Look for accompanying R code on the course web site.
Topics

- Definition
- Mining Frequent Itemsets (APRIORI)
- Concise Itemset Representation
- Alternative Methods to Find Frequent Itemsets
- Association Rule Generation
- Support Distribution
- Pattern Evaluation
Association Rule Mining

- Given a set of transactions, find rules that will predict the occurrence of an item based on the occurrences of other items in the transaction.

### Market-Basket transactions

<table>
<thead>
<tr>
<th>TID</th>
<th>Items</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Bread, Milk</td>
</tr>
<tr>
<td>2</td>
<td>Bread, Diaper, Beer, Eggs</td>
</tr>
<tr>
<td>3</td>
<td>Milk, Diaper, Beer, Coke</td>
</tr>
<tr>
<td>4</td>
<td>Bread, Milk, Diaper, Beer</td>
</tr>
<tr>
<td>5</td>
<td>Bread, Milk, Diaper, Coke</td>
</tr>
</tbody>
</table>

### Example of Association Rules

\[
\{\text{Diaper}\} \rightarrow \{\text{Beer}\}, \\
\{\text{Milk, Bread}\} \rightarrow \{\text{Eggs, Coke}\}, \\
\{\text{Beer, Bread}\} \rightarrow \{\text{Milk}\}, \\
\]

Implication means co-occurrence, not causality!
Definition: Frequent Itemset

- **Itemset**
  - A collection of one or more items
    - Example: \{Milk, Bread, Diaper\}
  - \(k\)-itemset
    - An itemset that contains \(k\) items

- **Support count (\(\sigma\))**
  - Frequency of occurrence of an itemset
  - E.g. \(\sigma(\{\text{Milk, Bread, Diaper}\}) = 2\)

- **Support**
  - Fraction of transactions that contain an itemset
  - E.g. \(s(\{\text{Milk, Bread, Diaper}\}) = \frac{\sigma(\{\text{Milk, Bread, Diaper}\})}{|T|} = 2/5\)

- **Frequent Itemset**
  - An itemset whose support is greater than or equal to a \textit{minsup} threshold

### Table: Itemsets

<table>
<thead>
<tr>
<th>TID</th>
<th>Items</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Bread, Milk</td>
</tr>
<tr>
<td>2</td>
<td>Bread, Diaper, Beer, Eggs</td>
</tr>
<tr>
<td>3</td>
<td>Milk, Diaper, Beer, Coke</td>
</tr>
<tr>
<td>4</td>
<td>Bread, Milk, Diaper, Beer</td>
</tr>
<tr>
<td>5</td>
<td>Bread, Milk, Diaper, Coke</td>
</tr>
</tbody>
</table>

\[
s(X) = \frac{\sigma(X)}{|T|}
\]
Definition: Association Rule

• **Association Rule**
  - An implication expression of the form \( X \rightarrow Y \), where \( X \) and \( Y \) are itemsets
  - Example: \( \{Milk, Bread\} \rightarrow \{Diaper\} \)

• **Rule Evaluation Metrics**
  - Support (s)
    - Fraction of transactions that contain both \( X \) and \( Y \)
  - Confidence (c)
    - Measures how often items in \( Y \) appear in transactions that contain \( X \)

\[
c(X \rightarrow Y) = \frac{s(X \cup Y)}{\sigma(X)} = \frac{s(X \cup Y)}{s(X)}
\]
Topics

• Definition
• Mining Frequent Itemsets (APRIORI)
• Concise Itemset Representation
• Alternative Methods to Find Frequent Itemsets
• Association Rule Generation
• Support Distribution
• Pattern Evaluation
Association Rule Mining Task

- Given a set of transactions $T$, the goal of association rule mining is to find all rules having
  - support $\geq \text{minsup}$ threshold
  - confidence $\geq \text{minconf}$ threshold

- Brute-force approach:
  - List all possible association rules
  - Compute the support and confidence for each rule
  - Prune rules that fail the $\text{minsup}$ and $\text{minconf}$ thresholds

$\Rightarrow$ Computationally prohibitive!
Mining Association Rules

Example of Rules:

\{\text{Milk,Diaper}\} \rightarrow \{\text{Beer}\} \ (s=0.4, \ c=0.67)
\{\text{Milk,Beer}\} \rightarrow \{\text{Diaper}\} \ (s=0.4, \ c=1.0)
\{\text{Diaper,Beer}\} \rightarrow \{\text{Milk}\} \ (s=0.4, \ c=0.67)
\{\text{Beer}\} \rightarrow \{\text{Milk,Diaper}\} \ (s=0.4, \ c=0.67)
\{\text{Diaper}\} \rightarrow \{\text{Milk,Beer}\} \ (s=0.4, \ c=0.5)
\{\text{Milk}\} \rightarrow \{\text{Diaper,Beer}\} \ (s=0.4, \ c=0.5)

Observations:

• All the above rules are binary partitions of the same itemset: \{\text{Milk, Diaper, Beer}\}
• Rules originating from the same itemset have identical support but can have different confidence
• Thus, we may decouple the support and confidence requirements
Mining Association Rules

- Two-step approach:
  1. Frequent Itemset Generation
     - Generate all itemsets whose support $\geq \text{minsup}$
  2. Rule Generation
     - Generate high confidence rules from each frequent itemset, where each rule is a binary partitioning of a frequent itemset

- Frequent itemset generation is still computationally expensive
Given $d$ items, there are $2^d$ possible candidate itemsets.
Reducing Number of Candidates

• Apriori principle:
  - If an itemset is frequent, then all of its subsets must also be frequent

• Apriori principle holds due to the following property of the support measure:

\[ \forall X, Y : (X \subseteq Y) \Rightarrow s(X) \geq s(Y) \]

- Support of an itemset never exceeds the support of its subsets
- This is known as the anti-monotone property of support
Illustrating Apriori Principle

Figure 6.4. An illustration of support-based pruning. If \( \{a, b\} \) is infrequent, then all supersets of \( \{a, b\} \) are infrequent.
Illustrating Apriori Principle

**Items (1-itemsets)**

<table>
<thead>
<tr>
<th>Item</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bread</td>
<td>4</td>
</tr>
<tr>
<td>Coke</td>
<td>2</td>
</tr>
<tr>
<td>Milk</td>
<td>4</td>
</tr>
<tr>
<td>Beer</td>
<td>3</td>
</tr>
<tr>
<td>Diaper</td>
<td>4</td>
</tr>
<tr>
<td>Eggs</td>
<td>1</td>
</tr>
</tbody>
</table>

**Pairs (2-itemsets)**

(No need to generate candidates involving Coke or Eggs)

<table>
<thead>
<tr>
<th>Itemset</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>{Bread, Milk}</td>
<td>3</td>
</tr>
<tr>
<td>{Bread, Beer}</td>
<td>2</td>
</tr>
<tr>
<td>{Bread, Diaper}</td>
<td>3</td>
</tr>
<tr>
<td>{Milk, Beer}</td>
<td>2</td>
</tr>
<tr>
<td>{Milk, Diaper}</td>
<td>3</td>
</tr>
<tr>
<td>{Beer, Diaper}</td>
<td>3</td>
</tr>
</tbody>
</table>

**Triplets (3-itemsets)**

<table>
<thead>
<tr>
<th>Itemset</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>{Bread, Milk, Diaper}</td>
<td>3</td>
</tr>
</tbody>
</table>

Minimum Support = 3

If every subset is considered, $\binom{6}{1} + \binom{6}{2} + \binom{6}{3} = 41$

With support-based pruning, $6 + 6 + 1 = 13$
Apriori Algorithm

• Method:
  – Let $k=1$
  – Generate frequent itemsets of length 1
  – Repeat until no new frequent itemsets are identified
    ◆ Generate length $(k+1)$ candidate itemsets from length $k$ frequent itemsets
    ◆ Prune candidate itemsets containing subsets of length $k$ that are infrequent
    ◆ Count the support of each candidate by scanning the DB
    ◆ Eliminate candidates that are infrequent, leaving only those that are frequent
Factors Affecting Complexity

• **Choice of minimum support threshold**
  - lowering support threshold results in more frequent itemsets
  - this may increase number of candidates and max length of frequent itemsets

• **Dimensionality (number of items) of the data set**
  - more space is needed to store support count of each item
  - if number of frequent items also increases, both computation and I/O costs may also increase

• **Size of database**
  - since Apriori makes multiple passes, run time of algorithm may increase with number of transactions

• **Average transaction width**
  - transaction width increases with denser data sets
  - This may increase max length of frequent itemsets and traversals of hash tree (number of subsets in a transaction increases with its width)
Topics

- Definition
- Mining Frequent Itemsets (APRIORI)
- Concise Itemset Representation
- Alternative Methods to Find Frequent Itemsets
- Association Rule Generation
- Support Distribution
- Pattern Evaluation
Maximal Frequent Itemset

An itemset is maximal frequent if none of its immediate supersets is frequent.

Figure 6.16. Maximal frequent itemset.
Closed Itemset

- An itemset is closed if none of its immediate supersets has the same support as the itemset (can only have smaller support - see APRIORI principle)

<table>
<thead>
<tr>
<th>TID</th>
<th>Items</th>
<th>Support</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>{A,B}</td>
<td>4</td>
</tr>
<tr>
<td>2</td>
<td>{B,C,D}</td>
<td>5</td>
</tr>
<tr>
<td>3</td>
<td>{A,B,C,D}</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>{A,B,D}</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>{A,B,C,D}</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Itemset</th>
<th>Support</th>
</tr>
</thead>
<tbody>
<tr>
<td>{A}</td>
<td>4</td>
</tr>
<tr>
<td>{B}</td>
<td>5</td>
</tr>
<tr>
<td>{C}</td>
<td>3</td>
</tr>
<tr>
<td>{D}</td>
<td>4</td>
</tr>
<tr>
<td>{A,B}</td>
<td>4</td>
</tr>
<tr>
<td>{A,C}</td>
<td>2</td>
</tr>
<tr>
<td>{A,D}</td>
<td>3</td>
</tr>
<tr>
<td>{B,C}</td>
<td>3</td>
</tr>
<tr>
<td>{B,D}</td>
<td>4</td>
</tr>
<tr>
<td>{C,D}</td>
<td>3</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Itemset</th>
<th>Support</th>
</tr>
</thead>
<tbody>
<tr>
<td>{A,B,C}</td>
<td>2</td>
</tr>
<tr>
<td>{A,B,D}</td>
<td>3</td>
</tr>
<tr>
<td>{A,C,D}</td>
<td>2</td>
</tr>
<tr>
<td>{B,C,D}</td>
<td>3</td>
</tr>
<tr>
<td>{A,B,C,D}</td>
<td>2</td>
</tr>
</tbody>
</table>
Maximal vs Closed Itemsets

<table>
<thead>
<tr>
<th>TID</th>
<th>Items</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>ABC</td>
</tr>
<tr>
<td>2</td>
<td>ABCD</td>
</tr>
<tr>
<td>3</td>
<td>BCE</td>
</tr>
<tr>
<td>4</td>
<td>ACDE</td>
</tr>
<tr>
<td>5</td>
<td>DE</td>
</tr>
</tbody>
</table>

Not supported by any transactions

Transaction Ids

null
Maximal vs Closed Frequent Itemsets

Minimum support = 2

Closed but not maximal

Closed and maximal

# Closed = 9
# Maximal = 4
Maximal vs Closed Itemsets

Figure 6.18. Relationships among frequent, maximal frequent, and closed frequent itemsets.
Topics

• Definition
• Mining Frequent Itemsets (APRIORI)
• Concise Itemset Representation
• Alternative Methods to Find Frequent Itemsets
• Association Rule Generation
• Support Distribution
• Pattern Evaluation
Alternative Methods for Frequent Itemset Generation

- Traversal of Itemset Lattice
  - Equivalent Classes

(a) Prefix tree.  
(b) Suffix tree.
Alternative Methods for Frequent Itemset Generation

- Representation of Database: horizontal vs vertical data layout

**Figure 6.23.** Horizontal and vertical data format.
Alternative Algorithms

• **FP-growth**
  - Use a compressed representation of the database using an **FP-tree**
  - Once an FP-tree has been constructed, it uses a recursive divide-and-conquer approach to mine the frequent itemsets

• **ECLAT**
  - Store transaction id-lists (vertical data layout).
  - Performs fast tid-list intersection (bit-wise XOR) to count itemset frequencies
Topics

- Definition
- Mining Frequent Itemsets (APRIORI)
- Concise Itemset Representation
- Alternative Methods to Find Frequent Itemsets
- Association Rule Generation
- Support Distribution
- Pattern Evaluation
Rule Generation

• Given a frequent itemset $L$, find all non-empty subsets $X = f \subset L$ and $Y = L - f$ such that $X \rightarrow Y$ satisfies the minimum confidence requirement

$$c(X \rightarrow Y) = \frac{\sigma(X \cup Y)}{\sigma(X)}$$

- If $\{A, B, C, D\}$ is a frequent itemset, candidate rules:

  - $ABC \rightarrow D$,
  - $ABD \rightarrow C$,
  - $ACD \rightarrow B$,
  - $BCD \rightarrow A$,
  - $A \rightarrow BCD$,
  - $B \rightarrow ACD$,
  - $C \rightarrow ABD$,
  - $D \rightarrow ABC$
  - $AB \rightarrow CD$,
  - $AC \rightarrow BD$,
  - $AD \rightarrow BC$,
  - $BC \rightarrow AD$,
  - $BD \rightarrow AC$,
  - $CD \rightarrow AB$

If $|L| = k$, then there are $2^k - 2$ candidate association rules (ignoring $L \rightarrow \emptyset$ and $\emptyset \rightarrow L$)
Rule Generation

• How to efficiently generate rules from frequent itemsets?
  - In general, confidence does not have an anti-monotone property
    
    \[ c(ABC \rightarrow D) \text{ can be larger or smaller than } c(AB \rightarrow D) \]

  - But confidence of rules generated from the same itemset has an anti-monotone property
  - e.g., \( L = \{A,B,C,D\} \):

    \[ c(ABC \rightarrow D) \geq c(AB \rightarrow CD) \geq c(A \rightarrow BCD) \]

• Confidence is anti-monotone w.r.t. number of items on the RHS of the rule
Rule Generation for Apriori Algorithm
Topics

• Definition
• Mining Frequent Itemsets (APRIORI)
• Concise Itemset Representation
• Alternative Methods to Find Frequent Itemsets
• Association Rule Generation
• Support Distribution
• Pattern Evaluation
Effect of Support Distribution

- Many real data sets have skewed support distribution
Effect of Support Distribution

• How to set the appropriate \textit{minsup} threshold?
  - If \textit{minsup} is set too high, we could miss itemsets involving interesting rare items (e.g., expensive products)

  - If \textit{minsup} is set too low, it is computationally expensive and the number of itemsets is very large

• Using a single minimum support threshold may not be effective
Topics

- Definition
- Mining Frequent Itemsets (APRIORI)
- Concise Itemset Representation
- Alternative Methods to Find Frequent Itemsets
- Association Rule Generation
- Support Distribution
- Pattern Evaluation
Pattern Evaluation

- Association rule algorithms tend to produce too many rules. Many of them are
  - uninteresting or
  - redundant

- Interestingness measures can be used to prune/rank the derived patterns

- A rule \{A,B,C\} → \{D\} can be considered redundant if \{A,B\} → \{D\} has the same or higher confidence.
Application of Interestingness Measure

Interestingness Measures

Preprocessed Data

Selected Data

Data

Mining

Preprocessing

Postprocessing

Patterns

Knowledge
Computing Interestingness Measure

• Given a rule $X \rightarrow Y$, information needed to compute rule interestingness can be obtained from a contingency table.

Contingency table for $X \rightarrow Y$

<table>
<thead>
<tr>
<th></th>
<th>$Y$</th>
<th>$\bar{Y}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$X$</td>
<td>$f_{11}$</td>
<td>$f_{10}$</td>
</tr>
<tr>
<td>$\bar{X}$</td>
<td>$f_{01}$</td>
<td>$f_{00}$</td>
</tr>
<tr>
<td></td>
<td>$f_{+1}$</td>
<td>$f_{+0}$</td>
</tr>
</tbody>
</table>

$f_{11}$: support of $X$ and $Y$
$f_{10}$: support of $X$ and not $Y$
$f_{01}$: support of not $X$ and $Y$
$f_{00}$: support of not $X$ and not $Y$

Used to define various measures
e.g., support, confidence, lift, Gini,
J-measure, etc.

$$sup\{X,Y\} = \frac{f_{11}}{|T|}$$ estimates $P(X,Y)$

$$conf(X \rightarrow Y) = \frac{f_{11}}{f_{1+}}$$ estimates $P(Y \mid X)$
Drawback of Confidence

|       | Coffee | Coffee
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Tea</td>
<td>15</td>
<td>5</td>
</tr>
<tr>
<td>Tea</td>
<td>75</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>90</td>
<td>10</td>
</tr>
</tbody>
</table>

Association Rule: Tea $\rightarrow$ Coffee

Support = $P(\text{Coffee}, \text{Tea}) = \frac{15}{100} = 0.15$

Confidence = $P(\text{Coffee} \mid \text{Tea}) = \frac{15}{20} = 0.75$

but $P(\text{Coffee}) = \frac{90}{100} = 0.9$

$\Rightarrow$ Although confidence is high, rule is misleading

$\Rightarrow P(\overline{\text{Coffee}} \mid \text{Tea}) = \frac{75}{80} = 0.9375$
Statistical Independence

- Population of 1000 students
  - 600 students know how to swim (S)
  - 700 students know how to bike (B)
  - 450 students know how to swim and bike (S,B)

- \( P(S, B) = \frac{450}{1000} = 0.45 \) (observed joint prob.)
- \( P(S) \times P(B) = 0.6 \times 0.7 = 0.42 \) (expected under indep.)

- \( P(S, B) = P(S) \times P(B) \Rightarrow \text{Statistical independence} \)
- \( P(S, B) > P(S) \times P(B) \Rightarrow \text{Positively correlated} \)
- \( P(S, B) < P(S) \times P(B) \Rightarrow \text{Negatively correlated} \)
Statistical-based Measures

• Measures that take statistical dependence into account for rule: $X \rightarrow Y$

$Lift = Interest = \frac{P(Y|X)}{P(Y)} = \frac{P(X,Y)}{P(X)P(Y)}$

$PS = P(X,Y) - P(X)P(Y)$

$\Phi = \frac{P(X,Y) - P(X)P(Y)}{\sqrt{P(X)[1-P(X)]P(Y)[1-P(Y)]}}$
**Example: Lift/Interest**

<table>
<thead>
<tr>
<th></th>
<th>Coffee</th>
<th>Coffee</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tea</td>
<td>15</td>
<td>5</td>
</tr>
<tr>
<td>Tea</td>
<td>75</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>90</td>
<td>10</td>
</tr>
</tbody>
</table>

**Association Rule: Tea → Coffee**

\[
\text{Conf(Tea → Coffee)} = \frac{P(\text{Coffee, Tea})}{P(\text{Tea})} = \frac{.15}{.2} = 0.75
\]

but \( P(\text{Coffee}) = 0.9 \)

\[
\Rightarrow \text{Lift(Tea → Coffee)} = \frac{P(\text{Coffee, Tea})}{(P(\text{Coffee})P(\text{Tea}))} = \frac{.15}{(.9 \times .2)} = 0.8333
\]

**Note:** Lift < 1, therefore Coffee and Tea are negatively associated
There are lots of measures proposed in the literature. Some measures are good for certain applications, but not for others.

<table>
<thead>
<tr>
<th>#</th>
<th>Measure</th>
<th>Formula</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$\phi$-coefficient</td>
<td>$\frac{P(A,B) - P(A)P(B)}{\sqrt{P(A)P(B)(1-P(A))(1-P(B))}}$</td>
</tr>
<tr>
<td>2</td>
<td>Goodman-Kruskal’s ($\lambda$)</td>
<td>$\frac{P(A,B)P(\bar{A},\bar{B}) - P(\bar{A},\bar{B})P(A,B)}{\sqrt{P(A)P(\bar{A})P(B)P(\bar{B})}}$</td>
</tr>
<tr>
<td>3</td>
<td>Odds ratio ($\alpha$)</td>
<td>$\frac{P(A,B)P(\bar{A},\bar{B}) - P(\bar{A},\bar{B})P(A,B)}{\sqrt{P(A)P(\bar{A})P(B)P(\bar{B})}}$</td>
</tr>
<tr>
<td>4</td>
<td>Yule’s $Q$</td>
<td>$\frac{(P(A,B)P(\bar{A},\bar{B}) - P(\bar{A},\bar{B})P(A,B))}{\sqrt{P(A)P(\bar{A})P(B)P(\bar{B})}}$</td>
</tr>
<tr>
<td>5</td>
<td>Yule’s $Y$</td>
<td>$\frac{P(A,B)P(\bar{A},\bar{B}) - P(\bar{A},\bar{B})P(A,B)}{\sqrt{P(A)P(\bar{A})P(B)P(\bar{B})}}$</td>
</tr>
<tr>
<td>6</td>
<td>Kappa ($\kappa$)</td>
<td>$\frac{P(A,B)P(\bar{A},\bar{B}) - P(\bar{A},\bar{B})P(A,B)}{\sqrt{P(A)P(\bar{A})P(B)P(\bar{B})}}$</td>
</tr>
<tr>
<td>7</td>
<td>Mutual Information ($M$)</td>
<td>$\frac{(P(A,B)P(\bar{A},\bar{B}) - P(\bar{A},\bar{B})P(A,B))}{\sqrt{P(A)P(\bar{A})P(B)P(\bar{B})}}$</td>
</tr>
<tr>
<td>8</td>
<td>J-Measure ($J$)</td>
<td>$\max\left{ P(A,B) \log\left( \frac{P(B</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$\max\left{ P(A,B) \log\left( \frac{P(B</td>
</tr>
<tr>
<td>9</td>
<td>Gini index ($G$)</td>
<td>$\max\left{ P(A) [P(B</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$\left. - P(B)^2 - P(\bar{B})^2 \right}$</td>
</tr>
<tr>
<td>10</td>
<td>Support ($s$)</td>
<td>$\max(P(B</td>
</tr>
<tr>
<td>11</td>
<td>Confidence ($c$)</td>
<td>$\max\left( \frac{NP(A,B)+1}{NP(A)+2}, \frac{NP(A,B)+1}{NP(B)+2} \right) \right.$</td>
</tr>
<tr>
<td>12</td>
<td>Laplace ($L$)</td>
<td>$\max\left( \frac{P(A,B)}{P(A)}, \frac{P(A,B)}{P(\bar{A})} \right) \right.$</td>
</tr>
<tr>
<td>13</td>
<td>Conviction ($V$)</td>
<td>$\max\left( \frac{P(A,B)}{P(A,B)}, \frac{P(A,B)}{P(B)} \right) \right.$</td>
</tr>
<tr>
<td>14</td>
<td>Interest ($I$)</td>
<td>$\max\left( P(\bar{A}) - P(A</td>
</tr>
<tr>
<td>15</td>
<td>cosine ($IS$)</td>
<td>$\max\left( P(\bar{A}) - P(A</td>
</tr>
<tr>
<td>16</td>
<td>Piatetksy-Shapiro’s ($PS$)</td>
<td>$\max\left( P(\bar{A}) - P(A</td>
</tr>
<tr>
<td>17</td>
<td>Certainty factor ($F$)</td>
<td>$\max\left( P(\bar{A}) - P(A</td>
</tr>
<tr>
<td>18</td>
<td>Added Value ($AV$)</td>
<td>$\max\left( P(\bar{A}) - P(A</td>
</tr>
<tr>
<td>19</td>
<td>Collective strength ($S$)</td>
<td>$\max\left( P(\bar{A}) - P(A</td>
</tr>
<tr>
<td>20</td>
<td>Jaccard ($\zeta$)</td>
<td>$\sqrt{P(A,B)} \max(P(B</td>
</tr>
<tr>
<td>21</td>
<td>Klosgen ($K$)</td>
<td>$\sqrt{P(A,B)} \max(P(B</td>
</tr>
</tbody>
</table>
Comparing Different Measures

10 examples of contingency tables:

<table>
<thead>
<tr>
<th>Example</th>
<th>$f_{11}$</th>
<th>$f_{10}$</th>
<th>$f_{01}$</th>
<th>$f_{00}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>E1</td>
<td>8123</td>
<td>83</td>
<td>424</td>
<td>1370</td>
</tr>
<tr>
<td>E2</td>
<td>8330</td>
<td>2</td>
<td>622</td>
<td>1046</td>
</tr>
<tr>
<td>E3</td>
<td>9481</td>
<td>94</td>
<td>127</td>
<td>298</td>
</tr>
<tr>
<td>E4</td>
<td>3954</td>
<td>3080</td>
<td>5</td>
<td>2961</td>
</tr>
<tr>
<td>E5</td>
<td>2886</td>
<td>1363</td>
<td>1320</td>
<td>4431</td>
</tr>
<tr>
<td>E6</td>
<td>1500</td>
<td>2000</td>
<td>500</td>
<td>6000</td>
</tr>
<tr>
<td>E7</td>
<td>4000</td>
<td>2000</td>
<td>1000</td>
<td>3000</td>
</tr>
<tr>
<td>E8</td>
<td>4000</td>
<td>2000</td>
<td>2000</td>
<td>2000</td>
</tr>
<tr>
<td>E9</td>
<td>1720</td>
<td>7121</td>
<td>5</td>
<td>1154</td>
</tr>
<tr>
<td>E10</td>
<td>61</td>
<td>2483</td>
<td>4</td>
<td>7452</td>
</tr>
</tbody>
</table>

Rankings of contingency tables using various measures:

<table>
<thead>
<tr>
<th>#</th>
<th>$\phi$</th>
<th>$\lambda$</th>
<th>$\alpha$</th>
<th>$Q$</th>
<th>$Y$</th>
<th>$\kappa$</th>
<th>$M$</th>
<th>$J$</th>
<th>$G$</th>
<th>$s$</th>
<th>$c$</th>
<th>$L$</th>
<th>$V$</th>
<th>$I$</th>
<th>$IS$</th>
<th>$PS$</th>
<th>$F$</th>
<th>$AV$</th>
<th>$S$</th>
<th>$\zeta$</th>
<th>$K$</th>
</tr>
</thead>
<tbody>
<tr>
<td>E1</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>3</td>
<td>5</td>
<td>5</td>
<td>4</td>
<td>6</td>
<td>2</td>
<td>2</td>
<td>4</td>
<td>6</td>
<td>1</td>
<td>2</td>
<td>5</td>
</tr>
<tr>
<td>E2</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>8</td>
<td>3</td>
<td>5</td>
<td>1</td>
<td>8</td>
<td>2</td>
<td>3</td>
<td>6</td>
</tr>
<tr>
<td>E3</td>
<td>3</td>
<td>3</td>
<td>4</td>
<td>4</td>
<td>3</td>
<td>3</td>
<td>8</td>
<td>7</td>
<td>1</td>
<td>4</td>
<td>4</td>
<td>6</td>
<td>10</td>
<td>1</td>
<td>8</td>
<td>6</td>
<td>10</td>
<td>3</td>
<td>1</td>
<td>10</td>
<td></td>
</tr>
<tr>
<td>E4</td>
<td>4</td>
<td>7</td>
<td>2</td>
<td>2</td>
<td>5</td>
<td>4</td>
<td>1</td>
<td>3</td>
<td>6</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>4</td>
<td>4</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>E5</td>
<td>5</td>
<td>4</td>
<td>8</td>
<td>8</td>
<td>4</td>
<td>7</td>
<td>5</td>
<td>4</td>
<td>7</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>3</td>
<td>6</td>
<td>3</td>
<td>9</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>E6</td>
<td>6</td>
<td>6</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>6</td>
<td>4</td>
<td>6</td>
<td>9</td>
<td>8</td>
<td>8</td>
<td>7</td>
<td>2</td>
<td>8</td>
<td>6</td>
<td>7</td>
<td>2</td>
<td>7</td>
<td>8</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>E7</td>
<td>7</td>
<td>5</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>6</td>
<td>6</td>
<td>5</td>
<td>4</td>
<td>7</td>
<td>7</td>
<td>8</td>
<td>5</td>
<td>5</td>
<td>4</td>
<td>8</td>
<td>5</td>
<td>6</td>
<td>4</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>E8</td>
<td>8</td>
<td>9</td>
<td>10</td>
<td>10</td>
<td>8</td>
<td>10</td>
<td>10</td>
<td>8</td>
<td>4</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>9</td>
<td>7</td>
<td>7</td>
<td>10</td>
<td>9</td>
<td>8</td>
<td>7</td>
<td>9</td>
<td></td>
</tr>
<tr>
<td>E9</td>
<td>9</td>
<td>9</td>
<td>5</td>
<td>5</td>
<td>9</td>
<td>9</td>
<td>7</td>
<td>9</td>
<td>8</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>7</td>
<td>9</td>
<td>9</td>
<td>3</td>
<td>7</td>
<td>9</td>
<td>9</td>
<td>8</td>
<td></td>
</tr>
<tr>
<td>E10</td>
<td>10</td>
<td>8</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>10</td>
<td>5</td>
<td>9</td>
<td>10</td>
<td>6</td>
<td>6</td>
<td>5</td>
<td>1</td>
<td>10</td>
<td>10</td>
<td>5</td>
<td>1</td>
<td>10</td>
<td>5</td>
<td>5</td>
<td></td>
</tr>
</tbody>
</table>

support & confidence

lift

Example $f_{11}$ $f_{10}$ $f_{01}$ $f_{00}$
Support-based Pruning

• Most of the association rule mining algorithms use support measure to prune rules and itemsets

• Study effect of support pruning on correlation of itemsets
  - Generate 10,000 random contingency tables
  - Compute support and pairwise correlation for each table
  - Apply support-based pruning and examine the tables that are removed
Support-based pruning eliminates mostly negatively correlated itemsets
Subjective Interestingness Measure

- **Objective measure:**
  - Rank patterns based on statistics computed from data
  - E.g., 21 measures of association (support, confidence, Laplace, Gini, mutual information, Jaccard, etc).

- **Subjective measure:**
  - Rank patterns according to user’s interpretation
    - A pattern is subjectively interesting if it **contradicts the expectation** of a user (Silberschatz & Tuzhilin)
    - A pattern is subjectively interesting if it is **actionable** (Silberschatz & Tuzhilin)
Interestingness via Unexpectedness

• Need to model expectation of users (domain knowledge)

- Pattern expected to be frequent
- Pattern expected to be infrequent
- Pattern found to be frequent
- Pattern found to be infrequent

+ Expected Patterns
- Unexpected Patterns

• Need to combine expectation of users with evidence from data
  (i.e., extracted patterns)
Applications for Association Rules

• **Market Basket Analysis**
  Marketing & Retail. E.g., frequent itemsets give information about "other customer who bought this item also bought X"

• **Exploratory Data Analysis**
  Find correlation in very large (= many transactions), high-dimensional (= many items) data

• **Intrusion Detection**
  Rules with low support but very high lift

• **Build Rule-based Classifiers**
  Class association rules (CARs)