CSCI 347: Introduction to Data Mining

Introduction to Itemset Mining

ITEMSET MINING/MARKET BASKET ANALYSIS

- Suppose we observe the following transactions in a supermarket:

Transaction ID	Items
1	Toilet paper, beans, rice, milk, baby wipes,
2	Oat milk, beans, toilet paper, orange juice
3	Oat milk, milk, orange juice, toilet paper
4	Beans, toilet paper, baby wipes, diapers
5	Toilet paper, butter, baby wipes, diapers
6	Milk, toilet paper
7	Milk, rice
8	Beans, milk, rice, toilet paper
9	Milk, butter, diapers
10	Beans, rice, toilet paper

ITEMSET MINING/MARKET BASKET ANALYSIS

- Suppose we observe the following transactions in a supermarket:

Transaction ID	Items
1	Toilet paper, beans, rice, milk, baby wipes,
2	Oat milk, beans, toilet paper, orange juice
3	Oat milk, milk, orange juice, toilet paper
4	Beans, toilet paper, baby wipes, diapers
5	Toilet paper, butter, baby wipes, diapers
6	Milk, toilet paper
7	Milk, rice
8	Beans, milk, rice, toilet paper
9	Milk, butter, diapers
10	Beans, rice, toilet paper

How can we find all sets of items that are frequently purchased together?

ITEMSET MINING/MARKET BASKET ANALYSIS

- Suppose we observe the following transactions in a supermarket:

Transaction ID	Items
1	Toilet paper, beans, rice, milk, baby wipes,
2	Oat milk, beans, toilet paper, orange juice
3	Oat milk, milk, orange juice, toilet paper
4	Beans, toilet paper, baby wipes, diapers
5	Toilet paper, butter, baby wipes, diapers
6	Milk, toilet paper
7	Milk, rice
8	Beans, milk, rice, toilet paper
9	Milk, butter, diapers
10	Beans, rice, toilet paper

How can we find all sets of items that are frequently purchased together?

For example: which sets of items are purchased at least 30% of the time?

ITEMSET MINING/MARKET BASKET ANALYSIS

Transaction ID	Items
1	Toilet paper, beans, rice, milk, baby wipes,
2	Oat milk, beans, toilet paper, orange juice
3	Oat milk, milk, orange juice, toilet paper
4	Beans, toilet paper, baby wipes, diapers
5	Toilet paper, butter, baby wipes, diapers
6	Milk, toilet paper
7	Milk, rice
8	Beans, milk, rice, toilet paper
9	Milk, butter, diapers
10	Beans, rice, toilet paper

For example: which sets of items are purchased at least 30% of the time?

***Brute-force approach**: count the number of times each item, pair of items, triple of items, etc… appears, then report those that appear 3 or more times*

ITEMSET MINING/MARKET BASKET ANALYSIS

Transaction ID	Items
1	Toilet paper, beans, rice, milk, baby wipes,
2	Oat milk, beans, toilet paper, orange juice
3	Oat milk, milk, orange juice, toilet paper
4	Beans, toilet paper, baby wipes, diapers
5	Toilet paper, butter, baby wipes, diapers
6	Milk, toilet paper
7	Milk, rice
8	Beans, milk, rice, toilet paper
9	Milk, butter, diapers
10	Beans, rice, toilet paper

For example: which sets of items are purchased at least 30% of the time?

Brute-force approach: *count the number of times each item, pair of items, triple of items, etc… appears, then report those that appear 3 or more times*

How long would this take?

ITEMSET MINING/MARKET BASKET ANALYSIS

Transaction ID	Items
1	Toilet paper, beans, rice, milk, baby wipes,
2	Oat milk, beans, toilet paper, orange juice
3	Oat milk, milk, orange juice, toilet paper
4	Beans, toilet paper, baby wipes, diapers
5	Toilet paper, butter, baby wipes, diapers
6	Milk, toilet paper
7	Milk, rice
8	Beans, milk, rice, toilet paper
9	Milk, butter, diapers
10	Beans, rice, toilet paper

For example: which sets of items are purchased at least 30% of the time?

A-Priori approach: *count the number of frequent items, use those to generate frequent pairs, use those to generate frequent triplets, etc..*

This will eliminate computing frequency of sets that have no chance of being frequent.

APRIORI ALGORITHM

Which sets of items are purchased at least 30% of the time? → "minsup" = 3

Transaction ID	Items
1	Toilet paper, beans, rice, milk, baby wipes, diapers
2	Oat milk, beans, toilet paper, orange juice
3	Oat milk, milk, orange juice, toilet paper
4	Beans, toilet paper, baby wipes, diapers
5	Toilet paper, butter, baby wipes, diapers
6	Milk, toilet paper
7	Milk, rice
8	Beans, milk, rice, toilet paper
9	Milk, butter, diapers
10	Beans, rice, toilet paper

Candidate Set	Support
{Baby Wipes}	3
{Beans}	5
{Butter}	2
{Diapers}	4
{Milk}	6
{Oat Milk}	2
{Orange Juice}	2
{Rice}	4
{Toilet Paper}	8

Start with frequent item sets of size k=1

APRIORI ALGORITHM

Which sets of items are purchased at least 30% of the time? → "minsup" = 3

Transaction ID	Items
1	Toilet paper (9), beans (2), rice (8), milk (5), baby wipes (1) , diapers (4)
2	Oat milk (6), beans (2), toilet paper (9), orange juice (7)
3	Oat milk (6), milk (5), orange juice (7), toilet paper (9)
4	Beans (2), toilet paper (9), baby wipes (1), diapers (4)
5	Toilet paper (9), butter (3), baby wipes (1), diapers (4)
6	Milk (5), toilet paper (9)
7	Milk (5), rice (8)
8	Beans (2), milk (5), rice (8), toilet paper (9)
9	Milk (5), butter (3) , diapers (4)
10	Beans (2), rice (8), toilet paper (9)

Candidate Set	Support
{1}	3
{2}	5
{3}	2
{4}	4
{5}	6
{6}	2
{7}	2
{8}	4
{9}	8

Assign an ID to each item

APRIORI ALGORITHM

Frequent item sets of size 1: {1}, {2}, {4}, {5}, {8},{9}

Which sets of items are purchased at least 30% of the time? → "minsup" = 3

Transaction ID	Items
1	Toilet paper (9), beans (2), rice (8), milk (5), baby wipes (1) , diapers (4)
2	Oat milk (6), beans (2), toilet paper (9), orange juice (7)
3	Oat milk (6), milk (5), orange juice (7), toilet paper (9)
4	Beans (2), toilet paper (9), baby wipes (1), diapers (4)
5	Toilet paper (9), butter (3), baby wipes (1), diapers (4)
6	Milk (5), toilet paper (9)
7	Milk (5), rice (8)
8	Beans (2), milk (5), rice (8), toilet paper (9)
9	Milk (5), butter (3) , diapers (4)
10	Beans (2), rice (8), toilet paper (9)

Candidate Set	Support
{1}	3
{2}	5
{3}	2
{4}	4
{5}	6
{6}	2
{7}	2
{8}	4
{9}	8

Select frequent itemsets (those that appear 3 times or more)

Frequent item sets of size 1: {1}, {2}, {4}, {5}, {8},{9}

APRIORI ALGORITHM

Which sets of items are purchased at least 30% of the time? → "minsup" = 3

Transaction ID	Items
1	Toilet paper (9), beans (2), rice (8), milk (5), baby wipes (1) , diapers (4)
2	Oat milk (6), beans (2), toilet paper (9), orange juice (7)
3	Oat milk (6), milk (5), orange juice (7), toilet paper (9)
4	Beans (2), toilet paper (9), baby wipes (1), diapers (4)
5	Toilet paper (9), butter (3), baby wipes (1), diapers (4)
6	Milk (5), toilet paper (9)
7	Milk (5), rice (8)
8	Beans (2), milk (5), rice (8), toilet paper (9)
9	Milk (5), butter (3) , diapers (4)
10	Beans (2), rice (8), toilet paper (9)

Candidate Set	Support
{1,2}	2
{1,4}	3
{1,5}	1
{1,8}	1
{1,9}	3
{2, 4}	2
{2, 5}	2
{2, 8}	3
{2, 9}	5
{4, 5}	2
{4, 8}	1
{4, 9}	3
{5,8}	3
{5, 9}	4
{8,9}	3

Generate new candidates of size k+1

Frequent item sets of size 1: {1}, {2}, {4}, {5}, {8},{9}

APRIORI ALGORITHM

Frequent item sets of size 2: {1,4}, {1,9}, {2,8}, {2,9}, {4,9}, {5,8},{5,9},{8,9}

Which sets of items are purchased at least 30% of the time? → "minsup" = 3

Transaction ID	Items
1	Toilet paper (9), beans (2), rice (8), milk (5), baby wipes (1) , diapers (4)
2	Oat milk (6), beans (2), toilet paper (9), orange juice (7)
3	Oat milk (6), milk (5), orange juice (7), toilet paper (9)
4	Beans (2), toilet paper (9), baby wipes (1), diapers (4)
5	Toilet paper (9), butter (3), baby wipes (1), diapers (4)
6	Milk (5), toilet paper (9)
7	Milk (5), rice (8)
8	Beans (2), milk (5), rice (8), toilet paper (9)
9	Milk (5), butter (3) , diapers (4)
10	Beans (2), rice (8), toilet paper (9)

Candidate Set	Support
{1,2}	2
{1,4}	3
{1,5}	1
{1,8}	1
{1,9}	3
{2, 4}	2
{2, 5}	2
{2, 8}	3
{2, 9}	5
{4, 5}	2
{4, 8}	1
{4, 9}	3
{5,8}	3
{5, 9}	4
{8,9}	3

Select frequent itemsets (those that appear 3 times or more)

APRIORI ALGORITHM

Frequent itemsets of size 1: {1}, {2}, {4}, {5}, {8},{9}

Frequent itemsets of size 2: {1,4}, {1,9}, {2,8}, {2,9}, {4,9}, {5,8},{5,9},{8,9}

Which sets of items are purchased at least 30% of the time? → "minsup" = 3

Transaction ID	Items
1	Toilet paper (9), beans (2), rice (8), milk (5), baby wipes (1) , diapers (4)
2	Oat milk (6), beans (2), toilet paper (9), orange juice (7)
3	Oat milk (6), milk (5), orange juice (7), toilet paper (9)
4	Beans (2), toilet paper (9), baby wipes (1), diapers (4)
5	Toilet paper (9), butter (3), baby wipes (1), diapers (4)
6	Milk (5), toilet paper (9)
7	Milk (5), rice (8)
8	Beans (2), milk (5), rice (8), toilet paper (9)
9	Milk (5), butter (3) , diapers (4)
10	Beans (2), rice (8), toilet paper (9)

Candidate Set	Support

Generate new candidates of size k+1

APRIORI ALGORITHM

Frequent itemsets of size 1: {1}, {2}, {4}, {5}, {8},{9}

Frequent itemsets of size 2: {1,4}, {1,9}, {2,8}, {2,9}, {4,9}, {5,8},{5,9},{8,9}

Which sets of items are purchased at least 30% of the time? → "minsup" = 3

Transaction ID	Items
1	Toilet paper (9), beans (2), rice (8), milk (5), baby wipes (1) , diapers (4)
2	Oat milk (6), beans (2), toilet paper (9), orange juice (7)
3	Oat milk (6), milk (5), orange juice (7), toilet paper (9)
4	Beans (2), toilet paper (9), baby wipes (1), diapers (4)
5	Toilet paper (9), butter (3), baby wipes (1), diapers (4)
6	Milk (5), toilet paper (9)
7	Milk (5), rice (8)
8	Beans (2), milk (5), rice (8), toilet paper (9)
9	Milk (5), butter (3) , diapers (4)
10	Beans (2), rice (8), toilet paper (9)

Candidate Set	Support
{1,4, 9}	3
{2,8,9}	3
{5,8,9}	2
{2,4,9}	2
{2,5,8}	2
{4,5,9}	1

Generate new candidates of size k+1

APRIORI ALGORITHM

Frequent itemsets of size 1: {1}, {2}, {4}, {5}, {8},{9}

Frequent itemsets of size 2: {1,4}, {1,9}, {2,8}, {2,9}, {4,9}, {5,8},{5,9},{8,9}

Frequent itemsets of size 3: {1,4,9}, {2,8,9}

Which sets of items are purchased at least 30% of the time? → "minsup" = 3

Transaction ID	Items
1	Toilet paper (9), beans (2), rice (8), milk (5), baby wipes (1) , diapers (4)
2	Oat milk (6), beans (2), toilet paper (9), orange juice (7)
3	Oat milk (6), milk (5), orange juice (7), toilet paper (9)
4	Beans (2), toilet paper (9), baby wipes (1), diapers (4)
5	Toilet paper (9), butter (3), baby wipes (1), diapers (4)
6	Milk (5), toilet paper (9)
7	Milk (5), rice (8)
8	Beans (2), milk (5), rice (8), toilet paper (9)
9	Milk (5), butter (3) , diapers (4)
10	Beans (2), rice (8), toilet paper (9)

Candidate Set	Support
{1,4, 9}	3
{2,8,9}	3
{5,8,9}	2
{2,4,9}	2
{2,5,8}	2
{4,5,9}	1

Select frequent itemsets (those that appear 3 times or more)

APRIORI ALGORITHM

Which sets of items are purchased at least 30% of the time? → "minsup" = 3

Transaction ID	Items
1	Toilet paper (9), beans (2), rice (8), milk (5), baby wipes (1) , diapers (4)
2	Oat milk (6), beans (2), toilet paper (9), orange juice (7)
3	Oat milk (6), milk (5), orange juice (7), toilet paper (9)
4	Beans (2), toilet paper (9), baby wipes (1), diapers (4)
5	Toilet paper (9), butter (3), baby wipes (1), diapers (4)
6	Milk (5), toilet paper (9)
7	Milk (5), rice (8)
8	Beans (2), milk (5), rice (8), toilet paper (9)
9	Milk (5), butter (3) , diapers (4)
10	Beans (2), rice (8), toilet paper (9)

Frequent itemsets of size 1: {1}, {2}, {4}, {5}, {8},{9}

Frequent itemsets of size 2: {1,4}, {1,9}, {2,8}, {2,9}, {4,9}, {5,8},{5,9},{8,9}

Frequent itemsets of size 3: {1,4,9}, {2,8,9}

At this point, no more candidates can be generated - we've found all frequent item sets

APRIORI ALGORITHM

APriori($D, I,$ minsup) :

1. $\mathscr{F} \leftarrow \varnothing$
2. $\mathscr{C}^{(1)} = \{\varnothing\}$
3. For each $i \in I$ do:

   Add i as a child of $\varnothing$ in $\mathscr{C}^{(1)}$ with $sup(i) \leftarrow 0$
4. $k \leftarrow 1$
5. While $\mathscr{C}^{(k)} \neq \varnothing$ do:
   - ComputeSupport$(\mathscr{C}^{(k)}, D)$
   - For each leaf $X \in \mathscr{C}^{(k)}$ do:
     - If $sup(X) \geq$ minsup then $\mathscr{F} \leftarrow \mathscr{F} \cup \{(X, sup(X))\}$
     - Else remove X from $\mathscr{C}^{(k)}$
   - $\mathscr{C}^{(k+1)} \leftarrow$ ExtendPrefixTree$(\mathscr{C}^{(k)})$
   - $k \leftarrow k + 1$
6. Return $\mathscr{F}^{(k)}$

APRIORI ALGORITHM

ComputeSupport($\mathscr{C}^{(k)}, D$):

For each leaf $< t, i(t) > \in D$ do:

For each k-subset $X \in i(t)$ do:

If $X \in \mathscr{C}^{(k)}$ then $sup(X) \leftarrow sup(X) + 1$

ExtendPrefixTree($\mathscr{C}^{(k)}$):

For each leaf $X_a \in \mathscr{C}^{(k)}$ do:

For each leaf $X_b \in \text{Sibling}(X_a)$ such that $b > a$ do:

$X_{ab} \leftarrow X_a \cup X_b$

If $X_j \in \mathscr{C}^{(k)}$, for all $X_j \subset X_{ab}$ such that $|X_j| = |X_{ab}| - 1$ then:

Add X_{ab} as a child of X_a with $\sup(X_{ab}) \leftarrow 0$

If no extensions from X_a then:

Remove X_a, and all ancestors of X_a with no extensions, from $\mathscr{C}^{(k)}$

Return $\mathscr{C}^{(k)}$

ASSOCIATION RULE MINING

What rules can we generate of the form $X \rightarrow Y$, where X and Y are itemsets, with enough support and enough confidence?

Transaction ID	Items
1	Toilet paper (9), beans (2), rice (8), milk (5), baby wipes (1) , diapers (4)
2	Oat milk (6), beans (2), toilet paper (9), orange juice (7)
3	Oat milk (6), milk (5), orange juice (7), toilet paper (9)
4	Beans (2), toilet paper (9), baby wipes (1), diapers (4)
5	Toilet paper (9), butter (3), baby wipes (1), diapers (4)
6	Milk (5), toilet paper (9)
7	Milk (5), rice (8)
8	Beans (2), milk (5), rice (8), toilet paper (9)
9	Milk (5), butter (3) , diapers (4)
10	Beans (2), rice (8), toilet paper (9)

Frequent itemsets of size 1: {1}, {2}, {4}, {5}, {8},{9}

Frequent itemsets of size 2: {1,4}, {1,9}, {2,8}, {2,9}, {4,9}, {5,8},{5,9},{8,9}

Frequent itemsets of size 3: {1,4,9}, {2,8,9}

ASSOCIATION RULE MINING

What rules can we generate of the form $X \rightarrow Y$, where X and Y are itemsets, with enough support and enough confidence?

Transaction ID	Items
1	Toilet paper (9), beans (2), rice (8), milk (5), baby wipes (1) , diapers (4)
2	Oat milk (6), beans (2), toilet paper (9), orange juice (7)
3	Oat milk (6), milk (5), orange juice (7), toilet paper (9)
4	Beans (2), toilet paper (9), baby wipes (1), diapers (4)
5	Toilet paper (9), butter (3), baby wipes (1), diapers (4)
6	Milk (5), toilet paper (9)
7	Milk (5), rice (8)
8	Beans (2), milk (5), rice (8), toilet paper (9)
9	Milk (5), butter (3) , diapers (4)
10	Beans (2), rice (8), toilet paper (9)

Frequent itemsets of size 1: {1}, {2}, {4}, {5}, {8},{9}

Frequent itemsets of size 2: {1,4}, {1,9}, {2,8}, {2,9}, {4,9}, {5,8},{5,9},{8,9}

Frequent itemsets of size 3: {1,4,9}, {2,8,9}

Example: {2,8}→{9}

ASSOCIATION RULE MINING

What rules can we generate of the form $X \rightarrow Y$, where X and Y are itemsets, with enough support and enough confidence?

Transaction ID	Items
1	Toilet paper (9), beans (2), rice (8), milk (5), baby wipes (1) , diapers (4)
2	Oat milk (6), beans (2), toilet paper (9), orange juice (7)
3	Oat milk (6), milk (5), orange juice (7), toilet paper (9)
4	Beans (2), toilet paper (9), baby wipes (1), diapers (4)
5	Toilet paper (9), butter (3), baby wipes (1), diapers (4)
6	Milk (5), toilet paper (9)
7	Milk (5), rice (8)
8	Beans (2), milk (5), rice (8), toilet paper (9)
9	Milk (5), butter (3) , diapers (4)
10	Beans (2), rice (8), toilet paper (9)

Frequent itemsets of size 1: {1}, {2}, {4}, {5}, {8},{9}

Frequent itemsets of size 2: {1,4}, {1,9}, {2,8}, {2,9}, {4,9}, {5,8},{5,9},{8,9}

Frequent itemsets of size 3: {1,4,9}, {2,8,9}

Example: {2,8}→{9}

{2,8} has support 3 and {2,8,9} has support 3

ASSOCIATION RULE MINING

What rules can we generate of the form $X \rightarrow Y$, where X and Y are itemsets, with enough support and enough confidence?

Transaction ID	Items
1	Toilet paper (9), beans (2), rice (8), milk (5), baby wipes (1) , diapers (4)
2	Oat milk (6), beans (2), toilet paper (9), orange juice (7)
3	Oat milk (6), milk (5), orange juice (7), toilet paper (9)
4	Beans (2), toilet paper (9), baby wipes (1), diapers (4)
5	Toilet paper (9), butter (3), baby wipes (1), diapers (4)
6	Milk (5), toilet paper (9)
7	Milk (5), rice (8)
8	Beans (2), milk (5), rice (8), toilet paper (9)
9	Milk (5), butter (3) , diapers (4)
10	Beans (2), rice (8), toilet paper (9)

Frequent itemsets of size 1: {1}, {2}, {4}, {5}, {8},{9}

Frequent itemsets of size 2: {1,4}, {1,9}, {2,8}, {2,9}, {4,9}, {5,8},{5,9},{8,9}

Frequent itemsets of size 3: {1,4,9}, {2,8,9}

Example: {2,8}→{9}

{2,8} has support 3 and {2,8,9} has support 3

So we say the rule {2,8}→{9} has support 3 and confidence $\frac{3}{3} = 1$

ASSOCIATION RULE MINING

What rules can we generate of the form $X \rightarrow Y$, where X and Y are itemsets, with enough support and enough confidence?

Transaction ID	Items
1	Toilet paper (9), beans (2), rice (8), milk (5), baby wipes (1) , diapers (4)
2	Oat milk (6), beans (2), toilet paper (9), orange juice (7)
3	Oat milk (6), milk (5), orange juice (7), toilet paper (9)
4	Beans (2), toilet paper (9), baby wipes (1), diapers (4)
5	Toilet paper (9), butter (3), baby wipes (1), diapers (4)
6	Milk (5), toilet paper (9)
7	Milk (5), rice (8)
8	Beans (2), milk (5), rice (8), toilet paper (9)
9	Milk (5), butter (3) , diapers (4)
10	Beans (2), rice (8), toilet paper (9)

Frequent itemsets of size 1: {1}, {2}, {4}, {5}, {8},{9}

Frequent itemsets of size 2: {1,4}, {1,9}, {2,8}, {2,9}, {4,9}, {5,8},{5,9},{8,9}

Frequent itemsets of size 3: {1,4,9}, {2,8,9}

Example: {9}→{2,8}

{9} has support 8 and {2,8,9} has support 3

So we say the rule {9}→{2,8} has support 3 and confidence $\frac{3}{8} = 0.375$

ASSOCIATION RULE MINING

What rules can we generate of the form $X \rightarrow Y$, where X and Y are itemsets, with enough support and enough confidence?

Transaction ID	Items
1	Toilet paper (9), beans (2), rice (8), milk (5), baby wipes (1) , diapers (4)
2	Oat milk (6), beans (2), toilet paper (9), orange juice (7)
3	Oat milk (6), milk (5), orange juice (7), toilet paper (9)
4	Beans (2), toilet paper (9), baby wipes (1), diapers (4)
5	Toilet paper (9), butter (3), baby wipes (1), diapers (4)
6	Milk (5), toilet paper (9)
7	Milk (5), rice (8)
8	Beans (2), milk (5), rice (8), toilet paper (9)
9	Milk (5), butter (3) , diapers (4)
10	Beans (2), rice (8), toilet paper (9)

Frequent itemsets of size 1: {1}, {2}, {4}, {5}, {8},{9}

Frequent itemsets of size 2: {1,4}, {1,9}, {2,8}, {2,9}, {4,9}, {5,8},{5,9},{8,9}

Frequent itemsets of size 3: {1,4,9}, {2,8,9}

Example: {4}→{1}

{4} has support 4 and {1,4} has support 3

So we say the rule {4}→{1} has support 3 and confidence $\frac{3}{4} = 0.75$