

Cluster Analysis

0.1 What is Cluster Analysis?

Cluster analysis is concerned with forming groups of similar objects based on several measurements of different kinds made on the objects. The key idea is to identify classifications of the objects that would be useful for the aims of the analysis. This idea has been applied in many areas including astronomy, archeology, medicine, chemistry, education, psychology, linguistics and sociology. For example, biological sciences have made extensive use of classes and sub-classes to organize species. A spectacular success of the clustering idea in chemistry was Mendeleev's periodic table of the elements. In marketing and political forecasting, clustering of neighborhoods using US postal Zip codes has been used successfully to group neighborhoods by lifestyles. Claritas, a company that pioneered this approach grouped neighborhoods into 40 clusters using various measures of consumer expenditure and demographics. Examining the clusters enabled Claritas to come up with evocative names, such as "Bohemian Mix," "Furs and Station Wagons" and "Money and Brains," for the groups that captured the dominant lifestyles in the neighborhoods. Knowledge of lifestyles can be used to estimate the potential demand for products such as sports utility vehicles and services such as pleasure cruises.

The objective of this chapter is to help you to understand the key ideas underlying the most commonly used techniques for cluster analysis and to appreciate their strengths and weaknesses. We cannot aspire to be comprehensive as there are literally hundreds of methods (there is even a journal dedicated to clustering ideas: "The Journal of Classification"!).

Typically, the basic data used to form clusters is a table of measurements on several variables where each column represents a variable and a row represents an object often referred to in statistics as a case. Thus the set of rows are to be grouped so that similar cases are in the same group. The number of groups may be specified or has to be determined from the data.

0.2 Example 1: Public Utilities Data

Table 1.1 below gives corporate data on 22 US public utilities.

We are interested in forming groups of similar utilities. The objects to be clustered are the utilities. There are 8 measurements on each utility described in Table 1.2. An example where clustering would be useful is a study to predict the cost impact of deregulation. To do the requisite analysis economists would need to build a detailed cost model of the various utilities. It would save a considerable amount of time and effort if we could cluster similar types of

utilities and to build detailed cost models for just one "typical" utility in each cluster and then scaling up from these models to estimate results for all utilities. The objects to be clustered are the utilities and there are 8 measurements on each utility.

Before we can use any technique for clustering we need to define a measure for distances between utilities so that similar utilities are a short distance apart and dissimilar ones are far from each other. A popular distance measure based on variables that take on continuous values is to standardize the values by dividing by the standard deviation (sometimes other measures such as range are used) and then to compute the distance between objects using the Euclidean metric.

The Euclidean distance d_{ij} between two cases, i and j with variable values $(x_{i1}, x_{i2}, \dots, x_{ip})$ and $(x_{j1}, x_{j2}, \dots, x_{jp})$ is defined by:

$$d_{ij} = \sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2 + \dots + (x_{ip} - x_{jp})^2}$$

All our variables are continuous in this example, so we compute distances using this metric. The result of the calculations is given in Table 1.2 below.

If we felt that some variables should be given more importance than others we would modify the squared difference terms by multiplying them by weights (positive numbers adding up to one) and use larger weights for the important variables. The Weighted Euclidean distance measure is given by:

$$d_{ij} = \sqrt{w_1(x_{i1} - x_{j1})^2 + w_2(x_{i2} - x_{j2})^2 + \dots + w_p(x_{ip} - x_{jp})^2}$$

where w_1, w_2, \dots, w_p are the weights for variables $1, 2, \dots, p$ so that $w_i \geq 0$, $\sum_{i=1}^p w_i = 1$.

0.3 Clustering Algorithms

A large number of techniques have been proposed for forming clusters from distance matrices. The most important types are hierarchical techniques, optimization techniques and mixture models. We discuss the first two types here. We will discuss mixture models in a separate note that includes their use in classification and regression as well as clustering.

0.3.1 Hierarchical Methods

There are two major types of hierarchical techniques: divisive and agglomerative. Agglomerative hierarchical techniques are the more commonly used. The

No.	Company	X1	X2	X3	X4	X5	X6	X7	X8
1	Arizona Public Service	1.06	9.2	151	54.4	1.6	9077	0	0.628
2	Boston Edison Company	0.89	10.3	202	57.9	2.2	5088	25.3	1.555
3	Central Louisiana Electric Co.	1.43	15.4	113	53	3.4	9212	0	1.058
4	Commonwealth Edison Co.	1.02	11.2	168	56	0.3	6423	34.3	0.7
5	Consolidated Edison Co. (NY)	1.49	8.8	1.92	51.2	1	3300	15.6	2.044
6	Florida Power and Light	1.32	13.5	111	60	-2.2	11127	22.5	1.241
7	Hawaiian Electric Co.	1.22	12.2	175	67.6	2.2	7642	0	1.652
8	Idaho Power Co.	1.1	9.2	245	57	3.3	13082	0	0.309
9	Kentucky Utilities Co.	1.34	13	168	60.4	7.2	8406	0	0.862
10	Madison Gas & Electric Co.	1.12	12.4	197	53	2.7	6455	39.2	0.623
11	Nevada Power Co.	0.75	7.5	173	51.5	6.5	17441	0	0.768
12	New England Electric Co.	1.13	10.9	178	62	3.7	6154	0	1.897
13	Northern States Power Co.	1.15	12.7	199	53.7	6.4	7179	50.2	0.527
14	Oklahoma Gas and Electric Co.	1.09	12	96	49.8	1.4	9673	0	0.588
15	Pacific Gas & Electric Co.	0.96	7.6	164	62.2	-0.1	6468	0.9	1.4
16	Puget Sound Power & Light Co.	1.16	9.9	252	56	9.2	15991	0	0.62
17	San Diego Gas & Electric Co.	0.76	6.4	136	61.9	9	5714	8.3	1.92
18	The Southern Co.	1.05	12.6	150	56.7	2.7	10140	0	1.108
19	Texas Utilities Co.	1.16	11.7	104	54	-2.1	13507	0	0.636
20	Wisconsin Electric Power Co.	1.2	11.8	148	59.9	3.5	7297	41.1	0.702
21	United Illuminating Co.	1.04	8.6	204	61	3.5	6650	0	2.116
22	Virginia Electric & Power Co.	1.07	9.3	1784	54.3	5.9	10093	26.6	1.306

Table 1: Public Utilities Data.

X1:	Fixed-charge covering ratio (income/debt)
X2:	Rate of return on capital
X3:	Cost per KW capacity in place
X4:	Annual Load Factor
X5:	Peak KWH demand growth from 1974 to 1975
X6:	Sales (KWH use per year)
X7:	Percent Nuclear
X8:	Total fuel costs (cents per KWH)

Table 2: Explanation of variables.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22
1	0.0	3.1	3.7	2.5	4.1	3.6	3.9	2.7	3.3	3.1	3.5	3.2	4.0	2.1	2.6	4.0	4.4	1.9	2.4	3.2	3.5	2.5
2	3.1	0.0	4.9	2.2	3.9	4.2	3.4	3.9	4.0	2.7	4.8	2.4	3.4	4.3	2.5	4.8	3.6	2.9	4.6	3.0	2.3	2.4
3	3.7	4.9	0.0	4.1	4.5	3.0	4.2	5.0	2.8	3.9	5.9	4.0	4.4	2.7	5.2	5.3	6.4	2.7	3.2	3.7	5.1	4.1
4	2.5	2.2	4.1	0.0	4.1	3.2	4.0	3.7	3.8	1.5	4.9	3.5	2.6	3.2	3.2	5.0	4.9	2.7	3.5	1.8	3.9	2.6
5	4.1	3.9	4.5	4.1	0.0	4.6	4.6	5.2	4.5	4.0	6.5	3.6	4.8	4.8	4.3	5.8	5.6	4.3	5.1	4.4	3.6	3.8
6	3.6	4.2	3.0	3.2	4.6	0.0	3.4	4.9	3.7	3.8	6.0	3.7	4.6	3.5	4.1	5.8	6.1	2.9	2.6	2.9	4.6	4.0
7	3.9	3.4	4.2	4.0	4.6	3.4	0.0	4.4	2.8	4.5	6.0	1.7	5.0	4.9	2.9	5.0	4.6	2.9	4.5	3.5	2.7	4.0
8	2.7	3.9	5.0	3.7	5.2	4.9	4.4	0.0	3.6	3.7	3.5	4.1	4.1	4.3	3.8	2.2	5.4	3.2	4.1	4.1	4.0	3.2
9	3.3	4.0	2.8	3.8	4.5	3.7	2.8	3.6	0.0	3.6	5.2	2.7	3.7	3.8	4.1	3.6	4.9	2.4	4.1	2.9	3.7	3.2
10	3.1	2.7	3.9	1.5	4.0	3.8	4.5	3.7	3.6	0.0	5.1	3.9	1.4	3.6	4.3	4.5	5.5	3.1	4.1	2.1	4.4	2.6
11	3.5	4.8	5.9	4.9	6.5	6.0	6.0	3.5	5.2	5.1	0.0	5.2	5.3	4.3	4.7	3.4	4.8	3.9	4.5	5.4	4.9	3.4
12	3.2	2.4	4.0	3.5	3.6	3.7	1.7	4.1	2.7	3.9	5.2	0.0	4.5	4.3	2.3	4.6	3.5	2.5	4.4	3.4	1.4	3.0
13	4.0	3.4	4.4	2.6	4.8	4.6	5.0	4.1	3.7	1.4	5.3	4.5	0.0	4.4	5.1	4.4	5.6	3.8	5.0	2.2	4.9	2.7
14	2.1	4.3	2.7	3.2	4.8	3.5	4.9	4.3	3.8	3.6	4.3	4.3	4.4	0.0	4.2	5.2	5.6	2.3	1.9	3.7	4.9	3.5
15	2.6	2.5	5.2	3.2	4.3	4.1	2.9	3.8	4.1	4.3	4.7	2.3	5.1	4.2	0.0	5.2	3.4	3.0	4.0	3.8	2.1	3.4
16	4.0	4.8	5.3	5.0	5.8	5.8	5.0	2.2	3.6	4.5	3.4	4.6	4.4	5.2	5.2	0.0	5.6	4.0	5.2	4.8	4.6	3.5
17	4.4	3.6	6.4	4.9	5.6	6.1	4.6	5.4	4.9	5.5	4.8	3.5	5.6	5.6	3.4	5.6	0.0	4.4	6.1	4.9	3.1	3.6
18	1.9	2.9	2.7	2.7	4.3	2.9	2.9	3.2	2.4	3.1	3.9	2.5	3.8	2.3	3.0	4.0	4.4	0.0	2.5	2.9	3.2	2.5
19	2.4	4.6	3.2	3.5	5.1	2.6	4.5	4.1	4.1	4.1	4.5	4.4	5.0	1.9	4.0	5.2	6.1	2.5	0.0	3.9	5.0	4.0
20	3.2	3.0	3.7	1.8	4.4	2.9	3.5	4.1	2.9	2.1	5.4	3.4	2.2	3.7	3.8	4.8	4.9	2.9	3.9	0.0	4.1	2.6
21	3.5	2.3	5.1	3.9	3.6	4.6	2.7	4.0	3.7	4.4	4.9	1.4	4.9	4.9	2.1	4.6	3.1	3.2	5.0	4.1	0.0	3.0
22	2.5	2.4	4.1	2.6	3.8	4.0	4.0	3.2	3.2	2.6	3.4	3.0	2.7	3.5	3.4	3.5	3.6	2.5	4.0	2.6	3.0	0.0

Table 3: Distances based on standardized variable values.

idea behind this set of techniques is to start with each cluster comprising of exactly one object and then progressively agglomerating (combining) the two nearest clusters until there is just one cluster left consisting of all the objects. Nearness of clusters is based on a measure of distance between clusters. All agglomerative methods require as input a distance measure between all the objects that are to be clustered. This measure of **distance between objects** is mapped into a metric for the **distance between clusters** (sets of objects) metrics for the distance between two clusters. The only difference between the various agglomerative techniques is the way in which this inter-cluster distance metric is defined. The most popular agglomerative techniques are:

1. **Nearest neighbor (also called single linkage).** Here the distance between two clusters is defined as the distance between the nearest pair of objects with one object in the pair belonging to a distinct cluster. If cluster A is the set of objects A_1, A_2, \dots, A_m and cluster B is B_1, B_2, \dots, B_n the single linkage distance between A and B is $Min(distance(A_i, B_j) | i = 1, 2 \dots m; j = 1, 2 \dots n)$. This method has a tendency to cluster together at an early stage objects that are distant from each other in the same cluster because of a chain of intermediate objects in the same cluster. Such clusters have elongated sausage-like shapes when visualized as objects in space.
2. **Farthest neighbor (also called complete linkage).** Here the distance between two clusters is defined as the distance between the far-

the pair of objects with one object in the pair belonging to a distinct cluster. If cluster A is the set of objects A_1, A_2, \dots, A_m and cluster B is B_1, B_2, \dots, B_n the single linkage distance between A and B is $\text{Max}(\text{distance}(A_i, B_j) | i = 1, 2 \dots m; j = 1, 2 \dots n)$. This method tends to produce clusters at the early stages that have objects that are within a narrow range of distances from each other. If we visualize them as objects in space the objects in such clusters would have a more spherical shape.

3. **Group average (also called average linkage).** Here the distance between two clusters is defined as the average distance between all possible pairs of objects with one object in each pair belonging to a distinct cluster. If cluster A is the set of objects A_1, A_2, \dots, A_m and cluster B is B_1, B_2, \dots, B_n the single linkage distance between A and B is $(1/mn)\sum \text{distance}(A_i, B_j)$ the sum being taken over $i = 1, 2 \dots m$ and $j = 1, 2 \dots n$.

Note that the results of the single linkage and the complete linkage methods depend only on the order of the inter-object distances and so are invariant to monotonic transformations of the inter-object distances.

The nearest neighbor clusters for the utilities are displayed in Figure 1 below in a useful graphic format called a dendrogram. For any given number of clusters we can determine the cases in the clusters by sliding a vertical line from left to right until the number of horizontal intersections of the vertical line equals the desired number of clusters. For example, if we wanted to form 6 clusters we would find that the clusters are:

$\{1, 18, 14, 19, 9, 10, 13, 4, 20, 2, 12, 21, 7, 15, 22, 6\}$; $\{3\}$; $\{8, 16\}$; $\{17\}$; $\{11\}$; and $\{5\}$. Notice that if we wanted 5 clusters they would be the same as for six with the exception that the first two clusters above would be merged into one cluster. In general all hierarchical methods have clusters that are nested within each other as we decrease the number of clusters we desire.

The average linkage dendrogram is shown in Figure 2. If we want six clusters using average linkage, they would be:

$\{1, 18, 14, 19, 6, 3, 9\}$; $\{2, 22, 4, 20, 10, 13\}$; $\{12, 21, 7, 15\}$; $\{17\}$; $\{5\}$; $\{8, 16, 11\}$. Notice that both methods identify $\{5\}$ and $\{17\}$ as small (“individualistic”) clusters. The clusters tend to group geographically – for example there is a southern group $\{1, 18, 14, 19, 6, 3, 9\}$, a east/west seaboard group: $\{12, 21, 7, 15\}$.

Figure1: Dendrogram - Single Linkage

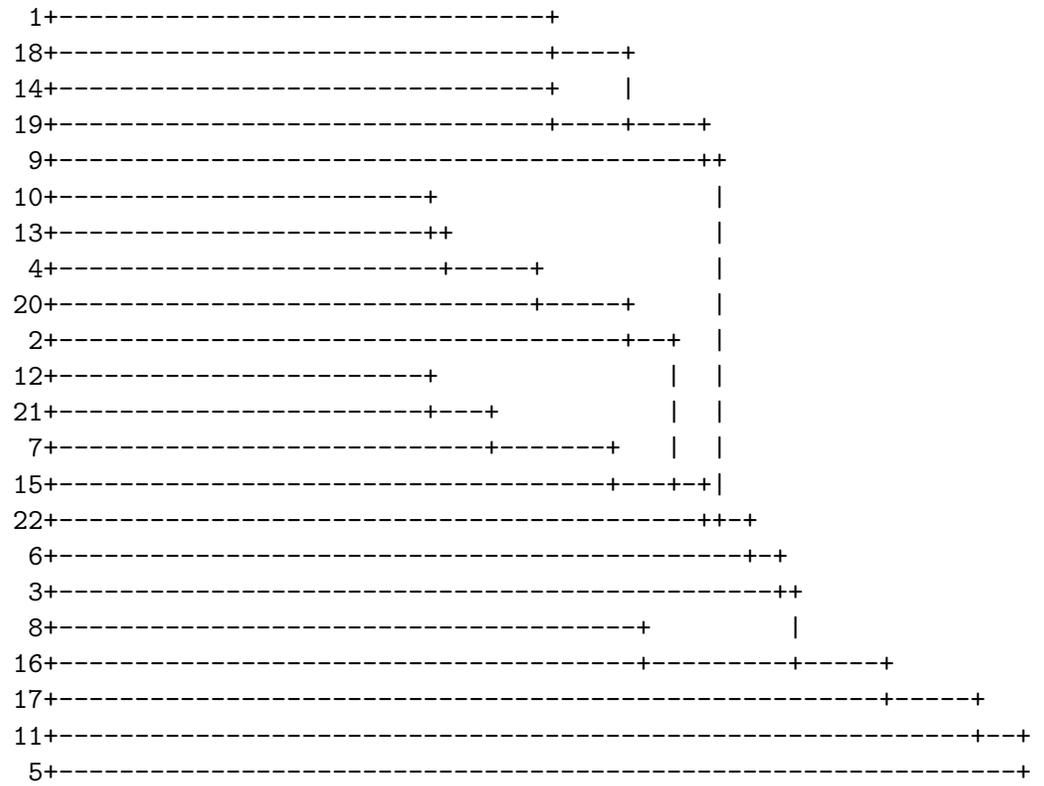
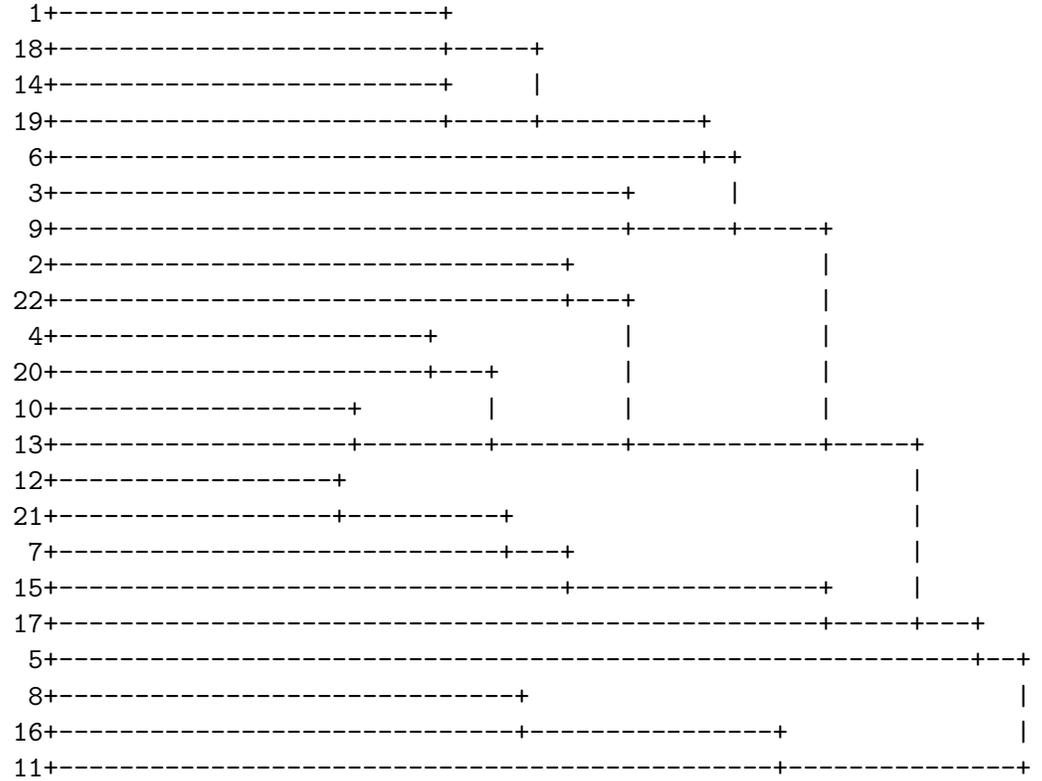


Figure2: Dendrogram - Average Linkage between groups



0.3.2 Similarity Measures

Sometimes it is more natural or convenient to work with a similarity measure between cases rather than distance which measures dissimilarity. An example is the square of the correlation coefficient, r_{ij}^2 , defined by

$$r_{ij}^2 \equiv \frac{\sum_{m=1}^p (x_{im} - \bar{x}_m)(x_{jm} - \bar{x}_m)}{\sqrt{\sum_{m=1}^p (x_{im} - \bar{x}_m)^2 \sum_{m=1}^p (x_{jm} - \bar{x}_m)^2}}$$

Such measures can always be converted to distance measures. In the above example we could define a distance measure $d_{ij} = 1 - r_{ij}^2$.

However, in the case of binary values of x it is more intuitively appealing to use similarity measures. Suppose we have binary values for all the x_{ij} 's and for individuals i and j we have the following 2×2 table:

		Individual j		
		0	1	
Individual i	0	a	b	$a + b$
	1	c	d	$c + d$
		$a + c$	$b + d$	p

The most useful similarity measures in this situation are:

- The matching coefficient, $(a + d)/p$
- Jaquard's coefficient, $d/(b + c + d)$. This coefficient ignores zero matches. This is desirable when we do not want to consider two individuals to be similar simply because they both do not have a large number of characteristics.

When the variables are mixed a similarity coefficient suggested by Gower is very useful. It is defined as

$$s_{ij} = \frac{\sum_{m=1}^p w_{ijm} s_{ijm}}{\sum_{m=1}^p w_{ijm}}$$

with $w_{ijm} = 1$ subject to the following rules:

- $w_{ijm} = 0$ when the value of the variable is not known for one of the pair of individuals or to binary variables to remove zero matches.
- For non-binary categorical variables $s_{ijm} = 0$ unless the individuals are in the same category in which case $s_{ijm} = 1$
- For continuous variables $s_{ijm} = 1 - |x_{im} - x_{jm}| / ((\max(xm) - \min(xm)))$

Other distance measures

Two useful measures of dissimilarity other than the Euclidean distance that satisfy the triangular inequality and so qualify as distance metrics are:

- Mahalanobis distance defined by

$$d_{ij} = \sqrt{(x_i - x_j)' S^{-1} (x_i - x_j)}$$

where x_i and x_j are p -dimensional vectors of the variable values for i and j respectively; and S is the covariance matrix for these vectors. This measure takes into account the correlation between the variable: variables that are highly correlated with other variables do not contribute as much as variables that are uncorrelated or mildly correlated.

- Manhattan distance defined by

$$d_{ij} = \sum_{m=1}^p |x_{im} - x_{jm}|$$

- Maximum co-ordinate distance defined by

$$d_{ij} = \max_{m=1,2,\dots,p} |x_{im} - x_{jm}|$$