Exercise I (Association Rules)

Consider the mining of association rules on the transactions:

transaction id	items
1	A, B, E
2	A, B, D, E
3	B, C, D, E
4	B, D, E
5	A, B, D
6	B, E
7	A, E

A. What is the support of the itemset {B,D,E}?

B. What is the support and confidence of the association rule BD → E?

C. Consider the application of the Apriori algorithm to find all the frequent itemsets whose counts are at least 3. Recall that the algorithm scans the transaction list a number of times, where the i^{th} scan generates the set F_i of all size-i frequent itemsets from a candidate set C_i. Show C_i and F_i for each possible i.

D. Find all the association rules with support at least 3 and confidence at least 3/4. For your convenience, all the itemsets with support at least 3 are {{A}, {B}, {D}, {E}, {A,B},{A,E},{B,D},{B,E},{D,E}, {B,D,E}}.

Problem 1. What is the support of the itemset $\{B, D, E\}$?

Answer.
The support count is 3 because transactions 2, 3 and 4 contain the itemset.

Problem 2. What is the support and confidence of the association rule $BD \to E$?

Answer.
The support $BD \to E$ is the support of $\{B, D, E\}$ which is 3. The confidence is

$$conf(BD \to E) = \frac{support(\{B, D, E\})}{support(\{B, D\})} = \frac{3}{4}.$$

Answer.
For the first scan, the candidate set C_1 contains all the singleton sets, i.e., C_1 includes $\{A\}$, $\{B\}$, $\{C\}$, $\{D\}$ and $\{E\}$. After the scan, only $\{A\}$, $\{B\}$, $\{D\}$ and $\{E\}$ remain in F_1. In particular, $\{C\}$ is eliminated because its count 1 is smaller than 3.

From F_1, the algorithm generates:

$$C_2 = \{\{A, B\}, \{A, D\}, \{A, E\}, \{B, D\}, \{B, E\}, \{D, E\}\}$$

The second scan produces:

$$F_2 = \{\{A, B\}, \{A, E\}, \{B, D\}, \{B, E\}, \{D, E\}\}$$

$\{A, D\}$ is removed because its count 2 is lower than 3.

From F_2, the algorithm generates:

$$C_3 = \{\{A, B, E\}, \{B, D, E\}\}$$

as follows. For each pair of distinct itemsets $\{a_1, a_2\}$ and $\{b_1, b_2\}$ in F_2, the algorithm adds to C_3 an itemset $\{a_1, a_2, b_2\}$ if and only if $a_1 = b_1$. Hence, $\{A, B\}$ and $\{A, E\}$ give rise to $\{A, B, E\}$, whereas $\{B, D\}$ and $\{B, E\}$ give rise to $\{B, D, E\}$.

Finally, the third scan produces:

$$F_3 = \{\{B, D, E\}\}$$

as you can verify easily by yourself. The algorithm terminates here.

Answer.
The following table lists all the possible association rules and their confidence values. The ones in bold are the final answers.

rule	confidence
$\boldsymbol{A \to B}$	3/4
$B \to A$	1/2
$\boldsymbol{A \to E}$	3/4
$E \to A$	1/2
$B \to D$	2/3
$\boldsymbol{D \to B}$	1
$\boldsymbol{B \to E}$	5/6
$\boldsymbol{E \to B}$	5/6
$\boldsymbol{D \to E}$	3/4
$E \to D$	1/2
$B \to DE$	1/2
$\boldsymbol{BD \to E}$	3/4
$BE \to D$	3/5
$\boldsymbol{D \to BE}$	3/4
$\boldsymbol{DE \to B}$	1
$E \to BD$	1/2

Exercise II (Association Rules)

The following is an example of customer purchase transaction data set.

CID	TID	Date	Items Purchased
1	1	01/01/2001	10,20
1	2	01/02/2001	10,30,50,70
1	3	01/03/2001	10,20,30,40
2	4	01/03/2001	20,30
2	5	01/04/2001	20,40,70
3	6	01/04/2001	10,30,60,70
3	7	01/05/2001	10,50,70
4	8	01/05/2001	10,20,30
4	9	01/06/2001	20,40,60
5	10	01/11/2001	10,20,30,60

Note: CID = Customer ID and TID = Transactions ID

Q.1 Calculate the *support*, *confidence* and *lift* of the following association rule. Indicate if the items in the association rule are independent of each other or have negative or positive impacts on each other.

{10} -> {50,70}

Q.2 The following is the list of large two item sets. Show the steps to apply the Apriori property to generate and prune the candidates for large three itemsets. Describe how the Apriori property is used is in the steps. Give the final list of candidates large three item sets.

{10,20} {10,30} {20,30} {20,40}

Answer to Q.1 Support = Support ({10,50,70}) = 2/10 = 20%
Confidence = Support ({10,50,70})/ Support({10}) = 0.2/0.7 = 2/7 = 29%
Lift = Confidence/Support({50,70}) = 2/7/0.2 = 10/7 = 1.43 > 1

Since lift is larger than 1, it's a positive rule.

Answer to Q.2
{10,20} {10,30} {20,30} {20,40}
***O: describe how the apriori property is used to decide which 2 large item sets are joined together and to determine which 3 item set should be pruned.

Join: {10,20,30} {20,30,40}
Prune: {10,20,30} ({20,30,40} is pruned)
Final list: {10,20,30}