Questions

1. Discuss whether or not each of the following activities is a data mining task.

(a) Dividing the customers of a company according to their gender.

(b) Dividing the customers of a company according to their profitability.

(c) Computing the total sales of a company.

(d) Sorting a student database based on student identification numbers.

(e) Predicting the outcomes of tossing a (fair) pair of dice.

(f) Predicting the future stock price of a company using historical records.

1. Discuss whether or not each of the following activities is a data mining task.

(g)Monitoring the heart rate of a patient for abnormalities.

(h) Monitoring seismic waves for earthquake activities.

(i) Extracting the frequencies of a sound wave.

2. Which of the following quantities is likely to show more temporal autocorrelation:

- daily rainfall or daily temperature?
- Why?

QUESTION 3

Table 5.1. Data set for Exercise 7.

Record	A	B	C	Class
1	0	0	0	+
2	0	0	1	−
3	0	1	1	−
4	0	1	1	−
5	0	0	1	+
6	1	0	1	+
7	1	0	1	−
8	1	0	1	−
9	1	1	1	+
10	1	0	1	+

QUESTION 3

(a) Estimate the conditional probabilities for P(A|+), P(B|+), P(C|+), P(A|–), P(B|–), and P(C|–).

(b) Use the estimate of conditional probabilities given in the previous question to predict the class label for a test sample (A = 0,B = 1, C = 0) using the naïve Bayes approach.

(c)Estimate the conditional probabilities using the m-estimate approach, with p = 1/2 and m = 4

(d) Repeat part (b) using the conditional probabilities given in part (c).

Answer a)

P(A = 1|−) = 2/5 = 0.4, P(B = 1|−) = 2/5 = 0.4,
P(C = 1|−) = 1, P(A = 0|−) = 3/5 = 0.6,
P(B = 0|−) = 3/5 = 0.6, P(C = 0|−) = 0; P(A = 1|+) = 3/5 = 0.6,
P(B = 1|+) = 1/5 = 0.2, P(C = 1|+) = 2/5 = 0.4,
P(A = 0|+) = 2/5 = 0.4, P(B = 0|+) = 4/5 = 0.8,
P(C = 0|+) = 3/5 = 0.6.

Answer b

Let $P(A = 0, B = 1, C = 0) = K$.

$$
\begin{aligned}
& P(+|A = 0, B = 1, C = 0) \\
= \quad & \frac{P(A = 0, B = 1, C = 0|+) \times P(+)}{P(A = 0, B = 1, C = 0)} \\
= \quad & \frac{P(A = 0|+)P(B = 1|+)P(C = 0|+) \times P(+)}{K} \\
= \quad & 0.4 \times 0.2 \times 0.6 \times 0.5/K \\
= \quad & 0.024/K.
\end{aligned}
$$

$$
\begin{aligned}
& P(-|A = 0, B = 1, C = 0) \\
= \quad & \frac{P(A = 0, B = 1, C = 0|-) \times P(-)}{P(A = 0, B = 1, C = 0)} \\
= \quad & \frac{P(A = 0|-) \times P(B = 1|-) \times P(C = 0|-) \times P(-)}{K} \\
= \quad & 0/K
\end{aligned}
$$

The class label should be '+'.

Answer c

$P(A = 0|+) = (2 + 2)/(5 + 4) = 4/9,$
$P(A = 0|-) = (3+2)/(5 + 4) = 5/9,$
$P(B = 1|+) = (1 + 2)/(5 + 4) = 3/9,$
$P(B = 1|-) = (2+2)/(5 + 4) = 4/9,$
$P(C = 0|+) = (3 + 2)/(5 + 4) = 5/9,$
$P(C = 0|-) = (0+2)/(5 + 4) = 2/9.$

Answer c

Let P(A = 0,B = 1, C = 0) = K

$$
\begin{aligned}
& P(+|A=0,B=1,C=0) \\
= & \frac{P(A=0,B=1,C=0|+) \times P(+)}{P(A=0,B=1,C=0)} \\
= & \frac{P(A=0|+)P(B=1|+)P(C=0|+) \times P(+)}{K} \\
= & \frac{(4/9) \times (3/9) \times (5/9) \times 0.5}{K} \\
= & 0.0412/K
\end{aligned}
$$

$$
\begin{aligned}
& P(-|A=0,B=1,C=0) \\
= & \frac{P(A=0,B=1,C=0|-) \times P(-)}{P(A=0,B=1,C=0)} \\
= & \frac{P(A=0|-) \times P(B=1|-) \times P(C=0|-) \times P(-)}{K} \\
= & \frac{(5/9) \times (4/9) \times (2/9) \times 0.5}{K} \\
= & 0.0274/K
\end{aligned}
$$

The class label should be '+'.

QUESTION 4

Consider the one-dimensional data set shown in Table 5.4.

Table 5.4. Data set for Exercise 13.

x	0.5	3.0	4.5	4.6	4.9	5.2	5.3	5.5	7.0	9.5
y	−	−	+	+	+	−	−	+	−	−

(a) Classify the data point x = 5.0 according to its 1-, 3-, 5-, and 9-nearest neighbors (using majority vote).
(b) Repeat the previous analysis using the distance-weighted voting approach described in Section 5.2.1.

ANSWERS

Answer A
1-nearest neighbor: +,
3-nearest neighbor: −,
5-nearest neighbor: +,
9-nearest neighbor: −.

ANSWER B.
1-nearest neighbor: +,
3-nearest neighbor: +,
5-nearest neighbor: +,
9-nearest neighbor: +.

QUESTION

Given the lattice structure shown in Figure 6.4 and the transactions given in Table 6.3, label each node with the following letter(s):

- M if the node is a maximal frequent itemset,
- C if it is a closed frequent itemset,
- N if it is frequent but neither maximal nor closed, and
- I if it is infrequent.

Assume that the support threshold is equal to 30%.

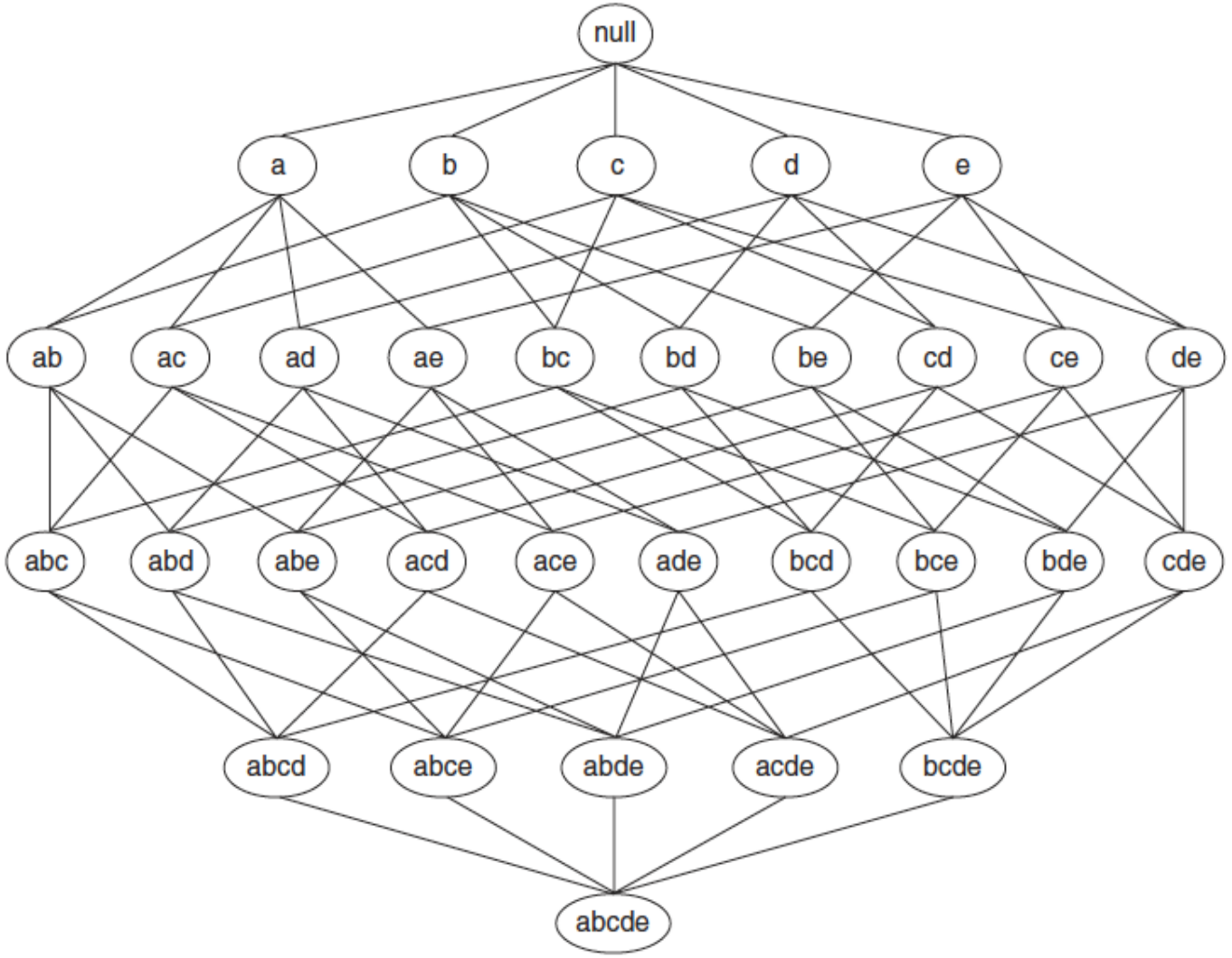


Figure 6.4. An itemset lattice

Table 6.3. Example of market basket transactions.

Transaction ID	Items Bought
1	$\{a, b, d, e\}$
2	$\{b, c, d\}$
3	$\{a, b, d, e\}$
4	$\{a, c, d, e\}$
5	$\{b, c, d, e\}$
6	$\{b, d, e\}$
7	$\{c, d\}$
8	$\{a, b, c\}$
9	$\{a, d, e\}$
10	$\{b, d\}$

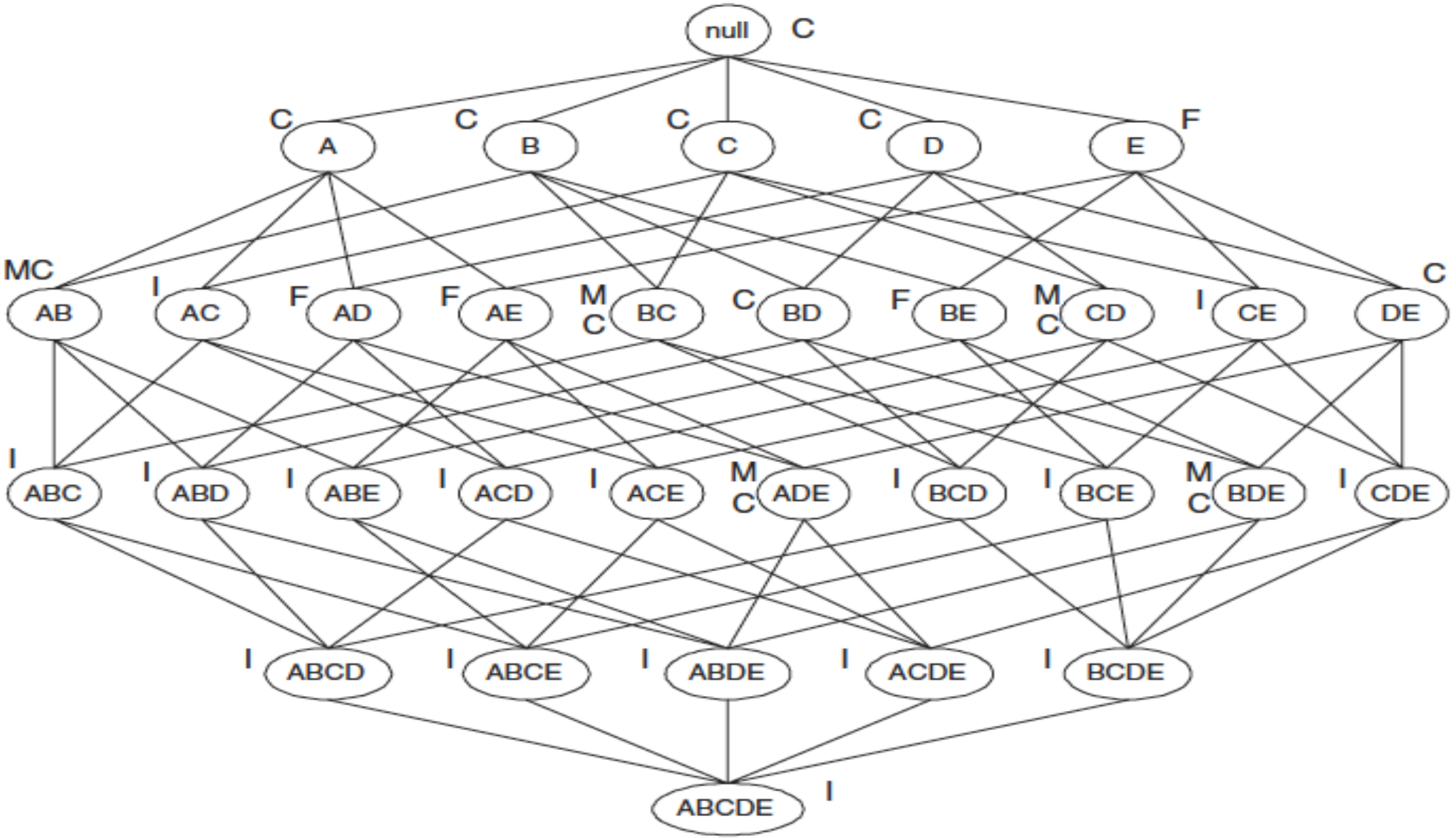


Figure 6.5. Solution for Exercise 11.

QUESTION

1. **Describe how you would create visualizations to display information that**

describes the following types of systems.Be sure to address the following issues:

- **Representation. How will you map objects, attributes, and relationships to visual elements?**
- **Arrangement. Are there any special considerations that need to be taken into account with respect to how visual elements are displayed? Specific examples might be the choice of viewpoint, the use of transparency, or the separation of certain groups of objects.**
- **Selection. How will you handle a large number of attributes and data objects?**

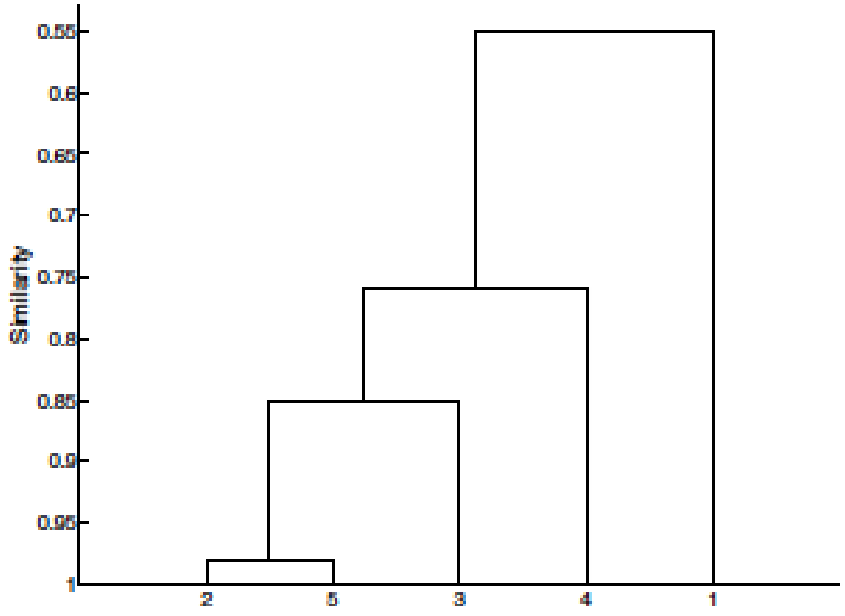
QUESTION

Use the similarity matrix in Table 8.1 to perform single and complete link hierarchical clustering. Show your results by drawing a dendrogram. The dendrogram should clearly show the order in which the points are merge

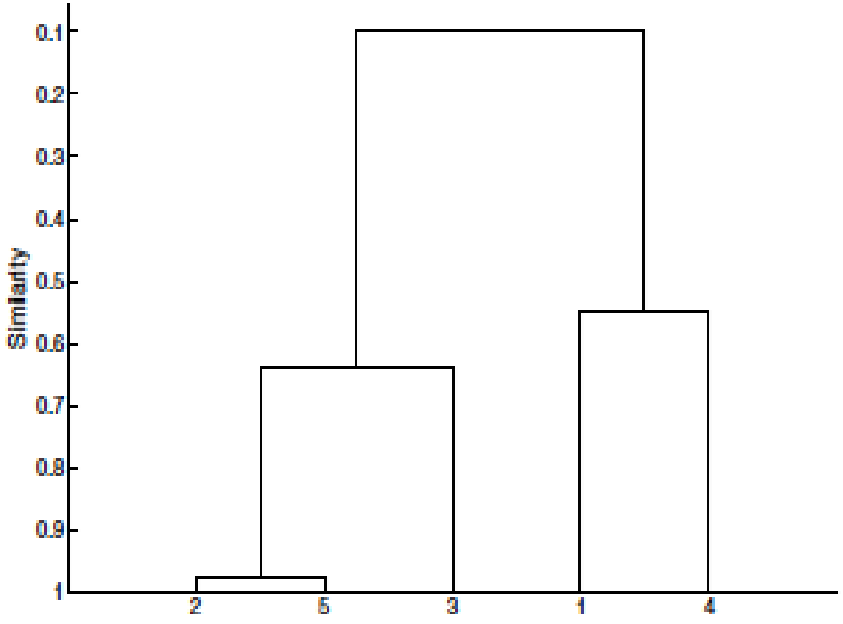
Table 8.1. Similarity matrix for Exercise 16.

	p1	p2	p3	p4	p5
p1	1.00	0.10	0.41	0.55	0.35
p2	0.10	1.00	0.64	0.47	0.98
p3	0.41	0.64	1.00	0.44	0.85
p4	0.55	0.47	0.44	1.00	0.76
p5	0.35	0.98	0.85	0.76	1.00

SOLUTION



(a) Single link.



(b) Complete link.

Figure 8.6. Dendrograms for Exercise 16.