Aaron Leung
ahleung2
leung.aa@northeastern.edu

Question 3

a) List all candidate 4-itemsets obtained by a candidate generation procedure using the $F_{k-1} \times F_1$ merging strategy.

To determine the items in F_1, we will use the Apriori principle which states that if an itemset is frequent, then all of its subsets must also be frequent. From this and given the set of frequent 3-itemsets we can conclude that F_1 includes {1}, {2}, {3}, {4} and {5}.

F_{k-1} (Frequent 3-itemset)

Itemset
{1, 2, 3}
{1, 2, 4}
{1, 2, 5}
{1, 3, 4}
{2, 3, 4}
{2, 3, 5}
{3, 4, 5}

F_1 (Frequent 1-itemset)

Itemset
{1}
{2}
{3}
{4}
{5}

Aaron Leung
ahleung2
leung.aa@northeastern.edu

In this merging strategy, we extend each frequent (k-1) itemset with frequent items that are not part of (k-1) itemset. In other words, every frequent k-itemset is composed of a frequent (k-1) itemset and frequent 1-itemset. To prevent duplicate candidate itemsets from being generated, itemsets will only be extended with frequent items that are in increasing numerical order.

- {1, 2, 3} can be extended by {4} and {5} respectively. Therefore, itemsets {1, 2, 3, 4} and {1, 2, 3, 5} are generated.
- {1, 2, 4} can only be extended by {5}. Therefore, itemset {1, 2, 4, 5} is generated.
- {1, 2, 5} can't be extended.
- {1, 3, 4} can only be extended by {5}. Therefore, itemset {1, 3, 4, 5} is generated.
- {2, 3, 4} can only be extended by {5}. Therefore, itemset {2, 3, 4, 5} is generated.
- {2, 3, 5} can't be extended.
- {3, 4, 5} can't be extended.

Therefore, the candidate 4-itemsets include the following: {1, 2, 3, 4}, {1, 2, 3, 5}, {1, 2, 4, 5}, {1, 3, 4, 5} and {2, 3, 4, 5}.

b) List all candidate 4-itemsets obtained by the candidate generation procedure in Apriori, using $F_{k-1} \times F_{k-1}$.

This method only merges a pair of (k-1) itemsets only if their first k-2 items are identical. Similar to part a), items are in increasing numerical order and itemsets will only be extended with frequent items that are in increasing numerical order.

- {1, 2, 3}, {1, 2, 4} and {1, 2, 5} all start with {1, 2} and thus we can merge these itemsets with one another.
  - {1, 2, 3} merged with {1, 2, 4} generates {1, 2, 3, 4}
  - {1, 2, 3) merged with {1, 2, 5} generates {1, 2, 3, 5}
  - {1, 2, 4} merged with {1, 2, 5} generates {1, 2, 4, 5}
- {2, 3, 4} and {2, 3, 5} both start with {2, 3} and can be merged.
  - {2, 3, 4} merged with {2, 3, 5} generates {2, 3, 4, 5}

Aaron Leung
ahleung2
leung.aa@northeastern.edu

Therefore, the candidate 4-itemsets include the following: {1, 2, 3, 4}, {1, 2, 3, 5}, {1, 2, 4, 5} and {2, 3, 4, 5}

c) List all candidate 4-itemsets that survive the candidate pruning step of the Apriori algorithm.

Now that we have generated the candidate 4-itemsets, we can now go through every possible subset (for each candidate) and see whether any of the subsets are infrequent. If any of the subsets of a candidate itemset are infrequent, we can prune that candidate itemset.

Candidate 4-itemset				
{1, 2, 3, 4}	{1, 2, 3}	{1, 2, 4}	{1, 3, 4}	{2, 3, 4}
{1, 2, 3, 5}	{1, 2, 3}	{1, 2, 5}	{1, 3, 5}	{2, 3, 5}
{1, 2, 4, 5}	{1, 2, 4}	{1, 2, 5}	{1, 4, 5}	{2, 4, 5}
{2, 3, 4, 5}	{2, 3, 4}	{2, 3, 5}	{2, 4, 5}	{3, 4, 5}

Itemsets that are highlighted in red are infrequent. Since these itemsets are infrequent, we know that the corresponding supersets must also be infrequent and we can prune these supersets. Based on this, the candidate itemset that survives the candidate pruning set is {1, 2, 3, 4}.

Question 4

a) What is the maximum number of association rules that can be extracted from this data (including rules that have zero support)?

The maximum number of association rules that can be extracted is given by the following formula: $3^d - 2^{d+1} + 1$ where d represents the items in the set. In this case, d = 7 ({Beer, Diapers, Milk, Bread, Butter, Cookies, Eggs}), so maximum number of association rules = $3^7 - 2^{7+1} + 1 = 1932$.

b) What is the confidence of the rule {Milk, Diapers} ⇒ {Butter}?

The rule's confidence is obtained by dividing the support count for {Milk, Diapers, Butter} by the support count for {Milk, Diapers}. The support counts for {Milk, Diapers, Butter} and {Milk, Diaper} are 2 and 4 respectively. Therefore, the confidence for this rule is 2/4 = 0.5.

Aaron Leung
ahleung2
leung.aa@northeastern.edu

c) What is the support for the rule {Milk, Diapers} ⇒ {Butter}?

Support is given by the following equation:

$$s(x \rightarrow y) = \frac{\sigma(x \cup y)}{N} \text{ where N is the total nuber of transactions}$$

In this case x U y = {Milk, Diapers, Butter} and N = 10. The support count for {Milk, Diapers, Butter} is 2. Therefore, the rule's support is 2/10 = 0.2.

d) True or False with an explanation: Given that {a,b,c,d} is a frequent itemset, {a,b} is always a frequent itemset.

True: The Apriori Principle states that if an itemset is frequent, then all of its subsets must also be frequent.

e) True or False with an explanation: Given that {a,b}, {b,c} and {a,c} are frequent itemsets, {a,b,c} is always frequent.

True: Based on the statement, we make the assumption that {a, b, c} is in the data set. Applying the Apriori Principle to the given frequent itemsets listed above, we can conclude that {a}, {b} and {c} are frequent size-1 itemsets. These size-1 itemsets along with {a, b}, {b,c} and {a, c} collectively make up all the subsets of {a, b, c} and are all frequent. Therefore, assuming {a, b, c} is in the dataset, {a, b, c} is always frequent.

f) True or False with an explanation: Given that the support of {a,b} is 20 and the support of {b,c} is 30, the support of {b} is larger than 20 but smaller than 30.

False: According to the anti-monotone property, the support of an itemset never exceeds the support for its subsets. In this scenario {b} is the subset of both {a,b} and {b,c}. If the support of {b} is between 20 and 30, this would violate the property above as {b,c} would exceed the support of subset {b}.

Aaron Leung
ahleung2
leung.aa@northeastern.edu

g) True or False with an explanation: In a dataset that has 5 items, the maximum number of size-2 frequent itemsets that can be extracted (assuming minsup > 0) is 20.

False: In order to determine the maximum number of size-2 frequent itemsets that can be extracted (asuming minus>0), we need to consider all possible unique size-2 frequent itemsets. The number of possible size-2 frequent itemsets is given by ${}_5C_2$ which comes out to be 10. If we now assume everyone one of these itemsets to be frequent, then the max number of unique size-2 frequent items would be 10 and not 20.

h) Draw the itemset lattice for the set of unique items I = {a, b, c}.