

# Knowledge Discovery and Data Mining

## Unit # 10

## Acknowledgement

- Most of the slides in this presentation are taken from course slides provided by
  - Han and Kimber (Data Mining Concepts and Techniques) and
  - Tan, Steinbach and Kumar (Introduction to Data Mining)

## 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

<i>TID</i>	<i>Items</i>
1	Bread, Milk
2	Bread, Diaper, Beer, Eggs
3	Milk, Diaper, Beer, Coke
4	Bread, Milk, Diaper, Beer
5	Bread, Milk, Diaper, Coke

### Example of Association Rules

{Diaper} → {Beer},  
 {Milk, Bread} → {Eggs,Coke},  
 {Beer, Bread} → {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.  $s(\{\text{Milk, Bread, Diaper}\}) = 2/5$
- Frequent Itemset**
  - An itemset whose support is greater than or equal to a *minsup* threshold

<i>TID</i>	<i>Items</i>
1	Bread, Milk
2	Bread, Diaper, Beer, Eggs
3	Milk, Diaper, Beer, Coke
4	Bread, Milk, Diaper, Beer
5	Bread, Milk, Diaper, Coke

## Definition: Association Rule

- Association Rule

- An implication expression of the form  $X \rightarrow Y$ , where  $X$  and  $Y$  are itemsets
- Example:  
 $\{\text{Milk, Diaper}\} \rightarrow \{\text{Beer}\}$

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

- 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  $X$

Example:

$\{\text{Milk, Diaper}\} \Rightarrow \text{Beer}$

$$s = \frac{\sigma(\text{Milk, Diaper, Beer})}{|T|} = \frac{2}{5} = 0.4$$

$$c = \frac{\sigma(\text{Milk, Diaper, Beer})}{\sigma(\text{Milk, Diaper})} = \frac{2}{3} = 0.67$$

## 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 *minsup* and *minconf* thresholds

$\Rightarrow$  **Computationally prohibitive!**

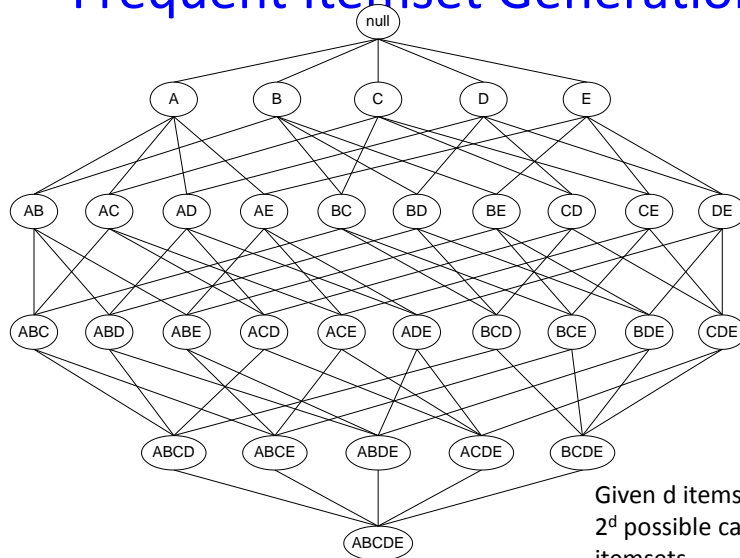
## Apriori

- Apriori is a seminal algorithm proposed by R. Agrawal and R. Srikant in 1994 for mining frequent itemsets for Boolean association rules.
- The name of the algorithm is based on the fact that the algorithm uses *prior knowledge of frequent itemset properties*.
- Two steps process
  - Join
  - Prune

## Mining Association Rules

- Two-step approach:
  1. **Frequent Itemset Generation**
    - Generate all itemsets whose support  $\geq$  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

## Frequent Itemset Generation



Sajjad Haider

Fall 2014

9

## 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:

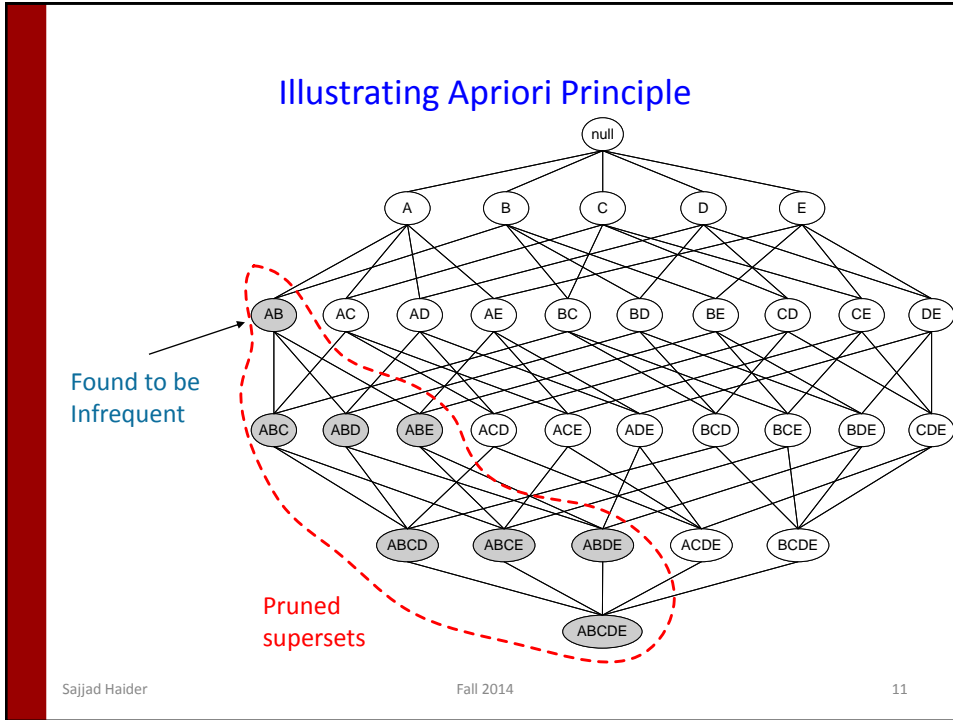
$$\forall X, Y : (X \subseteq Y) \Rightarrow s(X) \geq s(Y)$$

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

Sajjad Haider

Fall 2014

10



### Illustrating Apriori

Item	Count
Bread	4
Coke	2
Milk	4
Beer	3
Diaper	4
Eggs	1

Items (1-itemsets)

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

Pairs (2-itemsets)

Itemset	Count
{Bread,Milk}	3
{Bread,Beer}	2
{Bread,Diaper}	3
{Milk,Beer}	2
{Milk,Diaper}	3
{Beer,Diaper}	3

(No need to generate candidates involving Coke or Eggs)

Triplets (3-itemsets)

Itemset	Count
{Bread,Milk,Diaper}	3

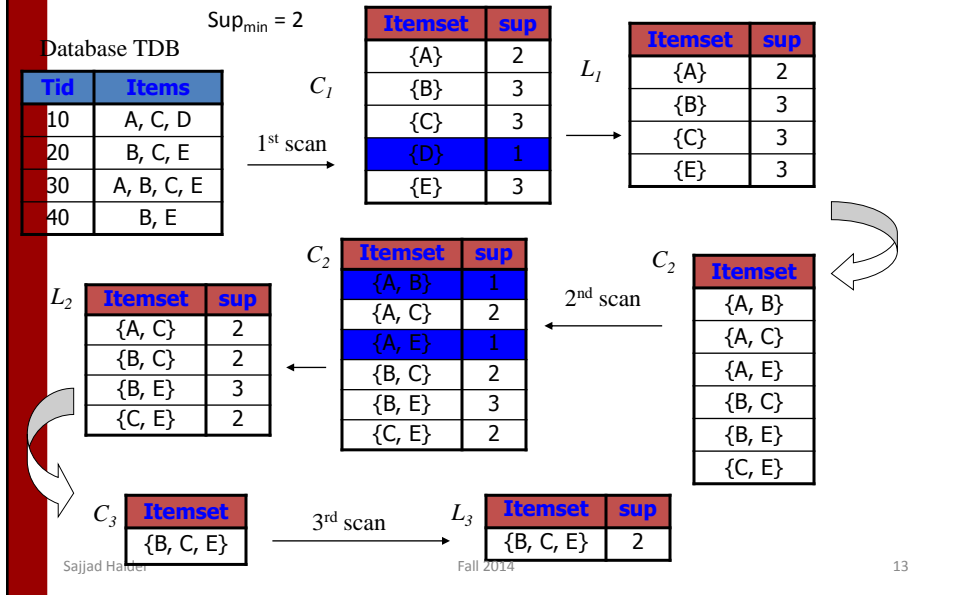
...

Minimum Support = 3

If every subset is considered,  
 ${}^6C_1 + {}^6C_2 + {}^6C_3 = 41$   
 With support-based pruning,  
 $6 + 6 + 1 = 13$

Sajjad Haider Fall 2014 12

## The Apriori Algorithm—An Example



## 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

## Effect of Support Distribution

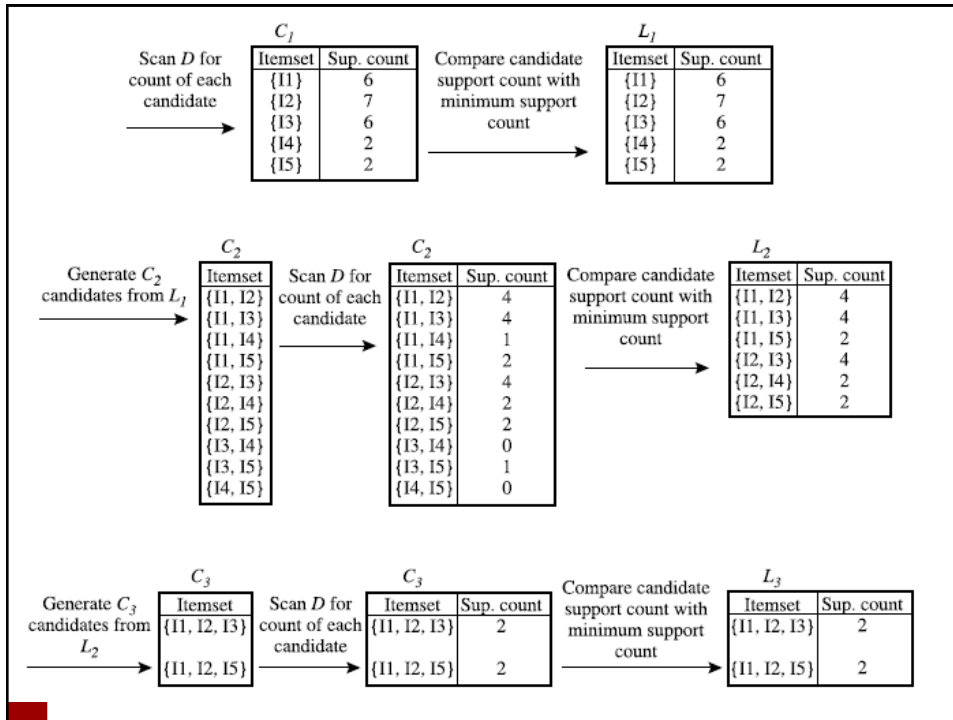
- How to set the appropriate *minsup* threshold?
  - If *minsup* is set too high, we could miss itemsets involving interesting rare items (e.g., expensive products)
  - If *minsup* is set too low, it is computationally expensive and the number of itemsets is very large

## Example

<i>TID</i>	<i>List of item_IDs</i>
T100	I1, I2, I5
T200	I2, I4
T300	I2, I3
T400	I1, I2, I4
T500	I1, I3
T600	I2, I3
T700	I1, I3
T800	I1, I2, I3, I5
T900	I1, I2, I3

Minimum Support = 2





## 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:
 

$ABC \rightarrow D,$	$ABD \rightarrow C,$	$ACD \rightarrow B,$	$BCD \rightarrow A,$
$A \rightarrow BCD,$	$B \rightarrow ACD,$	$C \rightarrow ABD,$	$D \rightarrow ABC$
$AB \rightarrow CD,$	$AC \rightarrow BD,$	$AD \rightarrow BC,$	$BC \rightarrow AD,$
$BD \rightarrow AC,$	$CD \rightarrow AB,$		
- 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  
 $c(ABC \rightarrow D)$  can be larger or smaller than  $c(AB \rightarrow D)$

- But confidence of rules generated from the same itemset has an anti-monotone property

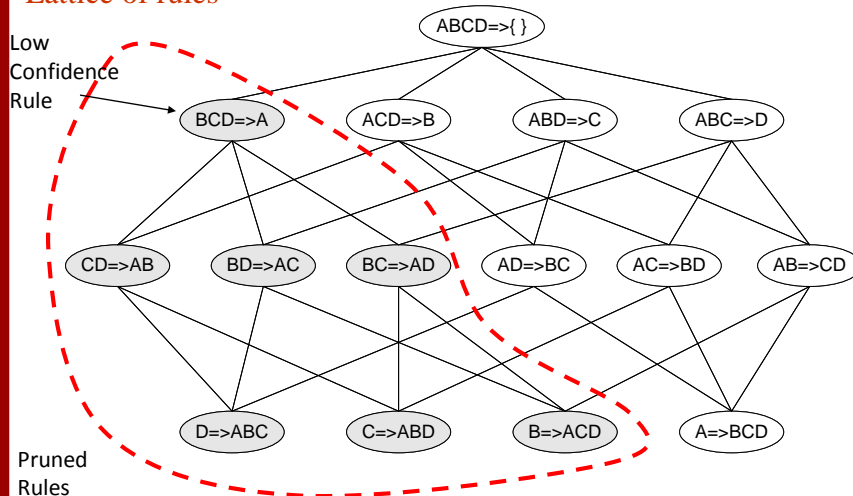
- e.g.,  $L = \{A,B,C,D\}$ :

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

- Confidence is anti-monotone w.r.t. number of items on the RHS of the rule

## Rule Generation for Apriori Algorithm

### Lattice of rules



## Rule Generation Example

- Suppose the data contains the frequent itemset  $I = \{I1, I2, I5\}$ . What are the association rules that can be generated from  $I$ ?
- The nonempty subsets of  $I$  are  $\{I1, I2\}$ ,  $\{I1, I5\}$ ,  $\{I2, I5\}$ ,  $\{I1\}$ ,  $\{I2\}$ , and  $\{I5\}$ .
- If the minimum confidence threshold is, say, 70%, then only the second, third, and last rules below are output

$I1 \wedge I2 \Rightarrow I5,$	$confidence = 2/4 = 50\%$
$I1 \wedge I5 \Rightarrow I2,$	$confidence = 2/2 = 100\%$
$I2 \wedge I5 \Rightarrow I1,$	$confidence = 2/2 = 100\%$
$I1 \Rightarrow I2 \wedge I5,$	$confidence = 2/6 = 33\%$
$I2 \Rightarrow I1 \wedge I5,$	$confidence = 2/7 = 29\%$
$I5 \Rightarrow I1 \wedge I2,$	$confidence = 2/2 = 100\%$

## Pattern Evaluation

- Association rule algorithms tend to produce too many rules
  - many of them are uninteresting or redundant
  - Redundant if  $\{A,B,C\} \rightarrow \{D\}$  and  $\{A,B\} \rightarrow \{D\}$  have same support & confidence
- Interestingness measures can be used to prune/rank the derived patterns
- In the original formulation of association rules, support & confidence are the only measures used

## Drawback of Confidence

	Coffee	$\overline{\text{Coffee}}$	
Tea	15	5	20
$\overline{\text{Tea}}$	75	5	80
	90	10	100

Association Rule: Tea  $\rightarrow$  Coffee

Confidence =  $P(\text{Coffee}|\text{Tea}) = 0.75$

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

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

$\Rightarrow P(\text{Coffee}|\overline{\text{Tea}}) = 0.9375$

## Statistical Independence

- Population of 1000 students
  - 600 students know how to swim (S)
  - 700 students know how to bike (B)
  - 420 students know how to swim and bike (S,B)
  - $P(S \wedge B) = 420/1000 = 0.42$
  - $P(S) \times P(B) = 0.6 \times 0.7 = 0.42$
  - $P(S \wedge B) = P(S) \times P(B) \Rightarrow$  Statistical independence
  - $P(S \wedge B) > P(S) \times P(B) \Rightarrow$  Positively correlated
  - $P(S \wedge B) < P(S) \times P(B) \Rightarrow$  Negatively correlated

## Statistical-based Measures

- Measures that take into account statistical dependence

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

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

## Example: Lift/Interest

	Coffee	<u>Coffee</u>	
Tea	15	5	20
<u>Tea</u>	75	5	80
	90	10	100

Association Rule: Tea → Coffee

Confidence =  $P(\text{Coffee}|\text{Tea}) = 0.75$

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

⇒ Lift =  $0.75/0.9 = 0.8333$  ( $< 1$ , therefore is negatively associated)

#	Measure	Formula
1	$\phi$ -coefficient	$\frac{P(A,B) - P(A)P(B)}{\sqrt{P(A)P(B)(1-P(A))(1-P(B))}}$
2	Goodman-Kruskal's ( $\lambda$ )	$\frac{\sum_{j_1} \max_k P(A_{j_1}, B_k) - \max_k P(A_{j_1}) - \max_k P(B_k)}{2 - \max_k P(A_{j_1}) - \max_k P(B_k)}$
3	Odds ratio ( $\alpha$ )	$\frac{P(A,B)P(\bar{A},\bar{B})}{P(\bar{A},B)P(A,\bar{B})}$
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 ( $\kappa$ )	$\frac{P(A,B) + P(\bar{A},\bar{B}) - P(A)P(B) - P(\bar{A})P(\bar{B})}{1 - P(A)P(B) - P(\bar{A})P(\bar{B})}$
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 ( $J$ )	$\max \left( P(A, B) \log \left( \frac{P(B A)}{P(B)} \right) + P(\bar{A}\bar{B}) \log \left( \frac{P(\bar{B} \bar{A})}{P(\bar{B})} \right), P(A, B) \log \left( \frac{P(A B)}{P(A)} \right) + P(\bar{A}\bar{B}) \log \left( \frac{P(\bar{A} \bar{B})}{P(\bar{A})} \right) \right)$
9	Gini index ( $G$ )	$\max \left( P(A)[P(B A)]^2 + P(\bar{B} A)]^2 + P(\bar{A})[P(B \bar{A})]^2 + P(\bar{B} \bar{A})]^2 - P(B)^2 - P(\bar{B})^2, P(\bar{B})[P(A \bar{B})]^2 + P(\bar{A} \bar{B})]^2 + P(A)[P(A B)]^2 + P(\bar{A} B)]^2 - P(A)^2 - P(\bar{A})^2 \right)$
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 ( $V$ )	$\max \left( \frac{P(A)P(\bar{B})}{P(A\bar{B})}, \frac{P(\bar{B})P(A)}{P(\bar{B}A)} \right)$
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 ( $S$ )	$\frac{P(A,B) + P(\bar{A}\bar{B})}{P(A,B)} \times \frac{1 - P(A)P(B) - P(\bar{A})P(\bar{B})}{1 - P(A,B) - P(\bar{A}\bar{B})}$
20	Jaccard ( $\zeta$ )	$\frac{P(A,B)}{P(A) + P(B) - P(A,B)}$
21	Klogsen ( $K$ )	$\sqrt{\frac{P(A,B)}{P(A)P(B)}} \max(P(B A) - P(B), P(A B) - P(A))$

There are lots of measures proposed in the literature

Some measures are good for certain applications, but not for others

What criteria should we use to determine whether a measure is good or bad?

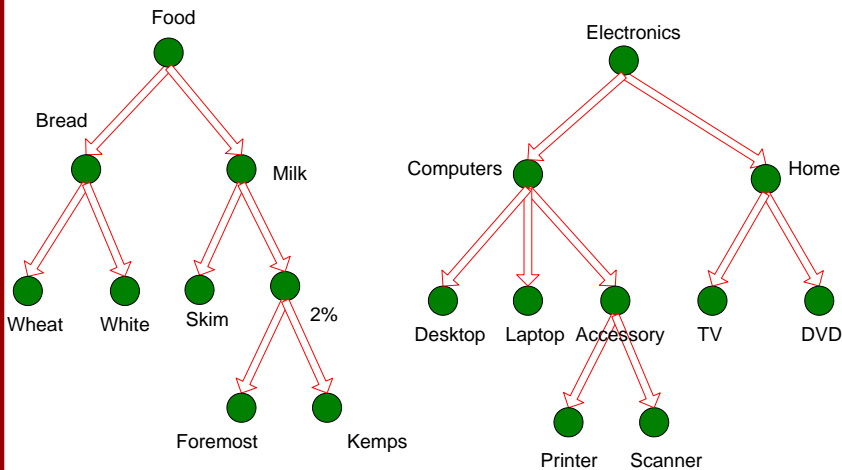
What about Apriori-style support based pruning? How does it affect these measures?

Sajjad Haider

Fall 2014

27

## Multi-level Association Rules



Sajjad Haider

Fall 2014

28

## Multi-level Association Rules

- Why should we incorporate concept hierarchy?
    - Rules at lower levels may not have enough support to appear in any frequent itemsets
    - Rules at lower levels of the hierarchy are overly specific
      - e.g., skim milk → white bread, 2% milk → wheat bread, skim milk → wheat bread, etc.
- are indicative of association between milk and bread

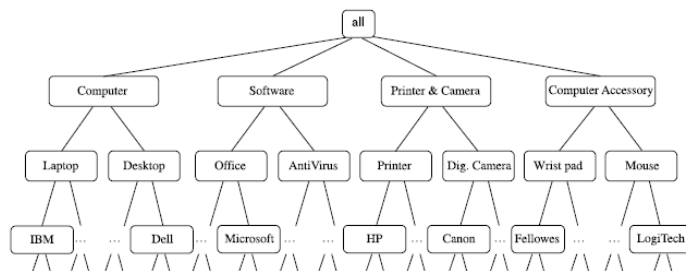
Sajjad Haider

Fall 2014

29

## Example

<i>TID</i>	<i>Items Purchased</i>
T100	IBM-ThinkPad-T40/2373, HP-Photosmart-7660
T200	Microsoft-Office-Professional-2003, Microsoft-Plus!-Digital-Media
T300	Logitech-MX700-Cordless-Mouse, Fellowes-Wrist-Rest
T400	Dell-Dimension-XPS, Canon-PowerShot-S400
T500	IBM-ThinkPad-R40/P4M, Symantec-Norton-Antivirus-2003
...	...



## Concept Hierarchies

- Multilevel association rules can be mined efficiently using concept hierarchies under a support-confidence framework.
- In general, a top-down strategy is employed, where counts are accumulated for the calculation of frequent itemsets at each concept level, starting at the concept level 1 and working downward in the hierarchy toward the more specific concept levels, until no more frequent itemsets can be found.
- For each level, any algorithm for discovering frequent itemsets may be used, such as Apriori or its variations.

Sajjad Haider

Fall 2014

31

## Uniform Minimum Support for All Levels

- The same minimum support threshold is used when mining at each level of abstraction.
- For example, a minimum support threshold of 5% is used throughout (e.g., for mining from “computer” down to “laptop computer”).
- Both “computer” and “laptop computer” are found to be frequent, while “desktop computer” is not.

Level 1  
*min\_sup* = 5%

computer [support = 10%]

Level 2  
*min\_sup* = 5%

laptop computer [support = 6%]

desktop computer [support = 4%]

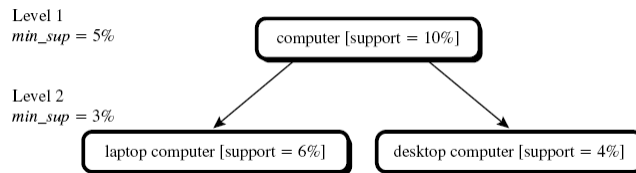
Sajjad Haider

32



## Reduced Minimum Support at Lower Levels

- Each level of abstraction has its own minimum support threshold. The deeper the level of abstraction, the smaller the corresponding threshold is.
- For example, the minimum support thresholds for levels 1 and 2 are 5% and 3%, respectively.
- In this way, “computer,” “laptop computer,” and “desktop computer” are all considered frequent.



Sajjad Haider

Fall 2014

33

## Continuous and Categorical Attributes

How to apply association analysis formulation to non-symmetric binary variables?

Session Id	Country	Session Length (sec)	Number of Web Pages viewed	Gender	Browser Type	Buy
1	USA	982	8	Male	IE	No
2	China	811	10	Female	Netscape	No
3	USA	2125	45	Female	Mozilla	Yes
4	Germany	596	4	Male	IE	Yes
5	Australia	123	9	Male	Mozilla	No
...	...	...	...	...	...	...

Example of Association Rule:

$\{\text{Number of Pages} \in [5,10) \wedge (\text{Browser}=\text{Mozilla})\} \rightarrow \{\text{Buy} = \text{No}\}$

Sajjad Haider

Fall 2014

34

## Handling Categorical Attributes

- Transform categorical attribute into asymmetric binary variables.
- Introduce a new “item” for each distinct attribute-value pair
  - Example: replace Browser Type attribute with
    - Browser Type = Internet Explorer
    - Browser Type = Mozilla
    - Browser Type = Netscape
- What if attribute has many possible values
  - Example: attribute country has more than 200 possible values
  - Many of the attribute values may have very low support
    - Potential solution: Aggregate the low-support attribute values

## Handling Continuous Data

- There are certain applications in which analysts are more interested in finding associations among the continuous attributes, rather than associations among discrete intervals of the continuous attributes.

## Min-Apriori (Han et al)

Document-term matrix:

TID	W1	W2	W3	W4	W5
D1	2	2	0	0	1
D2	0	0	1	2	2
D3	2	3	0	0	0
D4	0	0	1	0	1
D5	1	1	1	0	2

Example:

W1 and W2 tends to appear together in the same document

## Min-Apriori

- Data contains only continuous attributes of the same “type”
  - e.g., frequency of words in a document

TID	W1	W2	W3	W4	W5
D1	2	2	0	0	1
D2	0	0	1	2	2
D3	2	3	0	0	0
D4	0	0	1	0	1
D5	1	1	1	0	2

- Potential solution:
  - Convert into 0/1 matrix and then apply existing algorithms
    - lose word frequency information
  - Discretization does not apply as users want association among words not ranges of words

## Min-Apriori

- How to determine the support of a word?
  - If we simply sum up its frequency, support count will be greater than total number of documents!
    - Normalize the word vectors – e.g., using  $L_1$  norm
    - Each word has a support equals to 1.0

TID	W1	W2	W3	W4	W5
D1	2	2	0	0	1
D2	0	0	1	2	2
D3	2	3	0	0	0
D4	0	0	1	0	1
D5	1	1	1	0	2

→ Normalize

TID	W1	W2	W3	W4	W5
D1	0.40	0.33	0.00	0.00	0.17
D2	0.00	0.00	0.33	1.00	0.33
D3	0.40	0.50	0.00	0.00	0.00
D4	0.00	0.00	0.33	0.00	0.17
D5	0.20	0.17	0.33	0.00	0.33

Sajjad Haider

Fall 2014

39

## Min-Apriori

- New definition of support:

$$\text{sup}(C) = \sum_{i \in T} \min_{j \in C} D(i, j)$$

TID	W1	W2	W3	W4	W5
D1	0.40	0.33	0.00	0.00	0.17
D2	0.00	0.00	0.33	1.00	0.33
D3	0.40	0.50	0.00	0.00	0.00
D4	0.00	0.00	0.33	0.00	0.17
D5	0.20	0.17	0.33	0.00	0.33

Example:

$$\begin{aligned} \text{Sup}(W1, W2, W3) \\ &= 0 + 0 + 0 + 0 + 0.17 \\ &= 0.17 \end{aligned}$$

Sajjad Haider

Fall 2014

40

## Anti-monotone property of Support

TID	W1	W2	W3	W4	W5
D1	0.40	0.33	0.00	0.00	0.17
D2	0.00	0.00	0.33	1.00	0.33
D3	0.40	0.50	0.00	0.00	0.00
D4	0.00	0.00	0.33	0.00	0.17
D5	0.20	0.17	0.33	0.00	0.33

Example:

$$\text{Sup}(W1) = 0.4 + 0 + 0.4 + 0 + 0.2 = 1$$

$$\text{Sup}(W1, W2) = 0.33 + 0 + 0.4 + 0 + 0.17 = 0.9$$

$$\text{Sup}(W1, W2, W3) = 0 + 0 + 0 + 0 + 0.17 = 0.17$$

## Document Example

NATO	Cricket	Shareef	Missing	Court
4	0	3	1	0
1	0	1	4	3
0	5	2	0	1
3	0	4	0	0
1	0	2	3	3

## Principal Component Analysis

- Principal component analysis (PCA) or Factor Analysis (FA) statistical techniques applied to a single set of variables where the researcher is interested in discovering which variables in the set form coherent subsets that are relatively independent of one another.
- Variables that are correlated with one another but largely independent of other subsets of variables are combined into factors.
- Factors are thought to reflect underlying processes that have created the correlations among variables.

## Fundamental Steps

- Steps in PCA or FA include
  - Selecting and measuring a set of variables
  - Preparing the covariance matrix
  - Extracting a set of factors from the covariance matrix
  - Determining the number of factors
  - interpreting the results
- Although there are relevant statistical considerations to most of these steps, an important test of the analysis is its interpretability.

## Application

- PCA is a useful statistical technique that has found application in fields such as face recognition and image compression, and is a common technique for finding patterns in data of high dimension.

## Limitation

- One of the problems with PCA and FA is that there is no criterion variable against which to test the solution.
- In regression analysis, for instance, the dependent variable (DV) is a criterion and the correlation between observed and predicted DV scores serves as a test of the solution
- In classification, the solution is judged by how well it predicts group membership.
- But in PCA or FA there is no external criterion such as group membership against which to test the solution.

## Practical Issues

- Because FA and PCA are exquisitely sensitive to the sizes of correlations, it is critical that honest, reliable correlations be employed.
- Sensitivity to outlying cases, problems created by missing data, and degradation of correlations between poorly distributed variables all plague FA and PCA.

## Variance and Covariance

- Standard deviation and variance only operate on one dimension.
- However, it is useful to have a similar measure to find out how much the dimensions vary from the mean *with respect to each other*.
- Covariance is such a measure. Covariance is always measured *between 2* dimensions.
- If you calculate the covariance between one dimension and *itself*, you get the variance.
- So, if you had a 3-dimensional data set  $(x, y, z)$ , then you could measure the covariance between the x and y dimensions, the x and z dimensions, and the y and z dimensions.



## Variance and Covariance (Cont'd)

$$\text{var}(X) = \frac{\sum_{i=1}^n (X_i - \bar{X})(X_i - \bar{X})}{(n-1)}$$

$$\text{cov}(X, Y) = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{(n-1)}$$

Year	Gold	Dollar
2006	1233	60.8
2007	1581	61.5
2008	2075	79.1
2009	3070	83.5
2010	3835	86.5
2011	4720	90.2
2012	5500	97.8
2013	4250	110.1

## Eigen Values and Eigen Vectors

- Let  $A$  be a square matrix. A non-zero vector  $x$  is called an **eigenvector** of  $A$  if and only if there exists a number (real or complex)  $\lambda$  such that

$$Ax = \lambda x$$

- If such a number  $\lambda$  exists, it is called an **eigenvalue** of  $A$ . The vector  $x$  is called eigenvector associated to the eigenvalue .

## Eigen Vectors

- Eigenvectors can only be found for square matrices.
- And, not every square matrix has eigenvectors.
- And, given an nxn matrix that does have eigenvectors, there are n of them.
- For example, given a 3x3 matrix, there are 3 eigenvectors.

## Example (Source: Wikipedia)

For the matrix  $A$

$$A = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}.$$

the vector

$$\mathbf{x} = \begin{bmatrix} 3 \\ -3 \end{bmatrix}$$

is an eigenvector with eigenvalue 1. Indeed,

$$A\mathbf{x} = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix} \begin{bmatrix} 3 \\ -3 \end{bmatrix} = \begin{bmatrix} 2 \cdot 3 + 1 \cdot (-3) \\ 1 \cdot 3 + 2 \cdot (-3) \end{bmatrix} = \begin{bmatrix} 3 \\ -3 \end{bmatrix} = 1 \cdot \begin{bmatrix} 3 \\ -3 \end{bmatrix}.$$

On the other hand the vector

$$\mathbf{x} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

is *not* an eigenvector, since

$$\begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix} \begin{bmatrix} 0 \\ 1 \end{bmatrix} = \begin{bmatrix} 2 \cdot 0 + 1 \cdot 1 \\ 1 \cdot 0 + 2 \cdot 1 \end{bmatrix} = \begin{bmatrix} 1 \\ 2 \end{bmatrix}.$$

and this vector is not a multiple of the original vector  $\mathbf{x}$ .

## Eigen Value Computation

- When a transformation is represented by a square matrix  $A$ , the eigen value equation can be expressed as  $Ax - \lambda x = 0$
- Where  $I$  is the identify matrix. This can be rearranged to  $(A - \lambda I)x = 0$
- If there exists an inverse  $(A - \lambda I)^{-1}$  then both sides can be left multiplied by the inverse to obtain the trivial solutions:  $x = 0$ . Thus we require there to be no inverse by assuming from linear algebra that the determinants equals zero:
- $\det(A - \lambda I) = 0$
- To compute eigen vectors, solve for  $Ax = \lambda x$  for all values of  $\lambda$ .

## Example

### Exercises

For the following square matrix:

$$\begin{pmatrix} 3 & 0 & 1 \\ -4 & 1 & 2 \\ -6 & 0 & -2 \end{pmatrix}$$

Decide which, if any, of the following vectors are eigenvectors of that matrix and give the corresponding eigenvalue.

$$\begin{pmatrix} 2 \\ 2 \\ -1 \end{pmatrix} \begin{pmatrix} -1 \\ 0 \\ 2 \end{pmatrix} \begin{pmatrix} -1 \\ 1 \\ 3 \end{pmatrix} \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} \begin{pmatrix} 3 \\ 2 \\ 1 \end{pmatrix}$$

## Working of PCA

- The original data are projected onto a much smaller space, resulting in dimensionality reduction.
- Unlike attribute subset selection, which reduces the attribute set size by retaining a subset of the initial set of attributes, PCA “combines” the essence of attributes by creating an alternative, smaller set of variables.
- The initial data can then be projected onto this smaller set.

## Steps 1 and 2

- Step 1:
  - The input data are normalized, so that each attribute falls within the same range. This step helps ensure that attributes with large domains will not dominate attributes with smaller domains.
- Step 2:
  - PCA computes  $k$  orthonormal vectors that provide a basis for the normalized input data.
  - These vectors are referred to as the *principal components*.
  - *The input data are a linear combination of the principal components.*

## Steps 3 and 4

- Step 3:
  - The principal components are sorted in order of decreasing “significance” or strength.
  - The principal components essentially serve as a new set of axes for the data, providing important information about variance.
  - That is, the sorted axes are such that the first axis shows the most variance among the data, the second axis shows the next highest variance, and so on.
- Step 4:
  - Because the components are sorted according to decreasing order of “significance,” the size of the data can be reduced by eliminating the weaker components, that is, those with low variance.
  - Using the strongest principal components, it should be possible to reconstruct a good approximation of the original data.

Sajjad Haider

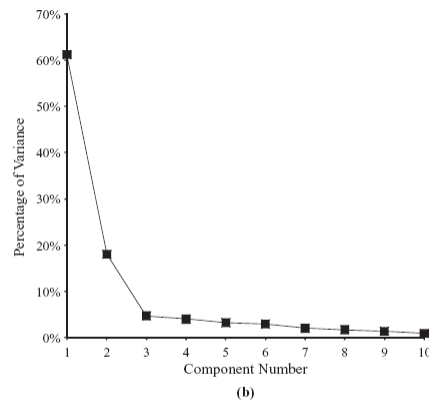
Fall 2014

57

## Example (Source: Witten et al.)

Axis	Variance	Cumulative
1	61.2%	61.2%
2	18.0%	79.2%
3	4.7%	83.9%
4	4.0%	87.9%
5	3.2%	91.1%
6	2.9%	94.0%
7	2.0%	96.0%
8	1.7%	97.7%
9	1.4%	99.1%
10	0.9%	100.0%

(a)

**FIGURE 7.5**

Principal components transform of a dataset: (a) variance of each component and (b) variance plot.

Sajjad Haider

Fall 2014

58

# KNIME: PCA DEMO