Mining Large Scale Datasets

2024 / 2025

Frequent Itemsets and Association Rules

(Adapted from CS246@Starford.edu; http://www.mmds.org)

Association Rule Discovery

Supermarket shelf management – Market-basket model:

- **Goal:** Identify items that are bought together by sufficiently many customers
- **Approach:** Process the sales data collected with barcode scanners to find dependencies among items
- **A classic rule:**
  - If someone buys diaper and milk, then he/she is likely to buy beer
  - Don’t be surprised if you find six-packs next to diapers!

The Market-Basket Model

- A **large set** of **items**
  - e.g., things sold in a supermarket
- A **large set** of **baskets**
  - Each basket is a **small subset of items**
    - e.g., the things one customer buys on one day
- **Discover association rules:**

People who bought {x,y,z} tend to buy {v,w}

- Example application: Amazon

Input:

Basket	*Items*
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

Output:

Rules Discovered:

{Milk} --> {Coke}

{Diaper, Milk} --> {Beer}

More generally

- **A general many-to-many mapping (association) between two kinds of things**
  - But we are interested in connections among “items”, not “baskets”
- **Items and baskets are abstract:**
  - **For example:**
    - Items/baskets can be products/shopping basket
    - Items/baskets can be words/documents
    - Items/baskets can be base-pairs/genes
    - Items/baskets can be drugs/patients

Example Applications

- Related words: items are words, baskets are documents
- Plagiarism: items are documents, baskets are sentences
- Biomarkers: items are diseases and biomarkers (genes/blood proteins), baskets are sets of data about each patient
- Side-effects: items are drugs and side-effects, baskets are patients

- Note: baskets should contain small number of items; items can be in a large number of baskets

Applications (1)

- **Items** = products; **Baskets** = sets of products someone bought in one trip to the store
- **Real market baskets:** Chain stores keep TBs of data about what customers buy together
  - Tells how typical customers navigate stores, lets them position tempting items together:
    - Apocryphal story of “diapers and beer” discovery
    - Used to position potato chips between diapers and beer to enhance sales of potato chips
- **Amazon’s ‘people who bought *X* also bought *Y*’**

Applications (2)

- **Baskets** = sentences; **Items** = documents in which those sentences appear
  - Items that appear together too often could represent plagiarism
  - Notice items do not have to be "in" baskets
- **Baskets** = patients; **Items** = drugs & side-effects
  - Has been used to detect combinations of drugs that result in particular side-effects
  - **But requires extension:** Absence of an item needs to be observed as well as presence

Outline

First: Define

Frequent itemsets

Association rules:

Confidence, Support, Interestingness

Then: Algorithms for finding frequent itemsets

Finding frequent pairs

A-Priori algorithm

PCY algorithm

Frequent Itemsets

- **Simplest question:** Find sets of items that appear together "frequently" in baskets
- ***Support*** for itemset I: Number of baskets containing all items in I
  - (Often expressed as a fraction of the total number of baskets)
- Given a ***support threshold s***, then sets of items that appear in at least s baskets are called ***frequent itemsets***

TID	*Items*
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

Support of {Beer, Bread} = 2

Frequent Itemsets - Example

Support threshold = 3 baskets

B1: {beer, coke, milk}
B2: {juice, milk, pepsi}
B3: {beer, milk}
B4: {coke, juice}
B5: {beer, milk, pepsi}
B6: {beer, coke, juice, milk}
B7: {beer, coke, juice}
B8: {beer, coke}

Frequent Itemsets - Example

Support threshold = 3 baskets

B1: {beer, coke, milk}
B2: {juice, milk, pepsi}
B3: {beer, milk}
B4: {coke, juice}
B5: {beer, milk, pepsi}
B6: {beer, coke, juice, milk}
B7: {beer, coke, juice}
B8: {beer, coke}

Frequent itemsets: {beer}, {coke}, {juice}, {milk}, {beer, coke}, {beer, milk}, {coke, juice}

Define: Association Rules

- **Define: Association Rules:**
  If-then rules about the contents of baskets
- $\{i_1, i_2, \ldots, i_k\} \to j$ means: "if a basket contains all of $i_1, \ldots, i_k$ then it is ***likely*** to contain j"
- **In practice there are many rules, want to find significant/interesting ones!**
- ***Confidence*** of association rule is the probability of j given $I = \{i_1, \ldots, i_k\}$

$$\mathrm{conf}(I \to j) = \frac{\mathrm{support}(I \cup j)}{\mathrm{support}(I)}$$

Association Rules

- **Not all high-confidence rules are interesting**
  - The rule $\boldsymbol{X} \rightarrow \boldsymbol{milk}$ may have high confidence for many itemsets $\boldsymbol{X}$, because milk is just purchased very often (independent of $\boldsymbol{X}$)

Association rules: Interest

$I \rightarrow j$

if all items in I appear in a basket then it is likely that j appears in the same basket

<u>Interest</u> of a rule $I \rightarrow j$ is given by the <u>probability of j given I minus the probability of j</u>

$$interest\ I \rightarrow j\ = p(j|I) - p(j)$$

$$interest\ I \rightarrow j\ = confidence\ I \rightarrow j\ - \text{baskets containing } j / \text{ baskets}$$

high positive interest: presence of I indicates the presence of j
high negative interest: presence of I discourages the presence of j

Example: Confidence and Interest

B_1 = {m, c, b}
B_2 = {m, p, j}
B_3 = {m, b}
B_4= {c, j}
B_5 = {m, p, b}
B_6 = {m, c, b, j}
B_7 = {c, b, j}
B_8 = {b, c}

- **Association rule: {m, b} →c**
  - **Support** = 2
  - **Confidence** = 2/4 = 0.5
  - **Interest** = |0.5 – 5/8| = 1/8
    - Item ***c*** appears in 5/8 of the baskets
    - The rule is not very interesting!

Association rules: Lift

$I \rightarrow j$

if all items in I appear in a basket then it is likely that j appears in the same basket

$$Lift\ I \rightarrow j \quad = \frac{confidence(I \rightarrow j)}{P(j)} = \frac{P(I \mid j)}{P(I)\ P(j)}$$

$P\ j$

Lift (also known as the observed/expected ratio) is a measure of the degree of dependence between I and j.

A lift of 1 indicates that I and j are independent.

Association Rule Mining

- **Problem: Find all association rules with support $\geq s$ and confidence $\geq c$**
  - **Note:** Support of an association rule is the support of the entire set of items in the rule (left side + right side)
- **Hard part: Finding the frequent itemsets!**
  - If $\{i_1, i_2,\ldots, i_k\} \rightarrow \{j\}$ has high support and confidence, then both $\{i_1, i_2,\ldots, i_k\}$ and $\{i_1, i_2,\ldots,i_k, j\}$ will be "frequent"

$$\text{conf}(I \rightarrow j) = \frac{\text{support}(I \cup j)}{\text{support}(I)}$$

$$\text{conf}(I \rightarrow j) = \frac{\text{support}(I \cup j)}{\text{support}(I)}$$

Mining Association Rules

- **Step 1:** Find all frequent itemsets I
  - (we will explain this next)
- **Step 2: Rule generation**
  - For every subset A of I, generate a rule $A \rightarrow I \setminus A$
    - Since I is frequent, A is also frequent
    - **Variant 1:** Single pass to compute the rule confidence
      - confidence(***A,B***→***C,D***) = support(**A,B,C,D**) / support(**A,B**)
    - **Variant 2:**
      - **Observation:** If **A,B,C**→**D** is below confidence, then so is **A,B**→**C,D**
      - Can generate "bigger" rules from smaller ones!
  - **Output the rules above the confidence threshold**

Mining Association Rules

- This process:
- Finds combinations of items that occur frequently.
- Tries to turn those into “If…then…” rules.
- Measures how confident we are in those rules.
- Uses optimization tricks to skip bad rules and speed things up.

Example

B_1 = {m, c, b} B_2 = {m, p, j}
B_3 = {m, c, b, n} B_4= {c, j}
B_5 = {m, p, b} B_6 = {m, c, b, j}
B_7 = {c, b, j} B_8 = {b, c}

- **Support threshold $s = 3$, confidence $c = 0.75$**
- **Step 1) Find frequent itemsets:**
  - {b,m} {b,c} {c,m} {c,j} {m,c,b}
- **Step 2) Generate rules:**
  - b→m: c=4/6 b→c: c=5/6 b,c→m: c=3/5
  - m→b: c=4/5 … b,m→c: c=3/4

  b→c,m: c=3/6

Compacting the Output

- **To reduce the number of rules, we can post-process them and only output:**
  - **Maximal frequent itemsets:**
    No immediate superset is frequent
    - Gives more pruning

or

  - **Closed itemsets:**
    No immediate superset has the same support (> 0)
    - Stores not only frequent information, but exact supports/counts

Example: Maximal/Closed

	Support	Maximal(s= 3)	Closed
A	4	No	No
B	5	No	Yes
C	3	No	No
AB	4	Yes	Yes
AC	2	No	No
BC	3	Yes	Yes
ABC	2	No	Yes

Example: Maximal/Closed

	Support	Maximal(s= 3)	Closed
A	4	No	No
B	5	No	Yes
C	3	No	No
AB	4	Yes	Yes
AC	2	No	No
BC	3	Yes	Yes
ABC	2	No	No

Frequent, but superset BC also frequent

Frequent, and its only superset, ABC, not freq.

Superset BC has same support.

Its only super-set, ABC, has smaller support.