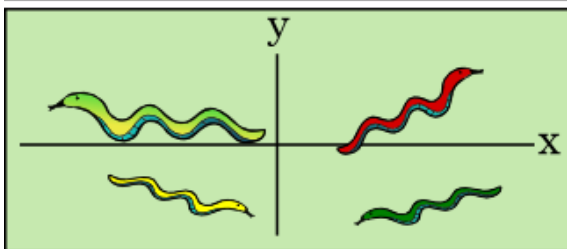
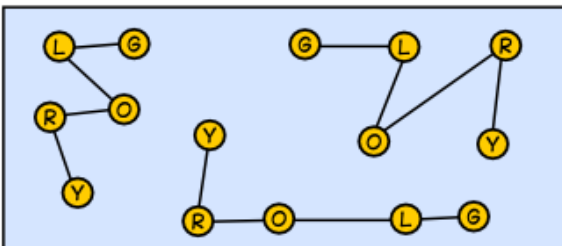


CAN YOU FIGURE OUT THESE MOVIE TITLES?



$$P(\text{Monday} \cap \text{Tuesday}) \\ = P(\text{Monday})P(\text{Tuesday})$$

$$\frac{1}{n} \sum_{i=1}^n \text{[Cartoon Girl Icon]}_i$$



12.874752 km

$$\begin{bmatrix} \cos \theta & -\sin \theta \\ \sin \theta & \cos \theta \end{bmatrix}$$

$\mathbb{D} = \{d : d \text{ is a dream}\}$
 \mathbb{D} HAS TWO OPERATIONS, NAMELY ADDITION AND MULTIPLICATION, SATISFYING THE CONDITIONS THAT MULTIPLICATION IS DISTRIBUTIVE OVER ADDITION, THAT THE SET IS A GROUP UNDER ADDITION, AND THAT THE ELEMENTS WITH THE EXCEPTION OF THE ADDITIVE IDENTITY FORM A GROUP UNDER MULTIPLICATION.

$$\alpha \wedge \omega$$

[13]

$$F = \{x : x \text{ is a fear}\}$$

$$\sum_{x \in F} x$$

Data Mining: Association Analysis: Part II

Laura Brown

Some slides adapted from G. Piatetsky-Shapiro;
Han, Kamber, & Pei; Tan, Steinbach, & Kumar

Outline

- Previously
 - Basic Terminology
 - Frequent Itemset Mining
 - Apriori
 - ECLAT
 - FPGrowth
- Today
 - Rule Generation
 - Rule Evaluation

Association Rule Generation

Mining Association Rules

Two-step approach

1. Frequent Itemset Generation

generate all itemsets whose support $\geq \textit{minsup}$

2. Rule Generation

generate high confidence rules from each frequent itemset, where each rule is a binary partitioning of a frequent itemset

Association Rules

- An **association rule** is an expression of the form

$$X \rightarrow Y$$

where X and Y are disjoint itemsets. Denote $X \cup Y$ as XY

- The **support** of a rule is the number of transactions in which X and Y co-occur

$$s = \text{sup}(X \rightarrow Y) = \text{sup}(XY)$$

- The **relative support** of a rule is the fraction of transactions in which X and Y co-occur

$$\text{rsup}(X \rightarrow Y) = \text{sup}(XY) / |\mathbf{D}| = P(X \wedge Y)$$

- The **confidence** of a rule is the conditional probability that a transaction contains Y given that it contains X

$$c = \text{conf}(X \rightarrow Y) = P(Y \mid X) = P(X \wedge Y) / P(X) = \text{sup}(XY) / \text{sup}(X)$$

From Freq Itemsets to Association Rules

- Given a frequent set {A, B, E}, what are possible association rules?

- {A} -> {B, E}
- {A, B} -> {E}
- {A, E} -> {B}
- {B} -> {A, E}
- {B, E} -> {A}
- {E} -> {A, B}

TID	List of items
1	A, B, E
2	B, D
3	B, C
4	A, B, D
5	A, C
6	B, C
7	A, C
8	A, B, C, E
9	A, B, C

The **confidence** of a rule is the conditional probability that a transaction contains Y given that it contains X

$$c = \text{conf}(X \rightarrow Y) = P(Y \mid X) = P(X \wedge Y) / P(X) = \text{sup}(XY) / \text{sup}(X)$$

From Freq Itemsets to Association Rules

- Given a frequent set {A, B, E}, what are possible association rules?

- $\{A\} \rightarrow \{B, E\}$, $c = 2/6 = 0.33$
- $\{A, B\} \rightarrow \{E\}$, $c = 2/4 = 0.50$
- $\{A, E\} \rightarrow \{B\}$, $c = 2/2 = 1.00$
- $\{B\} \rightarrow \{A, E\}$, $c = 2/7 = 0.28$
- $\{B, E\} \rightarrow \{A\}$, $c = 2/2 = 1.00$
- $\{E\} \rightarrow \{A, B\}$, $c = 2/2 = 1.00$

TID	List of items
1	A, B, E
2	B, D
3	B, C
4	A, B, D
5	A, C
6	B, C
7	A, C
8	A, B, C, E
9	A, B, C

- Each rule is binary partition of same itemset
 - have identical support, but different confidence
- We want to generate **strong** rules, that have both minimum support and minimum confidence

Rule Generation

- Given a frequent itemset L , find all non-empty subsets $f \subset L$ such that $f \rightarrow \{L - f\}$ satisfies the minimum confidence requirement
 - if $\{A, B, C, D\}$ is a frequent itemset, candidate rules:

ABC \rightarrow D	ABD \rightarrow C	ACD \rightarrow B	BDC \rightarrow A
D \rightarrow ABC	C \rightarrow ABD	B \rightarrow ACD	A \rightarrow ABC
AB \rightarrow CD	AC \rightarrow BD	AD \rightarrow BC	BC \rightarrow AD
BD \rightarrow AC	CD \rightarrow AB		

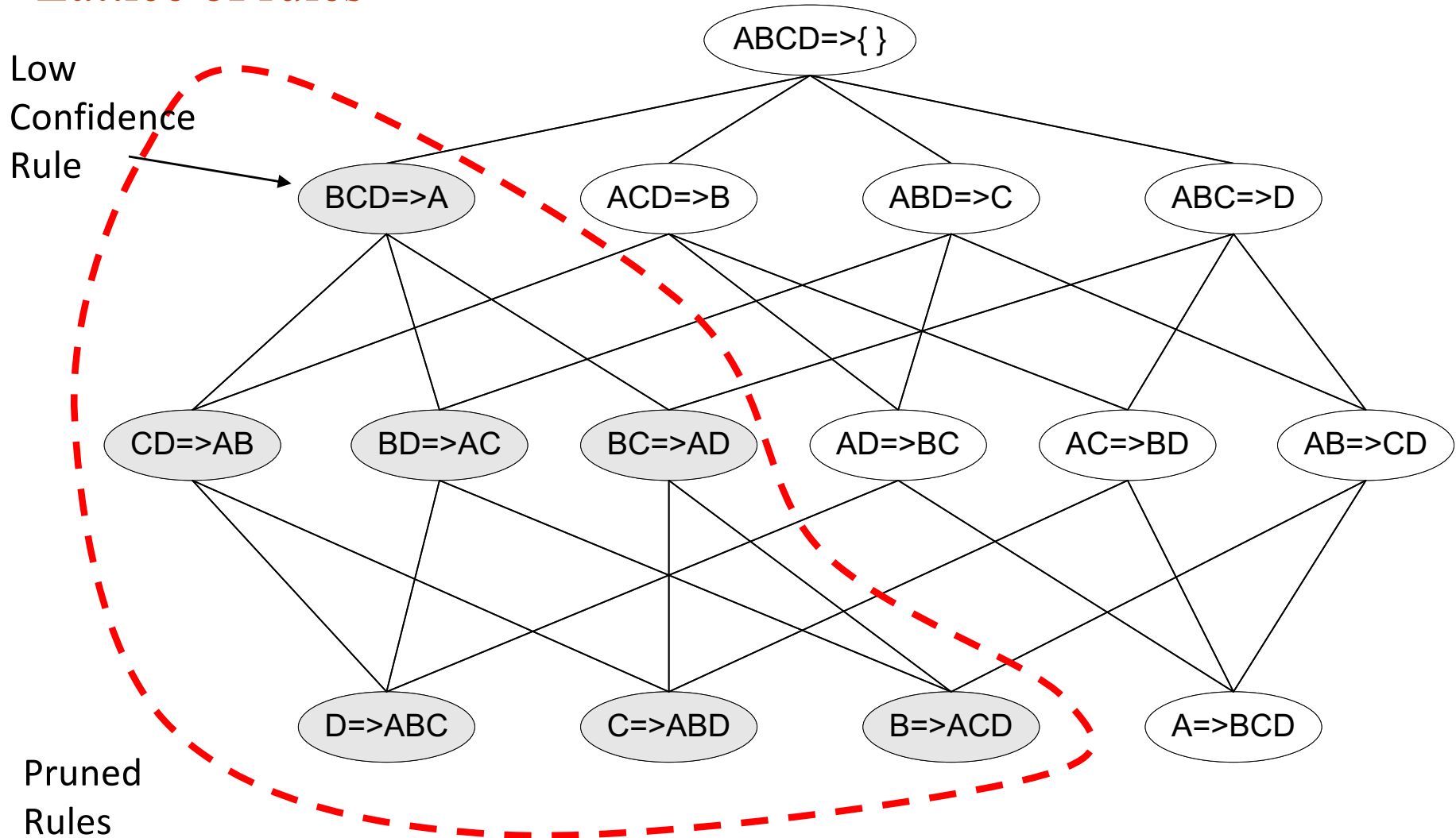
- If $|L| = k$, then there are $2^k - 2$ candidate rules (ignoring, $L \rightarrow \{\}$ and $\{\} \rightarrow L$)

Efficient Rule Generation

- How to efficiently generate rules from frequent analysis?
 - In general, confidence does not have an anti-monotone property
 $c(ABC \rightarrow D)$ can be larger or smaller than $c(AB \rightarrow D)$
- But confidence of rules generated from the same itemset has an anti-monotone property
 - For example, $L = \{A, B, C, D\}$
 $c(ABC \rightarrow D) \geq c(AB \rightarrow CD) \geq c(A \rightarrow BCD)$
 - confidence is anti-monotone w.r.t. inclusion on the RHS of the rule

Rule Generation

Lattice of rules



Example: Rule Generation

- $L = \{ \{I1\}, \{I2\}, \{I3\}, \{I4\}, \{I5\}, \{I1, I2\}, \{I1, I3\}, \{I1, I5\}, \{I2, I3\}, \{I2, I4\}, \{I2, I5\}, \{I1, I2, I3\}, \{I1, I2, I5\} \}$

- Look at $\{I1, I2, I5\}$

- *minconf* is 70%

- R1: $I1 \wedge I2 \rightarrow I5$

$$conf = \frac{s(\{I1, I2, I5\})}{s(\{I1, I2\})} = \frac{2}{4} = 50\%$$

- R2: $I1 \wedge I5 \rightarrow I2$

$$conf = \frac{s(\{I1, I2, I5\})}{s(\{I1, I5\})} = \frac{2}{2} = 100\%$$

- R3: $I2 \wedge I5 \rightarrow I1$

$$conf = \frac{s(\{I1, I2, I5\})}{s(\{I2, I5\})} = \frac{2}{2} = 100\%$$

TID	List of Items
T100	I1, I2, I5
T100	I2, I4
T100	I2, I3
T100	I1, I2, I4
T100	I1, I3
T100	I2, I3
T100	I1, I3
T100	I1, I2, I3, I5
T100	I1, I2, I3

Example: Rule Generation

- R4: I1 -> I2 ^ I5

$$conf = \frac{s(\{I1, I2, I5\})}{s(\{I1\})} = \frac{2}{6} = 33\%$$

- R5: I2 -> I1 ^ I5

$$conf = \frac{s(\{I1, I2, I5\})}{s(\{I2\})} = \frac{2}{7} = 29\%$$

- R6: I5 -> I1 ^ I2

$$conf = \frac{s(\{I1, I2, I5\})}{s(\{I5\})} = \frac{2}{2} = 100\%$$

- R2, R3, and R6 are selected

TID	List of Items
T100	I1, I2, I5
T100	I2, I4
T100	I2, I3
T100	I1, I2, I4
T100	I1, I3
T100	I2, I3
T100	I1, I3
T100	I1, I2, I3, I5
T100	I1, I2, I3

Rule Evaluation

Ch. 12 of book

Evaluation of Patterns

- A number of methods may be used to generate frequent itemsets and association rules
- Methods tend to produce too many rules
 - rules may be redundant or uninteresting
 - Ex. $\{A, B, C\} \rightarrow \{D\}$ and $\{A, B\} \rightarrow \{D\}$ are **redundant** if have same support and confidence
- Other “**interestingness**” measures can be used to prune or rank association rules

Example: Support and Confidence

Dataset

Tid	Items
1	<i>ABDE</i>
2	<i>BCE</i>
3	<i>ABDE</i>
4	<i>ABCE</i>
5	<i>ABCDE</i>
6	<i>BCD</i>

$$sup(X \rightarrow Y) = sup(XY)$$

$$conf(X \rightarrow Y) = sup(XY) / sup(X)$$

Rule	<i>sup</i>	<i>conf</i>
<i>A -> E</i>	4	1.00
<i>E -> A</i>	4	0.8
<i>B -> E</i>	5	0.83
<i>E -> B</i>	5	1.00
<i>E -> BC</i>	3	0.6
<i>BC -> E</i>	3	0.75

Frequent itemsets with minsup = 3

<i>sup</i>	<i>rsup</i>	Itemsets
3	0.5	<i>ABD, ABDE, AD, ADE, BCE, BDE, CE, DE</i>
4	0.67	<i>A, C, D, AB, ABE, AE, BC, BD</i>
5	0.83	<i>E, BE</i>
6	1.0	<i>B</i>

Other Measures: Lift

- Lift is the ratio of the observed joint probability of X and Y to the expected joint probability if they were statistically independent

$$\begin{aligned} lift(X \rightarrow Y) &= \frac{P(XY)}{P(X) \cdot P(Y)} \\ &= \frac{rsup(XY)}{rsup(X) \cdot rsup(Y)} \\ &= \frac{conf(X \rightarrow Y)}{rsup(Y)} \end{aligned}$$

$$lift(X \rightarrow Y) = \frac{P(Y | X)}{P(Y)}$$

Example: Lift

Dataset

Tid	Items
1	<i>ABDE</i>
2	<i>BCE</i>
3	<i>ABDE</i>
4	<i>ABCE</i>
5	<i>ABCDE</i>
6	<i>BCD</i>

Itemset *ABCE* has support = 2

Rule	<i>lift</i>
<i>AE -> BC</i>	0.75
<i>CE -> AB</i>	1.00
<i>BE -> AC</i>	1.20

$$lift(X \rightarrow Y) = \frac{rsup(XY)}{rsup(X) \cdot rsup(Y)} = \frac{conf(X \rightarrow Y)}{rsup(Y)}$$

$$lift(AE \rightarrow BC) = \frac{rsup(ABCE)}{rsup(AE) \cdot rsup(BC)} = \frac{2/6}{4/6 \cdot 4/6} = 0.75$$

Example: Support, Confidence, and Lift

Dataset

Tid	Items
1	<i>ABDE</i>
2	<i>BCE</i>
3	<i>ABDE</i>
4	<i>ABCE</i>
5	<i>ABCDE</i>
6	<i>BCD</i>

Rule		<i>rsup</i>	<i>conf</i>	<i>lift</i>
<i>E</i>	\longrightarrow <i>AC</i>	0.33	0.40	1.20
<i>E</i>	\longrightarrow <i>AB</i>	0.67	0.80	1.20
<i>B</i>	\longrightarrow <i>E</i>	0.83	0.83	1.00

Other Measures: Leverage

- Leverage measures the difference between the observed and expected joint probability of XY assuming that X and Y are independent

$$\begin{aligned} \text{leverage}(X \rightarrow Y) &= P(XY) - P(X) \cdot P(Y) \\ &= r_{sup}(XY) - r_{sup}(X) \cdot r_{sup}(Y) \end{aligned}$$

- Leverage gives an “absolute” measure of how surprising a rule is and can be used with lift

Example: Leverage

Dataset

Tid	Items
1	<i>ABDE</i>
2	<i>BCE</i>
3	<i>ABDE</i>
4	<i>ABCE</i>
5	<i>ABCDE</i>
6	<i>BCD</i>

Rule	<i>rsup</i>	<i>lift</i>	<i>leverage</i>
<i>ACD -> E</i>	0.17	1.20	0.03
<i>AC -> E</i>	0.33	1.20	0.06
<i>AB -> D</i>	0.50	1.12	0.06
<i>A -> E</i>	0.67	1.20	0.11

$$\begin{aligned} \text{leverage}(X \rightarrow Y) &= P(XY) - P(X) \cdot P(Y) \\ &= \text{rsup}(XY) - \text{rsup}(X) \cdot \text{rsup}(Y) \end{aligned}$$

$$\text{leverage}(ACD \rightarrow E) = P(ACDE) - P(ACD) \cdot P(E) = \frac{1}{6} - \frac{1}{6} \cdot \frac{5}{6} = \frac{1}{36}$$

Other Measures: Jaccard

- Jaccard coefficient measures the similarity between sets, or the similarity between the *tidsets* of X and Y

$$\begin{aligned} Jaccard(X \rightarrow Y) &= \frac{|t(X) \cap t(Y)|}{|t(X) \cup t(Y)|} \\ &= \frac{\sup(XY)}{\sup(X) + \sup(Y) - \sup(XY)} \\ &= \frac{P(XY)}{P(X) + P(Y) - P(XY)} \end{aligned}$$

Example: Jaccard

Dataset

Tid	Items
1	<i>ABDE</i>
2	<i>BCE</i>
3	<i>ABDE</i>
4	<i>ABCE</i>
5	<i>ABCDE</i>
6	<i>BCD</i>

Rule	<i>rsup</i>	<i>lift</i>	<i>jaccard</i>
<i>A -> C</i>	0.33	0.75	0.33
<i>A -> E</i>	0.67	1.20	0.80
<i>A -> B</i>	0.67	1.00	0.67

$$Jaccard(X \rightarrow Y) = \frac{|t(X) \cap t(Y)|}{|t(X) \cup t(Y)|} = \frac{\sup(XY)}{\sup(X) + \sup(Y) - \sup(XY)}$$

$$jaccard(A \rightarrow C) = \frac{\sup(AC)}{\sup(A) + \sup(B) - \sup(AC)} = \frac{2}{4 + 4 - 2} = 0.33$$

Contingency Table: $X \rightarrow Y$

	Y	$\neg Y$	
X	f_{11}	f_{10}	f_{1+}
$\neg X$	f_{01}	f_{00}	f_{0+}
	f_{+1}	f_{+0}	$ T $

f_{11} : support of X and Y

f_{10} : support of X and $\neg Y$

f_{01} : support of $\neg X$ and Y

f_{00} : support of $\neg X$ and $\neg Y$

	Y	$\neg Y$	
X	$sup(XY)$	$sup(X\neg Y)$	$sup(X)$
$\neg X$	$sup(\neg XY)$	$sup(\neg X\neg Y)$	$sup(\neg X)$
	$sup(Y)$	$sup(\neg Y)$	$ D $

Other Measures: Conviction

- Conviction measures the expected error of the rule. That is, how often X occurs in a transaction when Y does not.

$$\text{conv}(X \rightarrow Y) = \frac{P(X) \cdot P(\neg Y)}{P(X \neg Y)} = \frac{1}{\text{lift}(X \rightarrow \neg Y)}$$

$$\text{conv}(X \rightarrow Y) = \frac{P(X) \cdot P(\neg Y)}{P(X) - P(XY)} = \frac{P(\neg Y)}{1 - P(XY)/P(X)} = \frac{1 - \text{rsup}(Y)}{1 - \text{conf}(X \rightarrow Y)}$$

- Conviction is a measure of the strength of a rule with respect to the complement of the consequent

Example: Conviction

Dataset

Tid	Items
1	<i>ABDE</i>
2	<i>BCE</i>
3	<i>ABDE</i>
4	<i>ABCE</i>
5	<i>ABCDE</i>
6	<i>BCD</i>

Rule	<i>rsup</i>	<i>conf</i>	<i>lift</i>	<i>conv</i>
<i>A -> DE</i>	0.50	0.75	1.50	2.00
<i>DE -> A</i>	0.50	1.00	1.50	∞
<i>E -> C</i>	0.50	0.60	0.90	0.83
<i>C -> E</i>	0.50	0.75	0.90	0.68

$$conv(X \rightarrow Y) = \frac{P(X) \cdot P(\neg Y)}{P(X \neg Y)} = \frac{1}{lift(X \rightarrow \neg Y)}$$

$$conv(X \rightarrow Y) = \frac{P(X) \cdot P(\neg Y)}{P(X) - P(XY)} = \frac{P(\neg Y)}{1 - P(XY)/P(X)} = \frac{1 - rsup(Y)}{1 - conf(X \rightarrow Y)}$$

$$conv(A \rightarrow DE) = \frac{1 - rsup(DE)}{1 - conf(A \rightarrow DE)} = 2.00$$

Other Measures: Odds Ratio

- Odds ratio uses all four entries in the contingency tables. Divide the dataset into two groups of transactions, those with X and those $\neg X$

$$\text{odds}(Y | X) = \frac{P(XY)/P(X)}{P(X\neg Y)/P(X)} = \frac{P(XY)}{P(X\neg Y)}$$

$$\text{odds}(Y | \neg X) = \frac{P(\neg XY)/P(\neg X)}{P(\neg X\neg Y)/P(\neg X)} = \frac{P(\neg XY)}{P(\neg X\neg Y)}$$

- The odds ratio is the ratio of these two odds

$$\begin{aligned}\text{oddsratio}(X \rightarrow Y) &= \frac{\text{odds}(Y | X)}{\text{odds}(Y | \neg X)} = \frac{P(XY) \cdot P(\neg X\neg Y)}{P(X\neg Y) \cdot P(\neg XY)} \\ &= \frac{\text{sup}(XY) \cdot \text{sup}(\neg X\neg Y)}{\text{sup}(X\neg Y) \cdot \text{sup}(\neg XY)}\end{aligned}$$

Example: Odds Ratio

Dataset

Tid	Items
1	<i>ABDE</i>
2	<i>BCE</i>
3	<i>ABDE</i>
4	<i>ABCE</i>
5	<i>ABCDE</i>
6	<i>BCD</i>

Compare odds ratio for $C \rightarrow A$ and $D \rightarrow A$

	<i>C</i>	$\neg C$
<i>A</i>	2	2
$\neg A$	2	0

	<i>D</i>	$\neg D$
<i>A</i>	3	1
$\neg A$	1	1

$$\text{oddsratio}(C \rightarrow A) = \frac{\text{sup}(AC) \cdot \text{sup}(\neg A \neg C)}{\text{sup}(A \neg C) \cdot \text{sup}(\neg AC)} = \frac{2 \times 0}{2 \times 2} = \mathbf{0}$$

$$\text{oddsratio}(D \rightarrow A) = \frac{\text{sup}(AD) \cdot \text{sup}(\neg A \neg D)}{\text{sup}(A \neg D) \cdot \text{sup}(\neg AD)} = \frac{3 \times 1}{1 \times 1} = 3$$

Evaluate Rules: Example

Example: Measures (test your knowledge)

Tid	Items
1	ABC
2	BDE
3	ACDE
4	ACE
5	ABCE
6	BCD

Rule	<i>rsup</i>	<i>conf</i>	<i>lift</i>	<i>lev.</i>	<i>jac.</i>	<i>conv</i>	<i>OR</i>
$B \rightarrow C$							
$A \rightarrow E$							
$E \rightarrow A$							
$A \rightarrow C$							
$AE \rightarrow C$							
$AC \rightarrow E$							

Examine *rsup*, *conf*, *lift*, *leverage*, *jaccard*, *conviction*, and *odds ratio* for the rules above

Solution at End of Slides

Interestingness Measures

Interestingness Measures

Many different measures exists

Which measure is best?

- domain dependent

#	Measure	Formula
1	ϕ -coefficient	$\frac{P(A,B) - P(A)P(B)}{\sqrt{P(A)P(B)(1-P(A))(1-P(B))}}$
2	Goodman-Kruskal's (λ)	$\frac{\sum_j \max_k P(A_j, B_k) + \sum_k \max_j P(A_j, B_k) - \max_j P(A_j) - \max_k P(B_k)}{2 - \max_j P(A_j) - \max_k P(B_k)}$
3	Odds ratio (α)	$\frac{P(A,B)P(\bar{A},\bar{B})}{P(A,\bar{B})P(\bar{A},B)}$
4	Yule's Q	$\frac{P(A,B)P(\bar{A}\bar{B}) - P(A,\bar{B})P(\bar{A},B)}{P(A,B)P(\bar{A}\bar{B}) + P(A,\bar{B})P(\bar{A},B)} = \frac{\alpha - 1}{\alpha + 1}$
5	Yule's Y	$\frac{\sqrt{P(A,B)P(\bar{A}\bar{B})} - \sqrt{P(A,\bar{B})P(\bar{A},B)}}{\sqrt{P(A,B)P(\bar{A}\bar{B})} + \sqrt{P(A,\bar{B})P(\bar{A},B)}} = \frac{\sqrt{\alpha} - 1}{\sqrt{\alpha} + 1}$
6	Kappa (κ)	$\frac{P(A,B) + P(\bar{A},\bar{B}) - P(A)P(B) - P(\bar{A})P(\bar{B})}{1 - P(A)P(B) - P(\bar{A})P(\bar{B})}$
7	Mutual Information (M)	$\frac{\sum_i \sum_j P(A_i, B_j) \log \frac{P(A_i, B_j)}{P(A_i)P(B_j)}}{\min(-\sum_i P(A_i) \log P(A_i), -\sum_j P(B_j) \log P(B_j))}$
8	J-Measure (J)	$\max \left(P(A,B) \log \left(\frac{P(B A)}{P(B)} \right) + P(\bar{A}\bar{B}) \log \left(\frac{P(\bar{B} \bar{A})}{P(\bar{B})} \right), \right. \\ \left. P(\bar{A},B) \log \left(\frac{P(A B)}{P(A)} \right) + P(\bar{A}\bar{B}) \log \left(\frac{P(\bar{A} \bar{B})}{P(\bar{A})} \right) \right)$
9	Gini index (G)	$\max \left(P(A)[P(B A)^2 + P(\bar{B} A)^2] + P(\bar{A})[P(B \bar{A})^2 + P(\bar{B} \bar{A})^2] \right. \\ \left. - P(B)^2 - P(\bar{B})^2, \right. \\ \left. P(B)[P(A B)^2 + P(\bar{A} B)^2] + P(\bar{B})[P(A \bar{B})^2 + P(\bar{A} \bar{B})^2] \right. \\ \left. - P(A)^2 - P(\bar{A})^2 \right)$
10	Support (s)	$P(A,B)$
11	Confidence (c)	$\max(P(B A), P(A B))$
12	Laplace (L)	$\max \left(\frac{NP(A,B)+1}{NP(A)+2}, \frac{NP(A,B)+1}{NP(B)+2} \right)$
13	Conviction (V)	$\max \left(\frac{P(A)P(\bar{B})}{P(A\bar{B})}, \frac{P(B)P(\bar{A})}{P(\bar{A}B)} \right)$
14	Interest (I)	$\frac{P(A,B)}{P(A)P(B)}$
15	cosine (IS)	$\frac{P(A,B)}{\sqrt{P(A)P(B)}}$
16	Piatetsky-Shapiro's (PS)	$P(A,B) - P(A)P(B)$
17	Certainty factor (F)	$\max \left(\frac{P(B A) - P(B)}{1 - P(B)}, \frac{P(A B) - P(A)}{1 - P(A)} \right)$
18	Added Value (AV)	$\max(P(B A) - P(B), P(A B) - P(A))$
19	Collective strength (S)	$\frac{P(A,B) + P(\bar{A}\bar{B})}{P(A)P(B) + P(\bar{A})P(\bar{B})} \times \frac{1 - P(A)P(B) - P(\bar{A})P(\bar{B})}{1 - P(A,B) - P(\bar{A}\bar{B})}$
20	Jaccard (ζ)	$\frac{P(A,B)}{P(A) + P(B) - P(A,B)}$
21	Kloggen (K)	$\sqrt{P(A,B)} \max(P(B A) - P(B), P(A B) - P(A))$

Comparison of Measures

10 examples of
contingency tables:

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

Rankings of contingency tables using
various measures:

#	ϕ	λ	α	Q	Y	κ	M	J	G	s	c	L	V	I	IS	PS	F	AV	S	ζ	K
E1	1	1	3	3	3	1	2	2	1	3	5	5	4	6	2	2	4	6	1	2	5
E2	2	2	1	1	1	2	1	3	2	2	1	1	1	8	3	5	1	8	2	3	6
E3	3	3	4	4	4	3	3	8	7	1	4	4	6	10	1	8	6	10	3	1	10
E4	4	7	2	2	2	5	4	1	3	6	2	2	2	4	4	1	2	3	4	5	1
E5	5	4	8	8	8	4	7	5	4	7	9	9	9	3	6	3	9	4	5	6	3
E6	6	6	7	7	7	7	6	4	6	9	8	8	7	2	8	6	7	2	7	8	2
E7	7	5	9	9	9	6	8	6	5	4	7	7	8	5	5	4	8	5	6	4	4
E8	8	9	10	10	10	8	10	10	8	4	10	10	10	9	7	7	10	9	8	7	9
E9	9	9	5	5	5	9	9	7	9	8	3	3	3	7	9	9	3	7	9	9	8
E10	10	8	6	6	6	10	5	9	10	10	6	6	5	1	10	10	5	1	10	10	7

Properties of a Good Measure

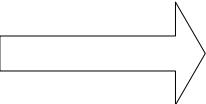
Piatetsky-Shapiro:

3 properties a good measure M must satisfy:

- $M(A,B) = 0$ if A and B are statistically independent
- $M(A,B)$ increase monotonically with $P(A,B)$ when $P(A)$ and $P(B)$ remain unchanged
- $M(A,B)$ decreases monotonically with $P(A)$ [or $P(B)$] when $P(A,B)$ and $P(B)$ [or $P(A)$] remain unchanged

Property under Variable Permutation

	B	$\overline{\mathbf{B}}$
A	p	q
$\overline{\mathbf{A}}$	r	s



	A	$\overline{\mathbf{A}}$
B	p	r
$\overline{\mathbf{B}}$	q	s

Does $M(A, B) = M(B, A)$?

- Symmetric Measures
 - support, lift, collective strength, cosine, Jaccard, ...
- Asymmetric Measures
 - confidence, conviction, Laplace, J-measure, ...

Property under Row/Column Scaling

- Grade-Gender Example (Mosteller, 1968)

	Male	Female	
High	2	3	5
Low	1	4	5
	3	7	10

	Male	Female	
High	4	30	34
Low	2	40	42
	6	70	76



2x




10x

Mosteller:

Underlying association should be independent of the relative number of male and female students in the samples

Property under Null Addition

	B	\bar{B}
A	p	q
\bar{A}	r	s



	B	\bar{B}
A	p	q
\bar{A}	r	s + k

- Invariant measures
 - support, cosine, Jaccard, ...
- Non-invariant measures
 - correlation, Gini, mutual information, odds ratio, etc.

Comparison of Measures

Symbol	Measure	Range	P1	P2	P3	O1	O2	O3	O3'	O4
Φ	Correlation	-1 ... 0 ... 1	Yes	Yes	Yes	Yes	No	Yes	Yes	No
λ	Lambda	0 ... 1	Yes	No	No	Yes	No	No*	Yes	No
α	Odds ratio	0 ... 1 ... ∞	Yes*	Yes	Yes	Yes	Yes	Yes*	Yes	No
Q	Yule's Q	-1 ... 0 ... 1	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No
Y	Yule's Y	-1 ... 0 ... 1	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No
κ	Cohen's	-1 ... 0 ... 1	Yes	Yes	Yes	Yes	No	No	Yes	No
M	Mutual Information	0 ... 1	Yes	Yes	Yes	Yes	No	No*	Yes	No
J	J-Measure	0 ... 1	Yes	No	No	No	No	No	No	No
G	Gini Index	0 ... 1	Yes	No	No	No	No	No*	Yes	No
s	Support	0 ... 1	No	Yes	No	Yes	No	No	No	No
c	Confidence	0 ... 1	No	Yes	No	Yes	No	No	No	Yes
L	Laplace	0 ... 1	No	Yes	No	Yes	No	No	No	No
V	Conviction	0.5 ... 1 ... ∞	No	Yes	No	Yes**	No	No	Yes	No
I	Interest	0 ... 1 ... ∞	Yes*	Yes	Yes	Yes	No	No	No	No
IS	IS (cosine)	0 .. 1	No	Yes	Yes	Yes	No	No	No	Yes
PS	Platetsky-Shapiro's	-0.25 ... 0 ... 0.25	Yes	Yes	Yes	Yes	No	Yes	Yes	No
F	Certainty factor	-1 ... 0 ... 1	Yes	Yes	Yes	No	No	No	Yes	No
AV	Added value	0.5 ... 1 ... 1	Yes	Yes	Yes	No	No	No	No	No
S	Collective strength	0 ... 1 ... ∞	No	Yes	Yes	Yes	No	Yes*	Yes	No
ζ	Jaccard	0 .. 1	No	Yes	Yes	Yes	No	No	No	Yes
K	Klosgen's	$\left(\sqrt{\frac{2}{\sqrt{3}}}-1\right)\left(2-\sqrt{3}-\frac{1}{\sqrt{3}}\right) \dots 0 \dots \frac{2}{3\sqrt{3}}$	Yes	Yes	Yes	No	No	No	No	No

Comparison of Measures

Symbol	Measure	Range	P1	P2	P3	O1	O2	O3	O3'	O4
Φ	Correlation	-1 ... 0 ... 1	Yes	Yes	Yes	Yes	No	Yes	Yes	No
λ	Lambda	0 ... 1	Yes	No	No	Yes	No	No*	Yes	No
α	Odds ratio	0 ... 1 ... ∞	Yes*	Yes	Yes	Yes	Yes	Yes*	Yes	No

where: P1: $O(M) = 0$ if $\det(M) = 0$, i.e., whenever A and B are statistically independent.

P2: $O(M_2) > O(M_1)$ if $M_2 = M_1 + [k \ -k; \ -k \ k]$.

P3: $O(M_2) < O(M_1)$ if $M_2 = M_1 + [0 \ k; \ 0 \ -k]$ or $M_2 = M_1 + [0 \ 0; \ k \ -k]$.

O1: Property 1: Symmetry under variable permutation.

O2: Property 2: Row and Column scaling invariance.

O3: Property 3: Antisymmetry under row or column permutation.

O3': Property 4: Inversion invariance.

O4: Property 5: Null invariance.

Yes*: Yes if measure is normalized.

No*: Symmetry under row or column permutation.

No**: No unless the measure is symmetrized by taking $\max(M(A, B), M(B, A))$.

IS	IS (cosine)	0 .. 1	No	Yes	Yes	Yes	No	No	No	Yes
PS	Platetsky-Shapiro's	-0.25 ... 0 ... 0.25	Yes	Yes	Yes	Yes	No	Yes	Yes	No
F	Certainty factor	-1 ... 0 ... 1	Yes	Yes	Yes	No	No	No	Yes	No
AV	Added value	0.5 ... 1 ... 1	Yes	Yes	Yes	No	No	No	No	No
S	Collective strength	0 ... 1 ... ∞	No	Yes	Yes	Yes	No	Yes*	Yes	No
ζ	Jaccard	0 .. 1	No	Yes	Yes	Yes	No	No	No	Yes
K	Klosgen's	$\left(\sqrt{\frac{2}{\sqrt{3}}}-1\right)\left(2-\sqrt{3}-\frac{1}{\sqrt{3}}\right) \dots 0 \dots \frac{2}{3\sqrt{3}}$	Yes	Yes	Yes	No	No	No	No	No

Subjective Interestingness Measures

- Objective measure
 - Rank patterns based on statistics computed from data
 - 21 measures reported (support, confidence, lift, Laplace, Gini, mutual information, ...)
- Subjective measure
 - Rank pattern's according to user's interpretation
 - A pattern is subjectively interesting if it contradicts the expectation of a user (Silberchatz & Tuzhilin)
 - A pattern is subjectively interesting if it is actionable

Solution to Examples

Example: Measures

Tid	Items
1	ABC
2	BDE
3	ACDE
4	ACE
5	ABCE
6	BCD

Rule	<i>rsup</i>	<i>conf</i>	<i>lift</i>	<i>lev.</i>	<i>jac.</i>	<i>conv</i>	<i>OR</i>
$B \rightarrow C$	0.50	0.75	0.90	-0.05	0.50	0.66	0
$A \rightarrow E$	0.50	0.75	1.13	0.05	0.60	1.33	3
$E \rightarrow A$	0.50	0.75	1.13	0.05	0.60	1.33	3
$A \rightarrow C$	0.67	1.00	1.20	0.11	0.80	inf	NA
$AE \rightarrow C$	0.50	1.00	1.20	0.08	0.60	inf	NA
$AC \rightarrow E$	0.5	0.75	1.13	0.05	0.60	1.33	3

Examine *rsup*, *conf*, *lift*, *leverage*, *jaccard*, *conviction*, and *odds ratio* for the rules above

Solution at End of Slides

