COEN281 -- Introduction to Pattern Recognition and Data Mining

Lecture 16: Association Rules

Instructor: Dr. Giovanni Seni
GSeni@scu.edu

Department of Computer Engineering
Santa Clara University

Fall/18

Syllabus

Week 1	Introduction; R (Ch.1)	
Week 2	Bayesian Decision Theory (Ch.2; DHS: 2.1-2.6, 2.9) Parameter Estimation (DHS: 3.1-3.4)	Supervized Learning
Week 3	Linear Discriminant Functions (Ch.3&4; DHS: 3.8.2, 5.1-5.8) Regularization (Ch.6; SE: Ch.3)	
Week 4	Neural Networks (DHS: 6.1-6.6, 6.8);	
Week 5	Support Vector Machines (Ch.9)	
Week 6	Decision Trees (Ch. 8.1; DHS: 8.3; Ch 2 SE)	
Week 7	Ensemble Methods (Ch. 8.2; SE: Ch 4, 5) ;	
Week 8	Clustering (Ch. 10; DHS: 10.6, 10.7) Clustering (DHS: 10.9); How many clusters are there? (DHS: 10.10)	Unsupervised Learning
Week 9	**Non-metric: Association Rules** Collaborative Filtering	
Week 10	Text Retrieval; Other topics	

Overview

- Market-Basket Data
  - Itemsets
  - Association rules
- Finding Itemsets and Rules
  - Problem definition
  - Example
- Apriori Algorithm - Itemset and Rule Mining
- Dealing with Non-Binary Data
- Software Resources

Market-Basket Data
Itemsets

- Indicator $n \times p$ matrix

t_{id}	*beer*	*chips*	*pizza*	*wine*
100	1	1	0	1
200	1	1	0	0
300	0	0	1	1
400	0	1	1	0

  - Rows correspond to "baskets"
  - Columns correspond to "items" (*I*)

- Itemset: $X=\{i_1,\ldots,i_k\} \subseteq I$ is called a *k-itemset*

Market-Basket Data

Association Rules

- A simple and interpretable probabilistic statement about the co-occurrence of events

$$\text{IF } \underbrace{A \text{ AND } B}_{\substack{\text{antecedent} \\ \text{(body)}}} \text{ THEN } \underbrace{C \text{ AND } D}_{\substack{\text{consequent} \\ \text{(head)}}} \text{ with probability } p$$

- Or, $X \Rightarrow Y$, where X and Y are itemsets, and $X \cap Y = \{\}$
  - If a transaction contains all the items in X, then it also contains all items in Y
- Use cases
  - Organize product stands, catalogs, web pages, promotions, adverse side effects, fraud detection

Market-Basket Data

Association Rules (2)

- Does $A \Rightarrow B$?
  - Easy if given specific *A* and specific B

		B: Y	B: N
A	Y		
	N		

and χ^2

  - Finding all A,B with interesting table is more difficult
  - p items, i.e., $|I| = p$, we need $p \times (p-1)$ table
    - Grocery store: $p \approx 25{,}000$
    - Amazon: $p \approx 15M$
- More generally, $2^{|I|}$ different itemsets, $3^{|I|}$ different rules
  - Quite sparse matrix – i.e, customers buy small subset of products
  - Will restrict to "frequent" itemsets

Finding Itemsets & Rules

Problem Definition

- *Cover* – set of transactions that "support" X

  $cover(X) = \{t_{id} \mid (t_{id}, \mathrm{I}) \in D \text{ and } X \subseteq \mathrm{I}\}$

- *Support* – number of transactions in the cover

  $support(X) = |\, cover(X) \,|$ note $|D| = support(\{\})$

- *Frequency* – probability of X occurring in a transaction

  $$frequency(X) = P(X) = \frac{support(X)}{|D|}$$

- *Frequent itemsets* – given a minimal support threshold σ

  $F(\sigma) = \{X \subseteq I \mid support(X) \geq \sigma\}$ // i.e., sets of items that occur reasonably often together

Finding Itemsets & Rules

Problem Definition (2)

- *Support* – of an association rule

  $support(X \Rightarrow Y) = support(X \cup Y)$

- *Confidence* – also referred to as accuracy

  $$confidence(X \Rightarrow Y) = \frac{support(X \cup Y)}{support(X)} = \frac{frrquency(X \wedge Y)}{frequency(X)} = P(Y \mid X)$$

  – Fraction of rows that satisfy Y among those rows that satisfy X

  – ML estimate of conditional probability

- Confident rules – given a minimal confidence threshold γ

  $R(\gamma) = \{X \Rightarrow Y \mid confidence(X \Rightarrow Y) \geq \gamma\}$

Finding Itemsets & Rules
Problem Definition (3)

- Association Rule Mining
  - Given a set of items I, a transaction database D over I, thresholds σ and γ, find $R(\sigma,\gamma) = \{X \Rightarrow Y \mid X,Y \subseteq I, X \cap Y = \{\},$
  $X \cup Y \in F(\sigma)$
  $confidence(X \Rightarrow Y) \geq \gamma\}$
  - i.e., restrict attention to well supported rules
  - warning – the confidence of a rule is not necessarily a very good indication of interestingness
    - E.g., *confidence*(pregnancy ⇒female) = P(female | pregnancy) = 1!
    - E.g., *confidence*(X ⇒"Harry Potter") ≈ 1

Finding Itemsets & Rules
Example

Itemset (σ = 1)	*Cover*	*Support*	*Frequency*
{}	{100, 200, 300, 400}	4	100%
{beer}	{100, 200}	2	50%
{chips}	{100, 200, 400}	3	75%
{pizza}	{300, 400}	2	50%
{wine}	{100, 300}	2	50%
{beer, chips}	{100, 200}	2	50%
{beer, wine}	{100}	1	25%
{chips, pizza}	{400}	1	25%
{chips, wine}	{100}	1	25%
{pizza, wine}	{300}	1	25%
{beer, chips, wine}	{100}	1	25%

Finding Itemsets & Rules

Example (2)

Rules ($\sigma = 1, \gamma = 50\%$)	*Support*	*Frequency*	*Confidence*
{beer} $\Rightarrow$ {chips}	2	50%	100%
{beer} $\Rightarrow$ {wine}	1	25%	50%
{chips} $\Rightarrow$ {beer}	2	50%	66%
{pizza} $\Rightarrow$ {chips}	1	25%	50%
{pizza} $\Rightarrow$ {wine}	1	25%	50%
{wine} $\Rightarrow$ {beer}	1	25%	50%
{wine} $\Rightarrow$ {chips}	1	25%	50%
{wine} $\Rightarrow$ {pizza}	1	25%	50%
{beer, chips} $\Rightarrow$ {wine}	1	25%	50%
{beer, wine} $\Rightarrow$ {chips}	1	25%	100%
{chips, wine} $\Rightarrow$ {beer}	1	25%	100%
{beer} $\Rightarrow$ {chips, wine}	1	25%	50%
{wine} $\Rightarrow$ {beer, chips}	1	25%	50%

Apriori Algorithm

Itemset Mining

- Key insight: support monotonicity
  - A set X of variables can be frequent only if all the subsets of X are frequent
  - Given itemsets $\boxed{X, Y \subseteq I, \quad X \subseteq Y \Rightarrow support(Y) \leq support(X)}$

    $\Rightarrow$ No need to find the frequency of any set X that has a non-frequent proper subset
- Breath-first search approach
  - Find all frequent sets of 1 variable
  - Build *candidate* sets of size 2: $\{A, B\}$ st. $\{A\}$ and $\{B\}$ are frequent
  - Count the support of candidate sets of size 2; keep frequent ones
  - Build candidate sets of size 3, etc.

Apriori Algorithm
Itemset Mining (2)

- Input: D, σ
  Output: $F(\sigma)$

```
k=1
C_1 = {{i} | i ∈ I}
while C_k ≠ {} do
  // database pass – compute support of candidate itemsets
  for each transaction (t_id, T) ∈ D do
    for each candidate itemset X ∈ C_k do
      if X ⊆ T then
        X.support++
      endif
    endfor
  endfor
```


Apriori Algorithm
Itemset Mining (3)

- Con't from previous page:

```
  // extract frequent itemsets
  F_k = {X | X.support ≥ σ}
  // generate new candidate itemsets
  for X, Y ∈ F_k , X[i]==Y[i] for 1≤i≤k-1, and X[k]<Y[k] do
    I = X ∪ {Y[k]}
    if ∀J⊂I, |J| == k and J ∈ F_k then
      C_{k+1} = C_{k+1} ∪ {I}
    endif
  endfor
  k++
endwhile
```

E.g., X={1 2 3} ⇒ I={1 2 3 5}
Y={1 2 5}

However, I is eliminated if {2 3 5} $\notin F_k$

Apriori Algorithm
Rule Mining

- For every frequent itemset I, there are up to $2^{|I|}$ rules

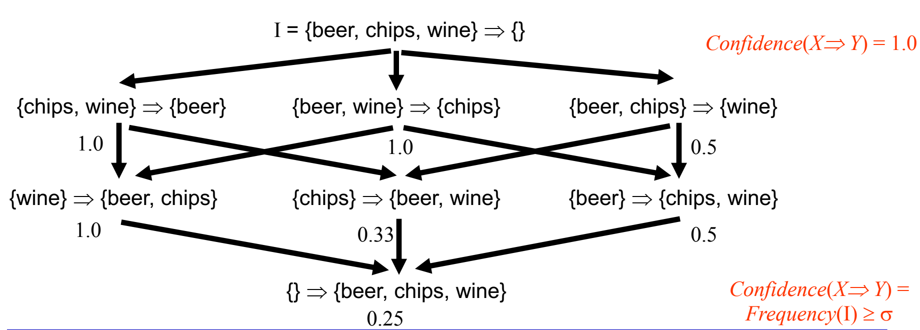
  I = {beer, chips, wine} →
  - I ⇒ {}
  - {chips, wine} ⇒ {beer}; {beer, wine} ⇒ {chips}; {beer, chips} ⇒ {wine}
  - {wine} ⇒ {beer, chips}; {chips} ⇒ {beer, wine}; {beer} ⇒ {chips, wine}
  - {} ⇒ {beer, chips, wine}

- Exploit similar property: confidence monotonicity
  - Given itemsets $\boxed{X, Y, Z \subseteq I,\ \text{st.}\ X \cap Y = \{\} \quad \text{then} \quad confidence(X \setminus Z \Rightarrow Y \cup Z) \leq confidence(X \Rightarrow Y)}$
  - Proof: Since $X \cup Y \subseteq X \cup Y \cup Z$, and $X \setminus Z \subseteq X$, we have

$$\frac{support(X \cup Y \cup Z)}{support(X \setminus Z)} \leq \frac{support(X \cup Y)}{support(X)}$$

Apriori Algorithm
Rule Mining (2)

- In other words, confidence is monotone decreasing with respect to extension of the "head" (consequent) of a rule
  - If a certain head causes a rule to be unconfident, all of the head's supersets must result in unconfident rules



Apriori Algorithm
Rule Mining (3)

- Input: D, $F(\sigma)$, γ
  Output: $R(\sigma, \gamma)$

$R=\{\}$
for all $\mathrm{I} \in F(\sigma)$ do
  $R = R \cup$ "$\mathrm{I} \Rightarrow \{\}$" // this rule always holds with confidence 1.0
  $k=1$
  $C_1 = \{\{i\} \mid i \in I\}$ // candidates head of length 1
  while $C_k \neq \{\}$ do
    // extract all heads of confident rules
    $H_k = \{X \in C_k \mid Confidence(\mathrm{I}\backslash X \Rightarrow X) \geq \gamma\}$
    // generate new candidate heads
    $GenerateHeads(H_k, C_{k+1})$ // exactly like candidate itemset generation
    $k{+}{+}$
  endwhile
  $R = R \cup \{\mathrm{I}\backslash X \Rightarrow X \mid X \in H_1 \cup H_2 \cup \ldots \cup H_k\}$ // cumulate all rules
endfor

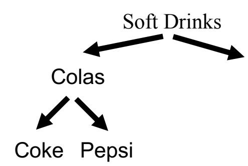
Apriori Algorithm
Rule Mining (4)

- $GenerateHeads(H_k, C_{k+1})$
  {
    for $X, Y \in H_k$, $X[i]{=}{=}Y[i]$ for $1 \leq i \leq k\text{-}1$, and $X[k] < Y[k]$ do
      $\mathrm{I} = X \cup \{Y[k]\}$
      if $\forall \mathrm{J} \subset \mathrm{I}$, $|\mathrm{J}| == \mathrm{k}$ and $\mathrm{J} \in H_k$ then // prune step
        $C_{k+1} = C_{k+1} \cup \{\mathrm{I}\}$
      endif
    endfor
  }

- Special data structures are used to efficiently find the itemsets contained in a transaction or in another itemset: *Hash-tree*

Market-Basket Analysis
Non-Binary Data

- Virtual items – represent an “is-a” taxonomy



  - Warning: need to screen for “duh” rules – e.g., *coke* ⇒ *cola*

- Set of indicator variables
  - Real-valued attribute: split into intervals and use one indicator variable for each range
  - Categorical variable
  - Time variable

Market-Basket Analysis
Software Resources

- **arules** package for R
  - Tutorial: http://michael.hahsler.net/research/arules_RUG_2015/demo/

- C++ source code
  - http://adrem.ua.ac.be/~goethals/software/