Sheet04: CHAPTER 28: DATA MINING CONCEPTS

Answers to Selected Problems

28.14 Apply the Apriori algorithm to the following data set:

Trans ID	Items Purchased
101	milk, bread, eggs
102	milk, juice
103	juice, butter
104	milk, bread, eggs
105	coffee, eggs
106	coffee
107	coffee, juice
108	milk, bread, cookies, eggs
109	cookies, butter
110	milk, bread

The set of items is {milk, bread, cookies, eggs, butter, coffee, juice}. Use 0.2 for the minimum support value.

Answer:

First, we compute the support for 1-item sets
(e.g., milk appears in 5 out of the 10 transactions, support is 0.5):

1-ITEM SETS	SUPPORT
milk	0.5
bread	0.4
eggs	0.4
coffee	0.3
juice	0.3
cookies	0.2
butter	0.2

The min support required is 0.2, so all 1-item sets satisfy this requirement, i.e. they are all frequent.

For the next iteration, we examine 2-item sets composed of the frequent 1-item sets. The number of potential 2-item sets is 21 (i.e., 7 items taken 2 at a time). The 2-item sets that satisfy the min support of 0.2 are the following:

2-ITEM SETS	SUPPORT
milk,bread	0.4
milk,eggs	0.3
bread,eggs	0.3

For the next iteration, we examine 3-item sets composed of the frequent 2-item sets. The 3-item sets that satisfy the min support of 0.2 are the following:

3-ITEM SETS	SUPPORT

milk,bread,eggs 0.3

28.15 Show two rules that have a confidence of 0.7 or greater for an itemset containing three items from Exercise 14.

Answer:

There is only one frequent itemset of size 3, i.e., {milk,bread,eggs}. We can try the rule milk,eggs -> bread. The confidence of this rule is 0.3/0.3 which exceeds the min confidence value of 0.7. Another rule we can try is bread -> milk,eggs. The confidence of this rule is 0.3/0.4 which again satisfies the min confidence requirement.

28.20 Consider the following set of two-dimensional records:

RID	Dimension 1	Dimension 2
1	8	4
2	5	4
3	2	4
4	2	6
5	2	8
6	8	6

Also consider two different clustering schemes: (1) where Cluster 1 contains records {1, 2, 3} and Cluster 2 contains records {4, 5, 6} and (2) where Cluster 1 contains records {1, 6} and Cluster 2 contains records {2, 3, 4, 5}. Which scheme is better and why?

Answer:

Compare the error of the two clustering schemes. The scheme with the smallest error is better.

For SCHEME (1) we have C1 = {1,2,3} and C2 = {4,5,6}
M1 = ((8+5+2)/3, (4+4+4)/3) = (5,4)

$C1\_error = (8-5)^2 + (4-4)^2 + (5-5)^2 + (4-4)^2 + (2-5)^2 + (4-4)^2$
$= 18$

For C2 we have
M2 = ((2+2+8)/3, (6+8+6)/3) = (4,6.66)

$C2\_error = (2-4)^2 + (6-6.66)^2 + (2-4)^2 + (8-6.66)^2 + (8-4)^2 + (6-6.66)^2$
$= 26.67$

C1_error + C2_error = 44.67

For SCHEME (2) we have C1 = {1,6} and C2 = {2,3,4,5}
M1 = ((8+8)/2, (4+6)/2) = (8,5)

C1_error = $(8-8)^2 + (4-5)^2 + (8-8)^2 + (6-5)^2$
= 2

For C2 we have
M2 = ((5+2+2+2)/4, (4+4+6+8)/4) = (2.75,5.5)

C2_error =
$(5-2.75)^2 + (4-5.5)^2 + (2-2.75)^2 + (4-5.5)^2 + (2-2.75)^2 + (6-5.5)^2 + (2-2.75)^2 + (8-5.5)^2$

= 17.74

C1_error + C2_error = 19.74

SCHEME 2 is better since the error associated with it is less than that of SCHEME (1).

28.21 Use the K-means algorithm to cluster the data from Exercise 28.20. We can use a value of 3 for K and can assume that the records with RIDs 1, 3, and 5 are used for the initial cluster centroids (means).

Answer:

We start by specifying the centroid for each of the 3 clusters.
C1's centroid is (8,4) , i.e., record with rid = 1
C2's centroid is (2,4) , i.e., record with rid = 3
C3's centroid is (2,8) , i.e., record with rid = 5

We now place the remaining records in the cluster whose centroid is closest.

The distance between record 2, i.e., point (5,4), and centroid for C1 is

SQROOT($|8-5|^2 + |4-4|^2$) = 3

The distance between record 2 and centroid for C2 is

SQROOT($|2-5|^2 + |4-4|^2$) = 3

The distance between record 2 and centroid for C3 is

SQROOT($|2-5|^2 + |8-4|^2$) = 5

Record 2 can be placed in either C1 or C2 since the distance from their respective centroids are the same. Let's choose to place the record in C1.

The distance between record 4, i.e., point (2,6), and centroid for C1 is

SQROOT($|8-2|^2 + |4-6|^2$) = 6.32

The distance between record 4 and centroid for C2 is

SQROOT($|2-2|^2$ + $|4-6|^2$) = 2

The distance between record 4 and centroid for C3 is

SQROOT($|2-2|^2$ + $|8-6|^2$) = 2

Record 4 can be placed in either C2 or C3 since the distance from their respective centroids are the same. Let's choose to place the record in C2.

The distance between record 6, i.e., point (8,6), and centroid for C1 is

SQROOT($|8-8|^2$ + $|4-6|^2$) = 2

The distance between record 6 and centroid for C2 is

SQROOT($|2-8|^2$ + $|4-6|^2$) = 6.32

The distance between record 6 and centroid for C3 is

SQROOT($|2-8|^2$ + $|8-6|^2$) = 6.32

Record 6 is closest to centroid of cluster C1 and is placed there.

We now recalculate the cluster centroids:

C1 contains records {1,2,6} with a centroid of
((8+5+8)/3, (4+4+6)/3) = (7, 4.67)

C2 contains records {3,4} with a centroid of
((2+2)/2, (4+6)/2) = (2, 5)

C3 contains record {5} with a centroid of (2, 8)

We now make a second iteration over the records, comparing the distance of each record with the new centroids and possibly moving records to new clusters. As it turns out, all records stay in their prior cluster assignment. Since there was no change, the algorithm terminates.