We see from Figure 12.32 that the neural network and the logistic regression have very similar predictive powers and they both do better, in this case, than the classification tree. The classification tree, in turn, outperforms a random assignment. If this represented the end of the model building and assessment effort, one model would be picked (say, the neural network) to score a new set of applicants (without a credit risk designation) as Good (accept) or Bad (reject).

In the decision flow diagram in Figure 12.30, the SAMPS10.DMAGESCR file contains 75 new applicants. Expected cost scores for these applicants were created using the neural network model. Of the 75 applicants, 33 were classified as Good credit risks (with negative expected costs). ■

Data mining procedures and software continue to evolve, and it is difficult to predict what the future might bring. Database packages with embedded data mining capabilities, such as SQL Server 2005, represent one evolutionary direction.

Exercises

12.1. Certain characteristics associated with a few recent U.S. presidents are listed in Table 12.11.

Table 12.11

President	Birthplace (region of United States)	Elected first term?	Party	Prior U.S. congressional experience?	Served as vice president?
1. R. Reagan	Midwest	Yes	Republican	No	No
2. J. Carter	South	Yes	Democrat	No	No
3. G. Ford	Midwest	No	Republican	Yes	Yes
4. R. Nixon	West	Yes	Republican	Yes	Yes
5. L. Johnson	South	No	Democrat	Yes	Yes
6. J. Kennedy	East	Yes	Democrat	Yes	No

(a) Introducing appropriate binary variables, calculate similarity coefficient 1 in Table 12.1 for pairs of presidents.

Hint: You may use birthplace as South, non-South.

(b) Proceeding as in Part a, calculate similarity coefficients 2 and 3 in Table 12.1 Verify the monotonicity relation of coefficients 1, 2, and 3 by displaying the order of the 15 similarities for each coefficient.

12.2. Repeat Exercise 12.1 using similarity coefficients 5, 6, and 7 in Table 12.1.

12.3. Show that the sample correlation coefficient [see (12-11)] can be written as

$$r = \frac{ad - bc}{[(a + b)(a + c)(b + d)(c + d)]^{1/2}}$$

for two 0–1 binary variables with the following frequencies:

		Variable 2	
		0	1
Variable 1	0	a	b
	1	c	d

12.4. Show that the monotonicity property holds for the similarity coefficients 1, 2, and 3 in Table 12.1.
Hint: $(b + c) = p - (a + d)$. So, for instance,

$$\frac{a + d}{a + d + 2(b + c)} = \frac{1}{1 + 2[p/(a + d) - 1]}$$

This equation relates coefficients 3 and 1. Find analogous representations for the other pairs.

12.5. Consider the matrix of distances

$$\begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \end{array} \begin{array}{c} \begin{array}{cccc} 1 & 2 & 3 & 4 \end{array} \\ \begin{bmatrix} 0 & & & \\ 1 & 0 & & \\ 11 & 2 & 0 & \\ 5 & 3 & 4 & 0 \end{bmatrix} \end{array}$$

Cluster the four items using each of the following procedures.
(a) Single linkage hierarchical procedure.
(b) Complete linkage hierarchical procedure.
(c) Average linkage hierarchical procedure.
Draw the dendrograms and compare the results in (a), (b), and (c).

12.6. The distances between pairs of five items are as follows:

$$\begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array} \begin{array}{c} \begin{array}{ccccc} 1 & 2 & 3 & 4 & 5 \end{array} \\ \begin{bmatrix} 0 & & & & \\ 4 & 0 & & & \\ 6 & 9 & 0 & & \\ 1 & 7 & 10 & 0 & \\ 6 & 3 & 5 & 8 & 0 \end{bmatrix} \end{array}$$

Cluster the five items using the single linkage, complete linkage, and average linkage hierarchical methods. Draw the dendrograms and compare the results.

12.7. Sample correlations for five stocks were given in Example 8.5. These correlations, rounded to two decimal places, are reproduced as follows:

	JP Morgan	Citibank	Wells Fargo	Royal DutchShell	Exxon Mobil
JP Morgan	1				
Citibank	.63	1			
Wells Fargo	.51	.57	1		
Royal DutchShell	.12	.32	.18	1	
ExxonMobil	.16	.21	.15	.68	1

Treating the sample correlations as similarity measures, cluster the stocks using the single linkage and complete linkage hierarchical procedures. Draw the dendrograms and compare the results.

12.8. Using the distances in Example 12.3, cluster the items using the average linkage hierarchical procedure. Draw the dendrogram. Compare the results with those in Examples 12.3 and 12.5.

12.9. The vocabulary "richness" of a text can be quantitatively described by counting the words used once, the words used twice, and so forth. Based on these counts, a linguist proposed the following distances between chapters of the Old Testament book Lamentations (data courtesy of Y. T. Radday and M. A. Pollatschek):

		Lamentations chapter				
		1	2	3	4	5
Lamentations chapter	1	0				
	2	.76	0			
	3	2.97	.80	0		
	4	4.88	4.17	.21	0	
	5	3.86	1.92	1.51	.51	0

Cluster the chapters of Lamentations using the three linkage hierarchical methods we have discussed. Draw the dendrograms and compare the results.

12.10. Use Ward's method to cluster the four items whose measurements on a single variable X are given in the following table.

	Measurements
Item	x
1	2
2	1
3	5
4	8

(a) Initially, each item is a cluster and we have the clusters

$$\{1\} \quad \{2\} \quad \{3\} \quad \{4\}$$

Show that ESS = 0, as it must.

(b) If we join clusters $\{1\}$ and $\{2\}$, the new cluster $\{12\}$ has

$$\text{ESS}_1 = \sum (x_j - \bar{x})^2 = (2 - 1.5)^2 + (1 - 1.5)^2 = .5$$

and the ESS associated with the grouping $\{12\}$, $\{3\}$, $\{4\}$ is ESS = .5 + 0 + 0 = .5. The *increase* in ESS (loss of information) from the first step to the current step in .5 − 0 = .5. Complete the following table by determining the increase in ESS for all the possibilities at step 2.

Clusters			Increase in ESS
{12}	{3}	{4}	.5
{13}	{2}	{4}	
{14}	{2}	{3}	
{1}	{23}	{4}	
{1}	{24}	{3}	
{1}	{2}	{34}	

(c) Complete the last two algamation steps, and construct the dendrogram showing the values of ESS at which the mergers take place.

12.11. Suppose we measure two variables X_1 and X_2 for four items A, B, C, and D. The data are as follows:

	Observations	
Item	x_1	x_2
A	5	4
B	1	−2
C	−1	1
D	3	1

Use the K-means clustering technique to divide the items into $K = 2$ clusters. Start with the initial groups (AB) and (CD).

12.12. Repeat Example 12.11, starting with the initial groups (AC) and (BD). Compare your solution with the solution in the example. Are they the same? Graph the items in terms of their (x_1, x_2) coordinates, and comment on the solutions.

12.13. Repeat Example 12.11, but start at the bottom of the list of items, and proceed up in the order D, C, B, A. Begin with the initial groups (AB) and (CD). [The first potential reassignment will be based on the distances $d^2(D, (AB))$ and $d^2(D, (CD))$.] Compare your solution with the solution in the example. Are they the same? Should they be the same?

The following exercises require the use of a computer.

12.14. Table 11.9 lists measurements on 8 variables for 43 breakfast cereals.

(a) Using the data in the table, calculate the Euclidean distances between pairs of cereal brands.

(b) Treating the distances calculated in (a) as measures of (dis)similarity, cluster the cereals using the single linkage and complete linkage hierarchical procedures. Construct dendrograms and compare the results.

12.15. Input the data in Table 11.9 into a K-means clustering program. Cluster the cereals into $K = 2, 3$, and 4 groups. Compare the results with those in Exercise 12.14.

12.16. The national track records data for women are given in Table 1.9.

(a) Using the data in Table 1.9, calculate the Euclidean distances between pairs of countries.

(b) Treating the distances in (a) as measures of (dis)similarity, cluster the countries using the single linkage and complete linkage hierarchical procedures. Construct dendrograms and compare the results.

(c) Input the data in Table 1.9 into a K-means clustering program. Cluster the countries into groups using several values of K. Compare the results with those in Part b.

12.17. Repeat Exercise 12.16 using the national track records data for men given in Table 8.6. Compare the results with those of Exercise 12.16. Explain any differences.

12.18. Table 12.12 gives the road distances between 12 Wisconsin cities and cities in neighboring states. Locate the cities in $q = 1, 2$, and 3 dimensions using multidimensional scaling. Plot the minimum stress (q) versus q and interpret the graph. Compare the two-dimensional multidimensional scaling configuration with the locations of the cities on a map from an atlas.

12.19. Table 12.13 on page 752 gives the "distances" between certain archaeological sites from different periods, based upon the frequencies of different types of potsherds found at the sites. Given these distances, determine the coordinates of the sites in $q = 3, 4$, and 5 dimensions using multidimensional scaling. Plot the minimum stress (q) versus q

Table 12.12 Distances Between Cities in Wisconsin and Cities in Neighboring States

	Appleton (1)	Beloit (2)	Fort Atkinson (3)	Madison (4)	Marshfield (5)	Milwaukee (6)	Monroe (7)	Superior (8)	Wausau (9)	Dubuque (10)	St. Paul (11)	Chicago (12)
(1)	0											
(2)	130	0										
(3)	98	33	0									
(4)	102	50	36	0								
(5)	103	185	164	138	0							
(6)	100	73	54	77	184	0						
(7)	149	33	58	47	170	107	0					
(8)	315	377	359	330	219	394	362	0				
(9)	91	186	166	139	45	181	186	223	0			
(10)	196	94	119	95	186	168	61	351	215	0		
(11)	257	304	287	258	161	322	289	162	175	274	0	
(12)	186	97	113	146	276	93	130	467	275	184	395	0

Table 12.13 Distances Between Archaeological Sites

	P1980918 (1)	P1931131 (2)	P1550960 (3)	P1530987 (4)	P1361024 (5)	P1351005 (6)	P1340945 (7)	P1311137 (8)	P1301062 (9)
(1)	0	.	.	.	.	.	.	.	.
(2)	2.202	0	.	.	.	.	.	.	.
(3)	1.004	2.025	0	.	.	.	.	.	.
(4)	1.108	1.943	0.233	0	.	.	.	.	.
(5)	1.122	1.870	0.719	0.541	0	.	.	.	.
(6)	0.914	2.070	0.719	0.679	0.539	0	.	.	.
(7)	0.914	2.186	0.452	0.681	1.102	0.916	0	.	.
(8)	2.056	2.055	1.986	1.990	1.963	2.056	2.027	0	.
(9)	1.608	1.722	1.358	1.168	0.681	1.005	1.719	1.991	0

KEY: P1980918 refers to site P198 dated A.D. 0918, P1931131 refers to site P193 dated A.D. 1131, and so forth.
Source: Data Courtesy of M. J. Tretter.

and interpret the graph. If possible, locate the sites in two dimensions (the first two principal components) using the coordinates for the $q = 5$-dimensional solution. (Treat the sites as variables.) Noting the periods associated with the sites, interpret the two-dimensional configuration.

12.20. A sample of $n = 1660$ people is cross-classified according to mental health status and socioeconomic status in Table 12.14.

Perform a correspondence analysis of these data. Interpret the results. Can the associations in the data be well represented in one dimension?

12.21. A sample of 901 individuals was cross-classified according to three categories of income and four categories of job satisfaction. The results are given in Table 12.15.

Perform a correspondence analysis of these data. Interpret the results.

12.22. Perform a correspondence analysis of the data on forests listed in Table 12.10, and verify Figure 12.28 given in Example 12.22.

12.23. Construct a biplot of the pottery data in Table 12.8. Interpret the biplot. Is the biplot consistent with the correspondence analysis plot in Figure 12.22? Discuss your answer. (Use the row proportions as a vector of observations at a site.)

12.24. Construct a biplot of the mental health and socioeconomic data in Table 12.14. Interpret the biplot. Is the biplot consistent with the correspondence analysis plot in Exercise 12.20? Discuss your answer. (Use the column proportions as the vector of observations for each status.)

Table 12.14 Mental Health Status and Socioeconomic Status Data

Mental Health Status	Parental Socioeconomic Status				
	A (High)	B	C	D	E (Low)
Well	121	57	72	36	21
Mild symptom formation	188	105	141	97	71
Moderate symptom formation	112	65	77	54	54
Impaired	86	60	94	78	71

Source: Adapted from data in Srole, L., T. S. Langner, S. T. Michael, P. Kirkpatrick, M. K. Opler, and T. A. C. Rennie, *Mental Health in the Metropolis: The Midtown Manhatten Study*, rev. ed. (New York: NYU Press, 1978).

Table 12.15 Income and Job Satisfaction Data

Income	Job Satisfaction			
	Very dissatisfied	Somewhat dissatisfied	Moderately satisfied	Very satisfied
< $25,000	42	62	184	207
$25,000–$50,000	13	28	81	113
> $50,000	7	18	54	92

Source: Adapted from data in Table 8.2 in Agresti, A., *Categorical Data Analysis* (New York: John Wiley, 1990).

12.25. Using the archaeological data in Table 12.13, determine the two-dimensional metric and nonmetric multidimensional scaling plots. (See Exercise 12.19.) Given the coordinates of the points in each of these plots, perform a Procrustes analysis. Interpret the results.

12.26. Table 8.7 contains the Mali family farm data (see Exercise 8.28). Remove the outliers 25, 34, 69 and 72, leaving at total of $n = 72$ observations in the data set. Treating the Euclidean distances between pairs of farms as a measure of similarity, cluster the farms using average linkage and Ward's method. Construct the dendrograms and compare the results. Do there appear to be several distinct clusters of farms?

12.27. Repeat Exercise 12.26 using standardized observations. Does it make a difference whether standardized or unstandardized observations are used? Explain.

12.28. Using the Mali family farm data in Table 8.7 with the outliers 25, 34, 69 and 72 removed, cluster the farms with the K-means clustering algorithm for $K = 5$ and $K = 6$. Compare the results with those in Exercise 12.26. Is 5 or 6 about the right number of distinct clusters? Discuss.

12.29. Repeat Exercise 12.28 using standardized observations. Does it make a difference whether standardized of unstandardized observations are used? Explain.

12.30. A company wants to do a mail marketing campaign. It costs the company \$1 for each item mailed. They have information on 100,000 customers. Create and interpret a cumulative lift chart from the following information.

Overall Response Rate: Assume we have no model other than the prediction of the overall response rate which is 20%. That is, if all 100,000 customers are contacted (at a cost of \$100,000), we will receive around 20,000 positive responses.

Results of Response Model: A response model predicts who will respond to a marketing campaign. We use the response model to assign a score to all 100,000 customers and predict the positive responses from contacting only the top 10,000 customers, the top 20,000 customers, and so forth. The model predictions are summarized below.

Cost ($)	Total Customers Contacted	Positive Responses
10000	10000	6000
20000	20000	10000
30000	30000	13000
40000	40000	15800
50000	50000	17000
60000	60000	18000
70000	70000	18800
80000	80000	19400
90000	90000	19800
100000	100000	20000

12.31. Consider the crude-oil data in Table 11.7. Transform the data as in Example 11.14. Ignore the known group membership. Using the special purpose software *MCLUST*,

(a) select a mixture model using the BIC criterion allowing for the different covariance structures listed in Section 12.5 and up to $K = 7$ groups.

(b) compare the clustering results for the best model with the known classifications given in Example 11.14. Notice how several clusters correspond to one crude-oil classification.