Assignment 5

Note: Show all your work.

Problem 1 (30 points). Consider the following transactional database.

TID	Items
100	1,2,3,5,6,8
200	1,2,4,5,6
300	1,2,3,4,5,6,8
400	1,2,3,5,7
500	1,2,3,5,8

(1) Mine all frequent itemsets using Apriori. Show all candidate itemsets, frequent itemsets, and all pruning you did. You should follow the process described in the book and lecture (i.e., C1 → L1 → C2 → L2 → …). Minimum support = 60% (or 3 or more transactions).

Answer:

C1	
Item set	Sub
{1}	5
{2}	5
{3}	4
{4}	2
{5}	5
{6}	3
{7}	1
{8}	3

From C1, we remove all 1-itemsets that do not satisfy the minimum support. The remaining item sets are called frequent 1-itemsets, or L1.

L1	
Item set	Sub
{1}	5
{2}	5
{3}	4
{5}	5
{6}	3
{8}	3

From L1 we generate candidate 2-itemsets by joining two frequent 1-itemsets. These itemsets are called candidate 2-itemsets, or C2.

C2
Item set
{1,2}
{1,3}
{1,5}
{1,6}
{1,8}
{2,3}
{2,5}
{2,6}
{2,8}
{3,5}
{3,6}
{3,8}
{5,6}
{5,8}
{6,8}

Next, we determine the support count of each candidate 2-itemset

C2 **(with counts)**	
Item set	Sub
{1,2}	5
{1,3}	4
{1,5}	5
{1,6}	3
{1,8}	3
{2,3}	4
{2,5}	5
{2,6}	3
{2,8}	3
{3,5}	4
{3,6}	2
{3,8}	3
{5,6}	3
{5,8}	3
{6,8}	2

Again, we remove all 2-itemsets which do not satisfy the minum support. The remaining itemsets are called frequent 2-itemsets, or L2.

L2	
Item set	Sub
{1,2}	5
{1,3}	4
{1,5}	5
{1,6}	3
{1,8}	3
{2,3}	4
{2,5}	5
{2,6}	3
{2,8}	3
{3,5}	4
{3,8}	3
{5,6}	3
{5,8}	3

The next iteration is to generate candidate 3-itemsets, or C3, from L2. Candidate 3-itemsets are generated by joining two frequent 2-itemsets.

C3 (before pruning)
Item set
{1,2,3}
{1,2,5}
{1,2,6}
{1,2,8}
{1,3,5}
{1,3,6}
{1,3,8}
{1,5,6}
{1,5,8}
{1,6,8}
{2,3,5}
{2,3,6}
{2,3,8}
{2,5,6}
{2,5,8}
{2,6,8}
{3,5,8}
{5,6,8}

C3(with count)	
Item set	Sub
{1,2,3}	4
{1,2,5}	5
{1,2,6}	3
{1,2,8}	3
{1,3,5}	4
{1,3,6}	2
{1,3,8}	3
{1,5,6}	3
{1,5,8}	3
{1,6,8}	2
{2,3,5}	4
{2,3,6}	2
{2,3,8}	3

After removing those data set which do not meet the minimum support, we have frequent 3-itemsets, or L3.

L3	
Item set	Sub
{1,2,3}	4
{1,2,5}	5
{1,2,6}	3
{1,2,8}	3
{1,3,5}	4
{1,3,8}	3
{1,5,6}	3
{1,5,8}	3
{2,3,5}	4
{2,3,8}	3

C4 (before pruning)
Item set
{1,2,3,5}
{1,2,3,6}
{1,2,3,8}
{1,2,5,6}
{1,2,5,8}
{1,2,6,8}
{1,3,5,8}
{1,5,6,8}
{2,3,5,8}

C4(with count)	
Item set	Sub
{1,2,3,5}	4
{1,2,3,6}	2
{1,2,3,8}	3
{1,2,5,6}	3
{1,2,5,8}	3
{1,2,6,8}	2
{1,3,5,8}	3
{1,5,6,8}	2
{2,3,5,8}	3

L4	
Item set	Sub
{1,2,3,5}	4
{1,2,3,8}	3
{1,2,5,6}	3
{1,2,5,8}	3
{1,3,5,8}	3
{2,3,5,8}	3

C5 (before pruning)
Item set
{1,2,3,5,8}
{1,2,5,6,8}

C5(with count)	
Item set	Sub
{1,2,3,5,8}	3
{1,2,5,6,8}	2

L5	
Item set	Sub
{1,2,3,5,8}	3

(2) Sort all frequent 4-itemsets by their item number. Then, select the first frequent 4-itemset form the sorted list of frequent 4-itemsets and mine all strong rules from this itemset that have the format {W, X} => {Y, Z}, where W, X, Y, and Z are individual items. Assume that minimum confidence = 70%.

L4	
Item set	Sub
{1,2,3,5}	4
{1,2,3,8}	3
{1,2,5,6}	3
{1,2,5,8}	3
{1,3,5,8}	3
{2,3,5,8}	3

R1: {1,2} => {3,5}
R2: {1,3} => {2,5}
R3: {1,5} => {2,3}
R4: {2,3} => {1,5}
R5: {2,5} => {1,3}
R6: {3,5} => {1,2}

confidence = sup(all items) / sup(antecedent)

conf(R6) = (sup({ 1,2,3,5})) / sup({3,5}) = 4/4 = 100%
conf(R5) = (sup({ 1,2,3,5})) / sup({2,5}) = 4/5 = 80%
conf(R4) = (sup({ 1,2,3,5})) / sup({2,3}) = 4/4 = 100%
conf(R3) = (sup({ 1,2,3,5})) / sup({1,5}) = 4/5 = 80%
conf(R2) = (sup({ 1,2,3,5})) / sup({1,3}) = 4/4 = 100%
conf(R1) = (sup({ 1,2,3,5})) / sup({1,2}) = 4/5 = 80%

Because of *min_conf* = 70%, then all are strong rules.

Problem 2. Consider the following contingency table.

	C (buys coffee = Yes)	*C* (buys coffee = No)	*Sum*
T (buys tea = Yes)	125	35	160
T (buys tea = No)	65	425	490
Sum	190	460	650

Compute the *lift*, *all-confidence*, *cosine*, *Kulczynski* and *imbalance ratio* measure, and determine whether buying coffee and buying tea are positively correlated, negatively correlated, or not correlated.

Answer:

$$lift = \frac{P(A \cup B)}{P(A)P(B)}$$

Lift(T,C) = (125/650) / [(160/650)*(190/650)] = 34.21 (says positively correlated)

Lift(T,-C) = (35/650) / [(160/650)*(460/650)] = 0.31 (says negatively correlated)

all_conf(T,C) = sup(T∪C) / max{sup(T),sup(c)} = 125 / 190 = 0.66

$$cosine(m,c) = \frac{sup(m \cup c)}{\sqrt{\sup(m) \times \sup(c)}}$$

Cosine(T,C) = $\frac{\sup(TUC)}{\sqrt{\sup(T) \times \sup(C)}} = \frac{125}{\sqrt{190 \times 160}} = 0.72$ greater than 0.5: positively correlated

$$Kulc(A, B) = (P(A|B) + P(B|A)) / 2$$

Kulc(T, C) = (P(T\C) + P(C\T)) / 2 = ½ [(T.C / C) + (T.C / T)] = ½[(125/190) + (125/160)]

Kulc(T, C) = 0.72 72 greater than 0.5: positive

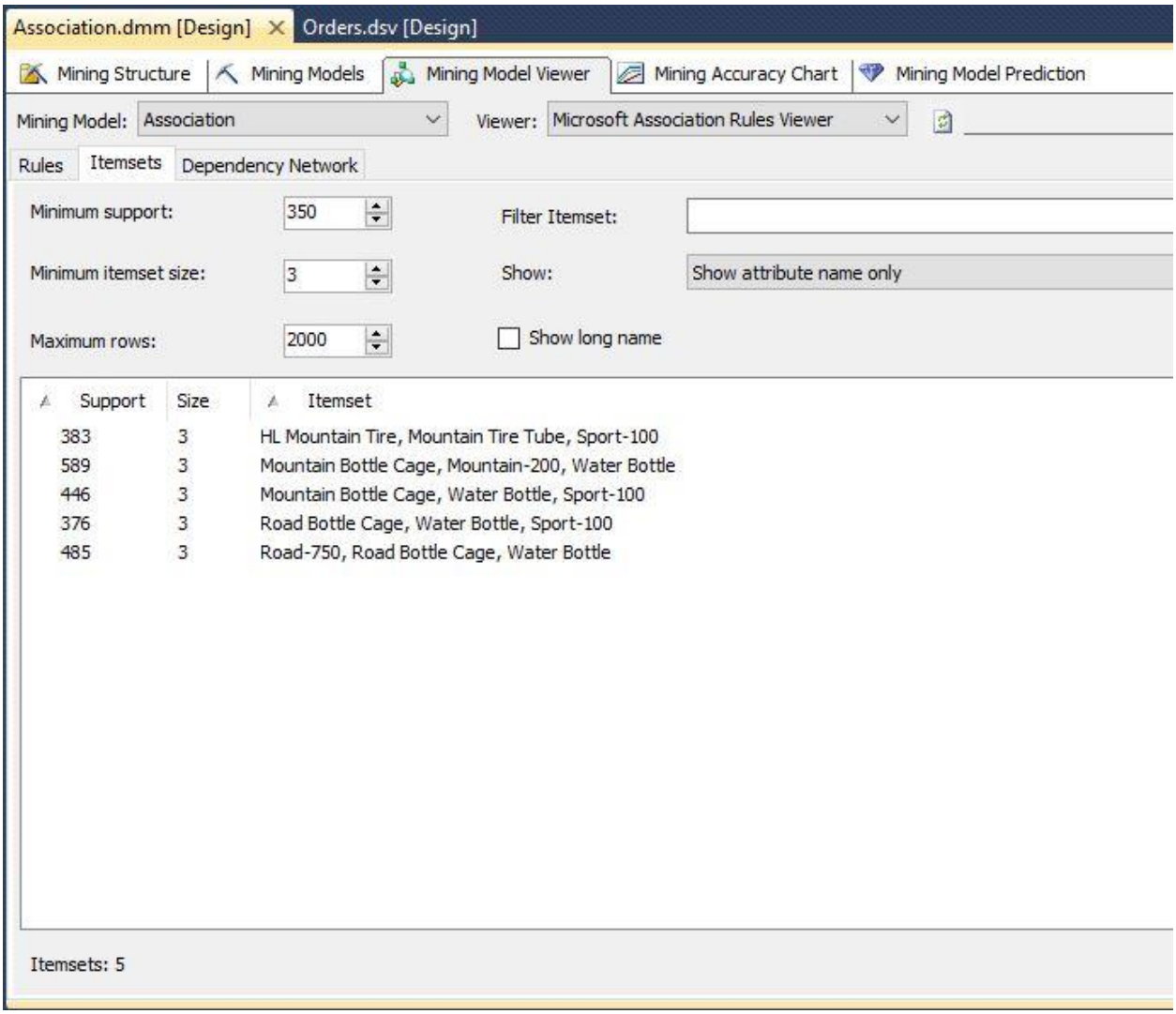
$$IR(A,B) = \frac{|\sup(A) - \sup(B)|}{\sup(A) + \sup(B) - \sup(A \cup B)} \quad , \quad 0 \le IR < 1$$

IR(T,C) = $\frac{|190-160|}{190+160-125} = 0.13$ less than 0.5: negative

Problem 3-1 (SQL Server)

This is a practice of association rule mining using SQL Server. You will follow Lesson 3 of SQL Server 2012 Intermediate Data Mining Tutorial. The link to Intermediate Data Mining Tutorial is: http://msdn.microsoft.com/en-us/library/cc879271(v=sql.110).aspx
But, before beginning Lesson 3, you have to create a new Analysis Project and a data source following the instruction in Lesson 1 of Intermediate Data Mining. Note that, when you create a data source, you may see AdventureWorks2012DW data source is already there in the *Select how to define the connection* page of *Data Source Wizard* (because you already created this data source when you were doing Assignment 4 and if you did not delete it). Then, you can skip to the next step of creating a data source instead of adding a new connection.
Once a project and a data source are created, perform the tasks in Lesson 3. Lesson 3 has six tasks (or topics). Perform the first four tasks (up to, and including, *Exploring the Market Basket Models*). At the end of the fourth task, there is a section titled *Generic Content Tree Viewer*. You don't have to do this section. If you do, keep in mind that the description there may not be the same as what you see on your model.
Requirements: In the fourth task *Exploring the Market Basket Models*, you will see

association rules and frequent itemsets mined by your model. Explore the rules and itemsets. Click *Itemsets* tab under *Mining Model Viewer* tab. Capture this screen and paste it onto your submission. Note that the tutorial tells you to change the minimum support to 100 in the section titled *To filter itemsets by support or size*. However, it may not allow you to change the minimum support below 213. If so, simply ignore that part. Change *Minimum support* to 350 and click an empty space inside the *Itemset* window. There will be about five or six 3-itemsets. In your submission, include all these 3-itemsets along with their supports. Then, pick the 3-itemset with the highest support and **manually** mine all rules from this itemset and compute their confidences. You need to show all your work.



All nonempty proper subsets are:

{Mountain Bottle Cage}, {Mountain-200}, {Water Bottle}, {Mountain Bottle Cage, Mountain-200}, {Mountain Bottle Cage, Water Bottle}, {Mountain-200, Water Bottle}

For each subset, we form a rule:
R1: {Mountain Bottle Cage} => {Mountain-200, Water Bottle}
R2: {Mountain-200} => {Mountain Bottle Cage, Water Bottle}
R3: {Water Bottle} => {Mountain Bottle Cage, Mountain-200}
R4: {Mountain Bottle Cage, Mountain-200} => {Water Bottle}
R5: {Mountain Bottle Cage, Water Bottle} => {Mountain-200}
R6: {Mountain-200, Water Bottle} => {Mountain Bottle Cage}

Confidence = sup(all items) / sup(antecedent)

conf(R1) = (sup({Mountain Bottle Cage, Mountain-200, Water Bottle })) / sup({Mountain Bottle Cage }) = 589 / 1941 = %30

conf(R2) = (sup({Mountain Bottle Cage, Mountain-200, Water Bottle })) / sup({Mountain-200}) = 589 / 2477 = %24

conf(R3) = (sup({Mountain Bottle Cage, Mountain-200, Water Bottle })) / sup({Water Bottle }) = 589 / 4076 = %14

conf(R4) = (sup({Mountain Bottle Cage, Mountain-200, Water Bottle })) / sup({Mountain Bottle Cage, Mountain-200}) = 589 / 725 = %81

conf(R5) = (sup({Mountain Bottle Cage, Mountain-200, Water Bottle })) / sup({Mountain Bottle Cage, Water Bottle }) = 589 / 1623 = %36

conf(R6) = (sup({Mountain Bottle Cage, Mountain-200, Water Bottle })) / sup({Mountain-200, Water Bottle }) = 589 / 589 = %100