage Example

- Using the distance values, we now consolidate our table so that (35) is now a single row/column.

- The distance from the (35) cluster to the remaining objects is given below:

$d_{(35)1} = \max\{d_{31},d_{51}\} = \max\{3,11\} = 11$
$d_{(35)2} = \max\{d_{32},d_{52}\} = \max\{7,10\} = 10$
$d_{(35)4} = \max\{d_{34},d_{54}\} = \max\{9,8\} = 9$

	(35)	1	2	4
(35)	0			
1	11	0		
2	10	9	0	
4	9	6	5	0

Notice our new computed distances with (35)

Complete Linkage Example

- We now repeat the process, by finding the smallest distance between within the set of remaining clusters.
- The smallest distance is between object 2 and object 4. Therefore, they form cluster (24).
- The distance from cluster (24) to the other clusters is then computed:

$d_{(24)(135)} = \max\{d_{2(35)}, d_{4(35)}\} = \max\{10,9\} = 10$

$d_{(24)1} = \max\{d_{21}, d_{41}\} = \max\{9,6\} = 9$

	(35)	1	2	4
(35)	0	11	10	9
1	11	0	9	6
2	10	9	0	5
4	9	6	5	0

So now we use our rule to combine 2 and 4

Complete Linkage Example

- Using the distance values, we now consolidate our table so that (24) is now a single row/column.

- The distance from the (24) cluster to the remaining objects is given below:

$d_{(24)(135)} = \max\{d_{2(35)}, d_{4(35)}\} = \max\{10,9\} = 10$
$d_{(24)1} = \max\{d_{21}, d_{41}\} = \max\{9,6\} = 9$

	(35)	(24)	1
(35)	0	10	11
(24)	10	0	9
1	11	9	0

Notice our 10 and 9

Complete Linkage Example

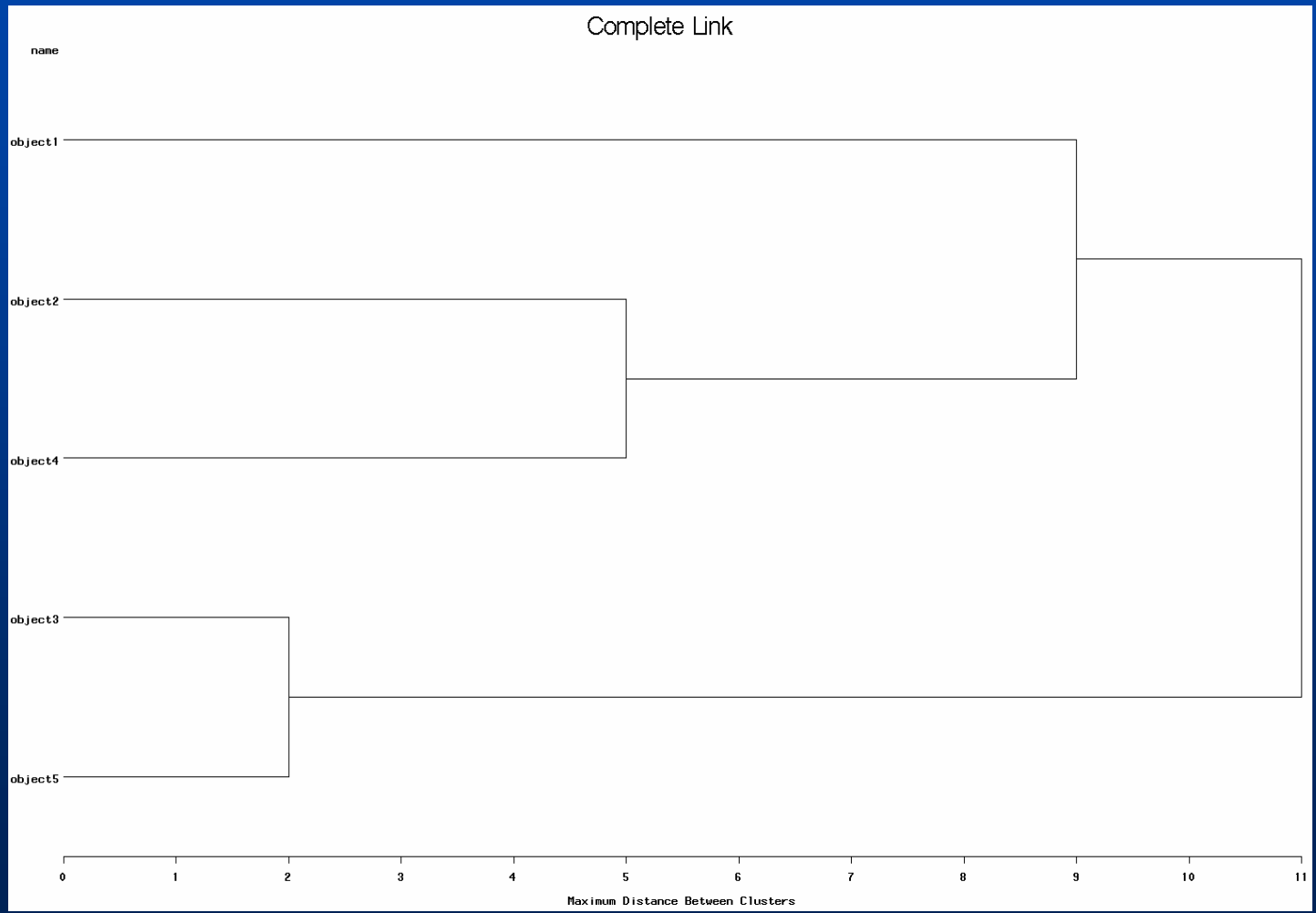
- We now repeat the process, by finding the smallest distance between within the set of remaining clusters.
- The smallest distance is between cluster (24) and object 1.
- These two objects will be joined to from cluster (124).
- The distance from (124) to (35) is then computed.

$d_{(35)(124)} = \max\{d_{1(35)}, d_{(24)(35)}\} = \max\{11,10\} = 11$

- The final cluster is formed (12345) with a distance of 11.

	(35)	(24)	1
(35)	0	10	
(24)	10	0	
1	11	9	0

The Dendogram



Average Linkage

- The average linkage method proceeds similarly to the single and complete linkage methods, with the exception that at the end of each agglomeration step, the distance between clusters is now represented by the average distance of all objects within each cluster.

- In reality, the average linkage method will produce very similar results to the complete linkage method.

So…

- To summarize…we have explained three of the methods to combine groups.
- Notice that once things are in the same group they cannot be separated
- The agglomeration method used is largely up to the user.

Wrapping Up

- Ultrametric trees are often used because of the wide-spread availability of computational routines.
- Additive trees are more general.
- What about fit?

Next Time

- Goodness of fit discussion – using multiple regression to estimate the goodness of fit of your clustering solution.
- How to do hierarchical clustering in R.
- Presentation and discussion of an empirical research article featuring hierarchical clustering.
- Punch and pie.