Mining Frequent Patterns, Associations and Correlations

Week 3
Team Homework Assignment #2

• Do Example 6.1. Prepare for the results of the homework assignment.
• Due date
  – beginning of the lecture on Friday February 18\textsuperscript{th}. 
Team Homework Assignment #3

• Prepare for the one-page description of your group project topic
• Prepare for presentation using slides
• Due date
  – beginning of the lecture on Friday February 11th.
<table>
<thead>
<tr>
<th>gene</th>
<th>function</th>
<th>Exp1</th>
<th>Exp2</th>
<th>Exp3</th>
<th>Exp4</th>
<th>Exp5</th>
<th>Exp6</th>
<th>Exp7</th>
<th>Exp8</th>
</tr>
</thead>
<tbody>
<tr>
<td>gene1</td>
<td>cell_cycle</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>gene2</td>
<td>cell_cycle</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>gene3</td>
<td>cell_cycle</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>gene4</td>
<td>cell_cycle</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>gene5</td>
<td>cell_cycle, apoptosis</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>gene6</td>
<td>cell_cycle, apoptosis</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>gene7</td>
<td>cell_cycle, apoptosis</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>gene8</td>
<td>cell_cycle, apoptosis</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>gene9</td>
<td>cell_cycle, apoptosis</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>gene10</td>
<td>apoptosis</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>gene11</td>
<td>apoptosis</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>gene12</td>
<td>apoptosis</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>gene13</td>
<td>apoptosis</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>gene14</td>
<td>apoptosis</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>gene15</td>
<td>apoptosis</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>gene16</td>
<td>apoptosis</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>gene17</td>
<td>apoptosis</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- cell_cycle -> [+Exp1, [+Exp2, [+Exp3, [+Exp4,
support=52.94% (9 genes)
apoptosis -> [+Exp6, [+Exp7, [+Exp8,
support=76.47% (13 genes)

http://www.cnb.uam.es/~pcarmona/assocrules/imag4.JPG
### Table 8.3
The substitution matrix of amino acids.

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>E</th>
<th>G</th>
<th>H</th>
<th>W</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>5</td>
<td>-1</td>
<td>0</td>
<td>-2</td>
<td>-3</td>
</tr>
<tr>
<td>E</td>
<td>-1</td>
<td>6</td>
<td>-3</td>
<td>0</td>
<td>-3</td>
</tr>
<tr>
<td>H</td>
<td>-2</td>
<td>0</td>
<td>-2</td>
<td>10</td>
<td>-3</td>
</tr>
<tr>
<td>P</td>
<td>-1</td>
<td>-1</td>
<td>-2</td>
<td>-2</td>
<td>-4</td>
</tr>
<tr>
<td>W</td>
<td>-3</td>
<td>-3</td>
<td>-3</td>
<td>-3</td>
<td>15</td>
</tr>
</tbody>
</table>

### Figure 8.8
Scoring two potential pairwise alignments, (a) and (b), of amino acids.
Figure 9.1 A sample graph data set.

\[ \begin{align*}
S & \quad \text{C} \quad \text{C} \quad \text{N} \\
| & \quad | \\
O & \quad S \\
\end{align*} \quad \text{(g}_1\text{)} \quad \begin{align*}
\text{C} & \quad \text{C} \quad \text{N} \quad \text{C} \quad \text{C} \quad \text{S} \quad \text{C} \quad \text{C} \\
| & \quad | \\
\text{S} & \quad \text{N} \quad \text{O} \\
\end{align*} \quad \text{(g}_2\text{)} \quad \begin{align*}
\text{S} & \quad \text{C} & \quad \text{C} \quad \text{N} \\
| & \quad | \\
\text{S} & \quad \text{N} \quad \text{O} \\
\end{align*} \quad \text{(g}_3\text{)}
\]

frequency: 2

Figure 9.2 Frequent graph.

\[ \begin{align*}
S & \quad \text{C} \quad \text{C} \quad \text{C} \quad \text{O} \\
| & \quad | \\
\text{N} & \quad \text{N} \\
\end{align*} \quad \text{frequency: 3} \]
Figure 9.14 A chemical database.

(a) caffeine

(b) thesal

(c) viagra
What Is Frequent Pattern Analysis?

- Frequent pattern: a pattern for itemsets, subsequences, substructures, etc. that occurs frequently in a data set
- First proposed by Agrawal, Imielinski, and Swami in 1993, in the context of frequent itemsets and association rule mining
Why Is Frequent Pattern Mining Important?

• Discloses an intrinsic and important property of data sets
• Forms the foundation for many essential data mining tasks and applications
  - What products were often purchased together?— Beer and diapers?
  - What are the subsequent purchases after buying a PC?
  - What kinds of DNA are sensitive to this new drug?
  - Can we automatically classify web documents?
Topics of Frequent Pattern Mining (1)

• Based on the kinds of patterns to be mined
  – Frequent itemset mining
  – Sequential pattern mining
  – Structured pattern mining
Topics of Frequent Pattern Mining (2)

• Based on the levels of abstraction involved in the rule set
  – Single-level association rules
  – Multi-level association rules
Topics of Frequent Pattern Mining (3)

• Based on the number of data dimensions involved in the rule
  – Single-dimensional association rules
  – Multi-dimensional association rules
Association Rule Mining Process

• Find all frequent itemsets
  – Join steps
  – Prune steps
• Generate “strong” association rules from the frequent itemsets
Basic Concepts of Frequent Itemsets

• Let \( I = \{I_1, I_2, \ldots, I_m\} \) be a set of items
• Let \( D, \) the task-relevant data, be a set of database transactions where each transaction \( T \) is a set of items such that \( T \subseteq I \)
• Each transaction is associated with an identifier, called \( TID. \)
• Let \( A \) be a set of items
• A transaction \( T \) is said to contain \( A \) if and only if \( A \subseteq T \)
How to Generate Frequent Itemset?

• Suppose the items in \( L_{k-1} \) are listed in an order.

• **The join step**: To find \( L_k \), a set of candidate \( k \)-itemsets, \( C_k \), is generated by joining \( L_{k-1} \) with itself. Let \( l_1 \) and \( l_2 \) be itemsets in \( L_{k-1} \). The resulting itemset formed by joining \( l_1 \) and \( l_2 \) is \( l_1[1], l_1[2], \ldots, l_1[k-2], l_1[k-1], l_2[k-1] \).

• **The prune step**: Scan data set \( D \) and compare candidate support count of \( C_k \) with minimum support count. Remove candidate itemsets that whose support count is less than minimum support count, resulting in \( L_k \).
Apriori Algorithm

- Initially, scan DB once to get frequent 1-itemset
- Generate length \((k+1)\) candidate itemsets from length \(k\) frequent itemsets
- Prune length \((k+1)\) candidate itemsets with Apriori property
  - Apriori property: All nonempty subsets of a frequent itemset must also be frequent
- Test the candidates against DB
- Terminate when no frequent or candidate set can be generated

Input:
- $D$, a database of transactions;
- $min\_sup$, the minimum support count threshold.

Output: $L$, frequent itemsets in $D$.

Method:
1. $L_1 = \text{find\_frequent\_1\_itemsets}(D)$;
2. for $k = 2$; $I_{k-1} = \pi; k++$ do
   3. $C_k = \text{apriori\_gen}(L_{k-1})$;
   4. for each transaction $t \in D$ do // scan $D$ for counts
      5. $C_t = \text{subset}(C_k, t)$; // get the subsets of $t$ that are candidates
      6. for each candidate $c \in C_t$
         7. $c.\text{count}++;
      8. }
   9. $L_k = \{ c \in C_k \mid c.\text{count} \geq min\_sup \}$
10. }
11. return $L = \cup_k L_k$;

procedure apriori\_gen($L_{k-1}$: frequent $(k-1)$-itemsets)
1. for each itemset $l_1 \in L_{k-1}$
2. for each itemset $l_2 \in L_{k-1}$
3. if $(l_1[1] = l_2[1]) \land (l_1[2] = l_2[2]) \land \ldots \land (l_1[k-2] = l_2[k-2] < l_2[k-1])$ then
4. $c = l_1 \bowtie l_2$; // join step: generate candidates
5. if has\_infrequent\_subset($c, L_{k-1}$) then
6. delete $c$; // prune step: remove unfruitful candidate
7. else add $c$ to $C_k$;
8. }
9. return $C_k$;

procedure has\_infrequent\_subset($c$: candidate $k$-itemset,
$L_{k-1}$: frequent $(k-1)$-itemsets); // use prior knowledge
1. for each $(k-1)$-subset $s$ of $c$
2. if $s \not\in L_{k-1}$ then
3. return TRUE;
4. return FALSE;
Table 5.1 Transactional data for an AllElectronics branch.
**Figure 5.2** Generation of candidate itemsets and frequent itemsets, where the minimum support count is 2.
Generating *Strong* Association Rules

• From the frequent itemsets
• For each frequent itemset $l$, generate all nonempty subset of $l$
• For every nonempty subset $s$ of $l$,
• Output the rule “$s \Rightarrow (l - s)$”
• If \( \frac{\text{support}_{\text{count}}(l)}{\text{support}_{\text{count}}(s)} \geq \text{min}_\text{conf} \), where \( \text{min}_\text{conf} \) is the minimum confidence threshold
• Rules that satisfy both a minimum support threshold and a minimum confidence threshold are called *strong*
Support

• The rule $A \Rightarrow B$ holds in the transaction set $D$ with support $s$
  
  – $support, s$, probability that a transaction contains $A$ and $B$
  
  – $support (A \Rightarrow B) = P (A \cup B)$
Confidence

• The rule $A \implies B$ has confidence $c$ in the transaction set $D$
  
  – confidence, $c$, conditional probability that a transaction having $A$ also contains $B$
  
  – confidence ($A \implies B$) = $P (B \mid A)$

\[
\text{Confidence} (A \implies B) = P(B \mid A) = \frac{\text{support} (A \cup B)}{\text{support} (A)} = \frac{\text{support\_count} (A \cup B)}{\text{support\_count} (A)}
\]
Generating Association Rules from Frequent Itemsets

• Example 5.4: Suppose the data contain the frequent itemset \( l = \{I1, I2, I5\} \). What are the association rules that can be generated from \( l \)? If the minimum confidence threshold is 70%, then which rules are strong?
  
  – \( I1 \land I2 \rightarrow I5 \), confidence = \( \frac{2}{4} = 50\% \)
  – \( I1 \land I5 \rightarrow I2 \), confidence = \( \frac{2}{2} = 100\% \)
  – \( I2 \land I5 \rightarrow I1 \), confidence = \( \frac{2}{2} = 100\% \)
  – \( I1 \rightarrow I2 \land I5 \), confidence = \( \frac{2}{6} = 33\% \)
  – \( I2 \rightarrow I1 \land I5 \), confidence = \( \frac{2}{7} = 29\% \)
  – \( I5 \rightarrow I1 \land I2 \), confidence = \( \frac{2}{2} = 100\% \)
Exercise

5.3 A database has five transactions. Let \( \text{min\_sup} = 60\% \) and \( \text{min\_conf} = 80\% \).

<table>
<thead>
<tr>
<th>TID</th>
<th>Items_bought</th>
</tr>
</thead>
<tbody>
<tr>
<td>T100</td>
<td>{M, O, N, K, E, Y}</td>
</tr>
<tr>
<td>T200</td>
<td>{D, O, N, K, E, Y}</td>
</tr>
<tr>
<td>T300</td>
<td>{M, A, K, E}</td>
</tr>
<tr>
<td>T400</td>
<td>{M, U, C, K, Y}</td>
</tr>
<tr>
<td>T500</td>
<td>{C, O, O, K, I, E}</td>
</tr>
</tbody>
</table>

(a) Find all frequent itemsets.

(b) List all of the **strong** association rules (with support \( s \) and confidence \( c \)) matching following meta-rule, where \( X \) is a variable representing customers, and \( \text{item}_i \) denotes variables representing items (e.g., “A”, “B”, etc.):

\[
\forall x \in \text{transaction}, \text{buys}(X, \text{item}_1) \land \text{buys}(X, \text{item}_2) \Rightarrow \text{buys}(X, \text{item}_3) \quad [s, c]
\]
Challenges of Frequent Pattern Mining

• Challenges
  – Multiple scans of transaction database
  – Huge number of candidates
  – Tedious workload of support counting for candidates

• Improving Apriori
  – Reduce passes of transaction database scans
  – Shrink number of candidates
  – Facilitate support counting of candidates
Advanced Methods for Mining Frequent Itemsets

• Mining frequent itemsets without candidate generation
  – Frequent-pattern growth (FP-growth—Han, Pei & Yin @SIGMOD’00)

• Mining frequent itemsets using vertical data format
  – Vertical data format approach (ECLAT—Zaki @IEEE-TKDE’00)
Mining Various Kinds of Association Rules

• Mining multilevel association rules

• Mining multidimensional association rules
Mining Multilevel Association Rules (1)

• Data mining systems should provide capabilities for mining association rules at multiple levels of abstraction

• Exploration of shared multi-level mining (Agrawal & Srikant@VLB’95, Han & Fu@VLDB’95)
Mining Multilevel Association Rules (2)

• For each level, any algorithm for discovering frequent itemsets may be used, such as Apriori or its variations
  – Using uniform minimum support for all levels (referred to as uniform support)
  – Using reduced minimum support at lower levels (referred to as reduced support)
  – Using item or group-based minimum support (referred to as group-based support)
<table>
<thead>
<tr>
<th>TID</th>
<th>items purchased</th>
</tr>
</thead>
<tbody>
<tr>
<td>T100</td>
<td>IBM-ThinkPad-T40/2373, HP-Photosmart-7660</td>
</tr>
<tr>
<td>T200</td>
<td>Microsoft-Office-Professional-2003, Microsoft-Plus!-Digital-Media</td>
</tr>
<tr>
<td>T300</td>
<td>Logitech-MX700-Cordless-Mouse, Fellowes-Wrist-Rest</td>
</tr>
<tr>
<td>T400</td>
<td>Dell-Dimension-XPS, Canon-PowerShot-S400</td>
</tr>
<tr>
<td>T500</td>
<td>IBM-ThinkPad-R40/P4M, Symantec-Norton-Antivirus-2003</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

Table 5.6 Task-relevant data D.
Figure 5.10 A concept hierarchy for AllElectronics computer items.
Figure 5.11 Multilevel mining with uniform support.
Figure 5.12  Multilevel mining with reduced support.
Multilevel mining with group-based support.
Mining Multilevel Association Rules (3)

• Side effect
  – The generation of many redundant rules across multiple levels of abstractions due to the ancestor relationships among items
  – \(\text{buys}(X, \text{“laptop computer”}) \Rightarrow \text{buys}(X, \text{“HP printer”})\)
    
    \[\text{support} = 8\%, \text{confidence} = 70\%\]
  – \(\text{buys}(X, \text{“IBM laptop computer”}) \Rightarrow \text{buys}(X, \text{“HP printer”})\)
    
    \[\text{support} = 2\%, \text{confidence} = 72\%\]
Mining Multidimensional Association Rules

• Single-dimensional rules:
  \[ \text{buys}(X, \text{“milk”}) \Rightarrow \text{buys}(X, \text{“bread”}) \]

• Multi-dimensional rules: \( \geq 2 \) dimensions or predicates
  – Inter-dimension assoc. rules (\textit{no repeated predicates})
    \[ \text{age}(X, \text{“19-25”}) \land \text{occupation}(X, \text{“student”}) \Rightarrow \text{buys}(X, \text{“coke”}) \]
  – Hybrid-dimension assoc. rules (\textit{repeated predicates})
    \[ \text{age}(X, \text{“19-25”}) \land \text{buys}(X, \text{“popcorn”}) \Rightarrow \text{buys}(X, \text{“coke”}) \]
Mining Quantitative Association Rules

• ARCS (Association Rule Clustering System): Cluster adjacent rules to form general association rules using a 2-D grid
  – \(\text{age}(X,\text{"34-35"}) \land \text{income}(X,\text{"31-50K"}) \Rightarrow \text{buys}(X,\text{"high resolution TV"})\)
  – Proposed by Lent, Swami and Widom ICDE’97
Figure 5.14  A 2-D grid for tuples representing customers who purchase high-definition TVs.

age(X,34) ∧ income(X,”31-40K”) ⇒ buys(X,”high resolution TV”)
age(X,35) ∧ income(X,”31-40K”) ⇒ buys(X,”high resolution TV”)
age(X,34) ∧ income(X,”41-50K”) ⇒ buys(X,”high resolution TV”)
age(X,35) ∧ income(X,”40-50K”) ⇒ buys(X,”high resolution TV”)
age(X,”34-35”) ∧ income(X,”31-50K”) ⇒ buys(X,”high resolution TV”)
Strong Rules Are Not Necessarily Interesting (1)

• Suppose we are interested in analyzing transaction in AllElectronics with respect to the purchase of computer games and videos. Let \textit{game} refer to the transactions containing computer games, and \textit{video} refer to those containing videos. Of the 10,000 transactions analyzed, the data show that 6,000 of the customer transactions included computer games, while 7,500 included videos, and 4,000 included both computer games and videos.
Strong Rules Are Not Necessarily Interesting (2)

• Suppose that a data mining program for discovering association rules is run on the data, using a minimum support of, say, 30% and a minimum confidence of 60%. Is the following association rule is strong?

• \( \text{buys}(X, \text{"computer games"}) \implies \text{buys}(X, \text{"videos"}) \)
Strong Rules Are Not Necessarily Interesting (3)

• The rule above is misleading because the probability of purchasing videos is 75%.

• It does not measure the real string of the correlation and implication computer games and videos.

• How can we tell which strong association rules are really interesting?
Correlation Analysis

• Correlation measure

\[ A \Rightarrow B \{support, confidence, correlation\} \]

• Correlation metrics
  – lift
  – chi-square
  – all_confidence
  – cosine measure
Correlation Analysis Using Lift

\[
\text{lift} = \frac{P(A \cup B)}{P(A)P(B)} = \frac{P(B|A)}{P(B)} = \frac{\text{conf}(A \Rightarrow B)}{\text{sup}(B)}
\]

- If the resulting value is greater than 1, then \( A \) and \( B \) are positively correlated, meaning that the occurrence of one implies the occurrence of the other.
- If the resulting value is equal to 1, then \( A \) and \( B \) are independent and there is no correlation between them.
- If the resulting value is less than 1, then the occurrence of \( A \) is negatively correlated with the occurrence of \( B \).
Correlation Analysis Using Lift

Table 5.7 A 2 X 2 contingency table summarizing the transactions with respect to game and video purchases.

<table>
<thead>
<tr>
<th></th>
<th>game</th>
<th>game</th>
<th>Σ_{row}</th>
</tr>
</thead>
<tbody>
<tr>
<td>video</td>
<td>4,000</td>
<td>3,500</td>
<td>7,500</td>
</tr>
<tr>
<td>video</td>
<td>2,000</td>
<td>500</td>
<td>2,500</td>
</tr>
<tr>
<td>Σ_{col}</td>
<td>6,000</td>
<td>4,000</td>
<td>10,000</td>
</tr>
</tbody>
</table>

\[ \text{lift} = \frac{P(A \cup B)}{P(A)P(B)} = \frac{P(B | A)}{P(B)} = \frac{\text{conf}(A \Rightarrow B)}{\text{sup}(B)} \]

\[ P(\{\text{game}\}) = \]
\[ P(\{\text{video}\}) = \]
\[ P(\{\text{game,video}\}) = \]
Correlation Analysis Using Chi-square

\[ \chi^2 = \sum \frac{(Observed - Expected)^2}{Expected} \]

\[ \chi^2 = \sum_{i=1}^{c} \sum_{j=1}^{r} \frac{(o_{ij} - e_{ij})^2}{e_{ij}} \]

\[ e_{ij} = \frac{\text{count}(A = a_i) \times \text{count}(B = b_j)}{N} \]

- The larger the \( \chi^2 \) value, the more likely the variables are related.
- If the observed value of the cell is less than the expected value, two variables associated with the cell is negatively correlated.
Correlation Analysis Using Chi-square

Table 5.8 The above contingency table, now shown with the expected value.

<table>
<thead>
<tr>
<th></th>
<th>game</th>
<th>game</th>
<th>$\Sigma_{row}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>video</td>
<td>4,000 (4,500)</td>
<td>3,500 (3,000)</td>
<td>7,500</td>
</tr>
<tr>
<td>video</td>
<td>2,000 (1,500)</td>
<td>500 (1,000)</td>
<td>2,500</td>
</tr>
<tr>
<td>$\Sigma_{col}$</td>
<td>6,000</td>
<td>4,000</td>
<td>10,000</td>
</tr>
</tbody>
</table>

\[ \chi^2 = \sum \frac{(Observed - Expected)^2}{Expected} = \]