CS548 Homework

Cate Dunham

April 7, 2025

Exercise 5.2

Consider the data set shown in Table 5.20.

Table 6.22. Example of market basket transactions.

Customer ID	Transaction ID	Items Bought
1	0001	$\{a, d, e\}$
1	0024	$\{a, b, c, e\}$
2	0012	$\{a, b, d, e\}$
2	0031	$\{a, c, d, e\}$
3	0015	$\{b, c, e\}$
3	0022	$\{b, d, e\}$
4	0029	$\{c, d\}$
4	0040	$\{a, b, c\}$
5	0033	$\{a, d, e\}$
5	0038	$\{a, b, e\}$

a. Compute the support for itemsets $\{e\}$, $\{b, d\}$, and $\{b, d, e\}$ by treating each transaction ID as a market basket.

- Support for $\{e\} = \frac{8}{10} = .8$
- Support for $\{b, d\} = \frac{2}{10} = .2$
- Support for $\{b, d, e\} = \frac{2}{10} = .2$

b. Use the results in part (a) to compute the confidence for the association rules $\{b, d\} \rightarrow \{e\}$ and $\{e\} \rightarrow \{b, d\}$. Is confidence a symmetric measure?

- Confidence for $\{b, d\} \rightarrow \{e\} = \frac{.2}{.2} = 1$
- Confidence for $\{e\} \rightarrow \{b, d\} = \frac{.2}{.8} = .25$

Confidence here is not a symmetric measure.

c. Repeat part (a) by treating each customer ID as a market basket. Each item should be treated as a binary variable (1 if an item appears in at least one transaction bought by the customer, and 0 otherwise.)

- Support for $\{e\} = \frac{4}{5} = .8$

Customer ID	Item a	Item b	Item c	Item d	Item e
1	1	1	1	1	1
2	1	1	1	1	1
3	0	1	1	1	1
4	1	1	1	1	0
5	1	1	0	1	1

Table 1: Customer ID as Market Basket

- Support for $\{b, d\} = \frac{5}{5} = 1$
- Support for $\{b, d, e\} = \frac{4}{5} = .8$

d. Use the results in part (c) to compute the confidence for the association rules $\{b, d\} \rightarrow \{e\}$ and $\{e\} \rightarrow \{b, d\}$.

- Confidence for $\{b, d\} \rightarrow \{e\} = \frac{.8}{1} = .8$
- Confidence for $\{e\} \rightarrow \{b, d\} = \frac{.8}{.8} = 1$

e. Suppose s_1 and c_1 are the support and confidence values of an association rule r when treating each transaction ID as a market basket. Also, let s_2 and c_2 be the support and confidence values of r when treating each customer ID as a market basket. Discuss whether there are any relationships between s_1 and s_2 or c_1 and c_2.

Support for $\{e\}$ is consistent across s_1 and s_2. Apart from that there are no obvious relationships between s_1 and s_2 or c_1 and c_2.

Exercise 5.3

a. What is the confidence for the rules $\emptyset \rightarrow A$ and $A \rightarrow \emptyset$?

- Confidence for $\emptyset \rightarrow A = \frac{\sigma(\emptyset \cup A)}{\sigma(\emptyset)} = \sigma(\emptyset \cup A)$
- Confidence for $A \rightarrow \emptyset = \frac{\sigma(A \cup \emptyset)}{\sigma(A)} = \frac{\sigma(A)}{\sigma(A)} = 1$

b. Let c_1, c_2, and c_3 be the confidence values of the rules $\{p\} \rightarrow \{q\}$, $\{p\} \rightarrow \{q, r\}$, and $\{p, r\} \rightarrow \{q\}$, respectively. If we assume that c_1, c_2, and c_3 have different values, what are the possible relationships that may exist among c_1, c_2, and c_3? Which rule has the lowest confidence?

We can evaluate each rule as follows:

- $c_1 = \frac{\sigma(p \cup q)}{\sigma(p)}$
- $c_2 = \frac{\sigma(p \cup q \cup r)}{\sigma(p)}$
- $c_3 = \frac{\sigma(p \cup q \cup r)}{\sigma(p \cup r)}$

Because we know that $\sigma(p \cup q) \geq \sigma(p \cup q \cup r)$, we can say that $c_1 \geq c_2$. Since we know that $\sigma(p) \geq \sigma(p \cup r)$, we can say that $c_3 \geq c_2$. Since $c_1 \geq c_2$ and $c_3 \geq c_2$, (and if we assume that there is a single rule with the lowest confidence, i.e., $c_3 \neq c_2$) we know that c_2, associated with the rule $\{p\} \rightarrow \{q, r\}$, has the lowest confidence.

c. Repeat the analysis in part (b) assuming that the rules have identical support. Which rule has the highest confidence?

From the three rules listed we can learn about $\sigma(p \cup q)$ and $\sigma(p \cup q \cup r)$, and based on this question we assume that $\sigma(p \cup q) = \sigma(p \cup q \cup r)$. The three listed rules do *not* tell us about $\sigma(p \cup r)$, which we previously established as $\leq \sigma(p)$. Based on this (and the assumption that there is a single rule with the highest confidence, i.e., $c_3 \neq c_2$), c_3 and the associated rule $\{p, r\} \rightarrow \{q\}$ have the highest confidence.

d. Transitivity: Suppose the confidence of the rules $A \rightarrow B$ and $B \rightarrow C$ are larger than some threshold, $minconf$. Is it possible that $A \rightarrow C$ has a confidence less than $minconf$?

Yes, this is entirely possible. We can use Table 2 as an example:

Transaction ID	A	B	C
1	1	1	1
2	1	1	1
3	1	1	0
4	1	0	0

Table 2: Transitivity Example

From Table 2, we can say that:

- $s(A) = 1$
- $s(B) = .75$
- $s(C) = .5$
- $s(A \rightarrow B) = .75$
- $s(B \rightarrow C) = .5$
- $s(A \rightarrow C) = .5$
- Confidence $A \rightarrow B = \frac{.75}{1} = .75$
- Confidence $B \rightarrow C = \frac{.5}{.75} = .67$
- Confidence $A \rightarrow C = \frac{.5}{1} = .5$

If $minconf = .6$, we would be able to say that Confidence $A \rightarrow C < minconf$.

Exercise 5.4

For each of the following measures, determine whether it is monotone, anti-monotone, or non-monotone (i.e., neither monotone nor anti-monotone).

Example: Support, $s = \frac{\sigma(X)}{|T|}$ is anti-monotone because $s(X) \geq s(Y)$ whenever $X \subset Y$.

a. A characteristic rule is a rule of the form $\{p\} \rightarrow \{q1, q2, ..., qn\}$, where the rule antecedent contains only a single item. An itemset of size k can produce up to k characteristic rules. Let ζ be the minimum confidence of all characteristic rules generated from a given itemset:

$$\zeta(\{p_1, p_2, ..., p_k\}) = \min[c(\{p_1\} \rightarrow \{p_2, p_3, ..., p_k\}), ... \quad (1)$$
$$c(\{p_k\} \rightarrow \{p_1, p_3 ..., p_{k-1}\})] \quad (2)$$

Is ζ monotone, anti-monotone, or non-monotone?

We know from the textbook that a measure is anti-monotone if for every itemset X that is a proper subset of the itemset Y, $f(Y) \leq f(X)$. Here we can consider $\zeta(\{A, B\})$ and $\zeta(\{A, B, C\})$:

$$\zeta(\{A, B\}) = \min(c(A \rightarrow B), c(B \rightarrow A)) \quad (3)$$
$$= \min(\frac{\sigma(A \cup B)}{\sigma(A)}, \frac{\sigma(A \cup B)}{\sigma(B)}) \quad (4)$$
$$= \frac{\sigma(A \cup B)}{\max(\sigma(A), \sigma(B))} \quad (5)$$

$$\zeta(\{A, B, C\}) = \min(c(A \rightarrow BC), c(B \rightarrow AC), c(C \rightarrow AB)) \quad (6)$$
$$= \min(\frac{\sigma(A \cup B \cup C)}{\sigma(A)}, \frac{\sigma(A \cup B \cup C)}{\sigma(B)}, \frac{\sigma(A \cup B \cup C)}{\sigma(C)}) \quad (7)$$
$$= \frac{\sigma(A \cup B \cup C)}{\max(\sigma(A), \sigma(B), \sigma(C))} \quad (8)$$

Since we know that $\sigma(A \cup B \cup C) \leq \sigma(A \cup B)$ and that $\max(\sigma(A), \sigma(B), \sigma(C)) \geq \max(\sigma(A), \sigma(B))$, we can say that $\zeta(\{A, B, C\}) \leq \zeta(\{A, B\})$, which means that ζ is **anti-monotone**.

b. A discriminant rule is a rule of the form $\{A, B, ..., p_n\} \rightarrow \{q\}$, where the rule consequent contains only a single item. An itemset of size k can produce up to k discriminant rules. Let η be the minimum confidence of all discriminant rules generated from a given itemset:

$$\eta(p_1, p_2, ..., p_k) = \min[c(p_2, p_3, ..., p_k \rightarrow p1), ... \quad (9)$$
$$c(p_1, p_2, ...p_{k-1} \rightarrow p_k)] \quad (10)$$

Is η monotone, anti-monotone, or non-monotone?

We can compare $\eta(\{A, B\})$ and $\eta(\{A, B, C\})$. $\eta(\{A, B\})$ is the same as $\zeta(\{A, B\})$. For $\eta(\{A, B, C\})$:

$$\eta(\{A, B, C\}) = \min(c(AB \to C), c(AC \to B), c(BC \to A)) \tag{11}$$

$$= \min(\frac{\sigma(A \cup B \cup C)}{\sigma(A \cup B)}, \frac{\sigma(A \cup B \cup C)}{\sigma(A \cup C)}, \frac{\sigma(A \cup B \cup C)}{\sigma(B \cup C)}) \tag{12}$$

$$= \frac{\sigma(A \cup B \cup C)}{\max(\sigma(A \cup B), \sigma(A \cup C), \sigma(B \cup C))} \tag{13}$$

We know that $\sigma(A \cup B) \geq \sigma(A \cup B \cup C)$ and we know that $\max(\sigma(A), \sigma(B)) \geq \max(\sigma(A \cup B), \sigma(A \cup C), \sigma(B \cup C))$, all we can say about $\eta(\{A \cup B \cup C\})$ is that it could be greater *or* less than $\eta(\{A \cup B\})$, making η **non-monotone**.

c. Repeat the analysis in parts (a) and (b) by replacing the min function with a max function.

- Part A - replacing the min function with the max function would result in a comparison between:

$$\zeta(\{A \cup B\}) = \frac{\sigma(A \cup B)}{\min(\sigma(A), \sigma(B))} \tag{14}$$

and

$$\zeta(\{A \cup B \cup C\}) = \frac{\sigma(A \cup B \cup C)}{\min(\sigma(A), \sigma(B), \sigma(C))}. \tag{15}$$

Since $\sigma(A \cup B) \geq \sigma(A \cup B \cup C)$ and $\min(\sigma(A), \sigma(B)) \geq \min(\sigma(A), \sigma(B), \sigma(C)$, $\zeta(\{A \cup B\})$ could be greater or less than $\zeta(\{A \cup B \cup C\})$, making ζ **non-monotone**.

- Part B - replacing the min function with the max function would result in a comparison between:

$$\eta(\{A \cup B\}) = \frac{\sigma(A \cup B)}{\min(\sigma(A), \sigma(B))} \tag{16}$$

and

$$\eta(\{A \cup B \cup C\}) = \frac{\sigma(A \cup B \cup C)}{\min(\sigma(A \cup B), \sigma(A \cup C), \sigma(B \cup C))}. \tag{17}$$

Given that $\sigma(A \cup B) \geq \sigma(A \cup B \cup C)$ and $\min(\sigma(A), \sigma(B)) \geq \min(\sigma(A \cup B), \sigma(A \cup C), \sigma(B \cup C))$, η is also **non-monotone**.

Exercise 5.5

Prove Equation 5.3. (Hint: First, count the number of ways to create an itemset that forms the left hand side of the rule. Next, for each size k itemset selected for the left-hand side, count the number of ways to choose the remaining $d-k$ items to form the right-hand side of the rule.)

$$R = 3^d - d^{d+1} + 1 \tag{18}$$

There are $\binom{d}{k}$ ways to choose an itemset of size k from d items and $\binom{d-k}{i}$ ways to choose the remaining items:

$$R = \sum_{k=1}^{d} \binom{d}{k} \sum_{i=1}^{d-k} \binom{d-k}{i} \tag{19}$$

$$\tag{20}$$

We can use the binomial theorem to evaluate $\sum_{i=1}^{d-k} \binom{d-k}{i}$ as $2^{d-k} - 1$:

$$R = \sum_{k=1}^{d} \binom{d}{k} (2^{d-k} - 1)$$

$$R = \sum_{k=1}^{d} \binom{d}{k} 2^{d-k} - \sum_{k=1}^{d} \binom{d}{k}$$

We know that:

$$(1+x)^d = \sum_{i=1}^{d} \binom{d}{i} x^{d-i} + x^d,$$

so:

$$3^d = \sum_{i=1}^{d} \binom{d}{i} 2^{d-i} + 2^d,$$

and:
$R = 3^d - 2^d - (2^d - 1) = 3^d - 2^{d+1} + 1$

Exercise 5.6

Consider the market basket transactions shown in Table 5.21.

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

   For a dataset with 6 items there is a maximum of $= 3^6 - 2^{6+1} + 1 = 729 - 128 + 1 = 602$ association rules.

Table 6.23. Market basket transactions.

Transaction ID	Items Bought
1	{Milk, Beer, Diapers}
2	{Bread, Butter, Milk}
3	{Milk, Diapers, Cookies}
4	{Bread, Butter, Cookies}
5	{Beer, Cookies, Diapers}
6	{Milk, Diapers, Bread, Butter}
7	{Bread, Butter, Diapers}
8	{Beer, Diapers}
9	{Milk, Diapers, Bread, Butter}
10	{Beer, Cookies}

b. What is the maximum size of frequent itemsets that can be extracted (assuming $minsup > 0$)?

The maximum size of frequent itemsets is 4 (based on {Milk, Diapers, Bread, Butter}).

c. Write an expression for the maximum number of size-3 itemsets that can be derived from this data set.

We can find the number of possible size-3 itemsets using:
$C(6,3) = \frac{6!}{3!(6-3)!} = \frac{720}{36} = 20.$

d. Find an itemset (of size 2 or larger) that has the largest support.

The itemset {Bread, Butter} has the highest support of .5.

e. Find a pair of items, a and b, such that the rules $\{a\} \rightarrow \{b\}$ and $\{b\} \rightarrow \{a\}$ have the same confidence.

For the following item counts:

Item	**Count**
Milk	5
Beer	4
Diapers	7
Bread	5
Butter	5
Cookies	4

We have the following pairs of items:

- c(Milk $\rightarrow$ Bread) = c(Bread $\rightarrow$ Milk): $\frac{3}{5} = \frac{3}{5}$
- c(Milk $\rightarrow$ Butter) = c(Butter $\rightarrow$ Milk): $\frac{3}{5} = \frac{3}{5}$
- c(Butter $\rightarrow$ Bread) = c(Bread $\rightarrow$ Butter): $\frac{5}{5} = \frac{5}{5}$
- c(Beer $\rightarrow$ Cookies) = c(Cookies $\rightarrow$ Beer): $\frac{2}{4} = \frac{2}{4}$

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	$\{b, d\}$

Table 5.22 Example of market basket transactions.

Exercise 5.9

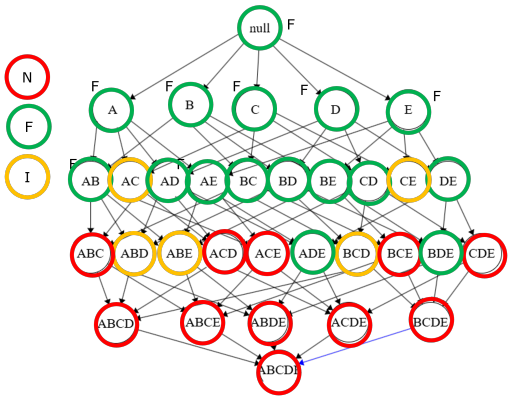
The Apriori algorithm uses a generate-and-count strategy for deriving frequent itemsets. Candidate itemsets of size k + 1 are created by joining a pair of frequent itemsets of size k (this is known as the candidate generation step). A candidate is discarded if any one of its subsets is found to be infrequent during the candidate pruning step. Suppose the Apriori algorithm is applied to the data set shown in Table 5.22 with $minsup = 30\%$, i.e., any itemset occurring in less than 3 transactions is considered to be infrequent.

a. Draw an itemset lattice representing the data set given in Table 6.24. Label each node in the lattice with the following letter(s):

   - N: If the itemset is not considered to be a candidate itemset by the Apriori algorithm. There are two reasons for an itemset not to be considered as a candidate itemset: (1) it is not generated at all during the candidate generation step, or (2) it is generated during the candidate generation step but is subsequently removed during the candidate pruning step because one of its subsets is found to be infrequent.
   - F: If the candidate itemset is found to be frequent by the Apriori algorithm.
   - I: If the candidate itemset is found to be infrequent after support counting.

b. What is the percentage of frequent itemsets (with respect to all itemsets in the lattice)?

   The percentage of frequent itemsets is 50% since 16 of the 32 itemsets in the lattice are frequent.



c. What is the pruning ratio of the Apriori algorithm on this data set? (Pruning ratio is defined as the percentage of itemsets not considered to be a candidate because (1) they are not generated during candidate generation or (2) they are pruned during the candidate pruning step.)

   The pruning ratio is 34% (11 out of 32 itemsets are labeled "N").

d. What is the false alarm rate (i.e, percentage of candidate itemsets that are found to be infrequent after performing support counting)?

   The false alarm rate is 16%, with 5 out of 32 itemsets labeled "I".

Exercise 5.10

The Apriori algorithm uses a hash tree data structure to efficiently count the support of candidate itemsets. Consider the hash tree for candidate 3-itemsets shown in Figure 6.32.

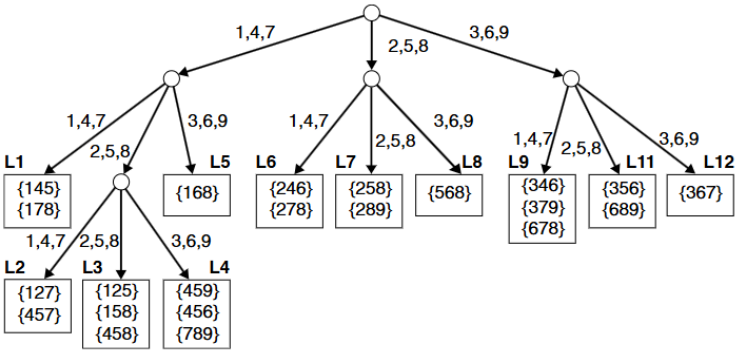


Figure 6.32. An example of a hash tree structure.

a. Given a transaction that contains items $\{1, 3, 4, 5, 8\}$, which of the hash tree leaf nodes will be visited when finding the candidates of the transaction?

   Nodes L5, L1, L3, L9 and L11 will be visited.

b. Use the visited leaf nodes in part (b) to determine the candidate itemsets that are contained in the transaction $\{1, 3, 4, 5, 8\}$.

   Candidate itemsets are: {1,4,5}, {1,5,8}, and {4,5,8}

Exercise 5.11

Consider the following set of candidate 3-itemsets:
$\{1, 2, 3\}, \{1, 2, 6\}, \{1, 3, 4\}, \{2, 3, 4\}, \{2, 4, 5\}, \{3, 4, 6\}, \{4, 5, 6\}$

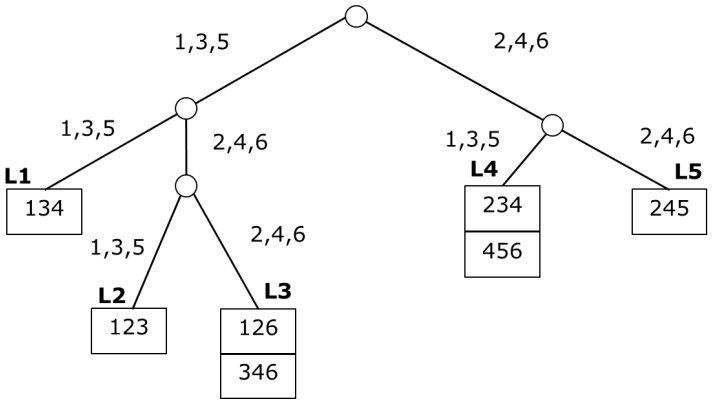
a. Construct a hash tree for the above candidate 3-itemsets. Assume the tree uses a hash function where all odd-numbered items are hashed to the left child of a node, while the even-numbered items are hashed to the right child. A candidate k-itemset is inserted into the tree by hashing on each successive item in the candidate and then following the appropriate branch of the tree according to the hash value. Once a leaf node is reached, the candidate is inserted based on one of the following conditions:

   - Condition 1: If the depth of the leaf node is equal to k (the root is assumed to be at depth 0), then the candidate is inserted regardless of the number of itemsets already stored at the node.
   - Condition 2: If the depth of the leaf node is less than k, then the candidate can be inserted as long as the number of itemsets stored at the node is less than $maxsize$. Assume $maxsize = 2$ for this question.
   - Condition 3: If the depth of the leaf node is less than k and the number of itemsets stored at the node is equal to maxsize, then the leaf node is converted into an internal node. New leaf nodes are created as children of the old leaf node. Candidate itemsets previously stored in the old leaf node are distributed to the children based on their hash values. The new candidate is also hashed to its appropriate leaf node.

b. How many leaf nodes are there in the candidate hash tree? How many internal nodes are there?

   In the hash tree we can see that there are 5 leaf nodes and 4 internal nodes.

c. Consider a transaction that contains the following items: $\{1, 2, 3, 5, 6\}$. Using the hash tree constructed in part (a), which leaf nodes will be checked



Hash Tree for Exercise 5.11

against the transaction? What are the candidate 3-itemsets contained in the transaction?

Nodes L2, L3, L4, and L1 would be checked. Itemsets {1,2,3} and {1,2,6} are contained in the transaction.