

Business Intelligence and Data Management

academic year 2019-2020

Dr. Ekaterini Ioannou, Department of Management, University of Tilburg

Lecture 11

topic: Association Rules

material: Chapters 14 (book “Data Mining for Business Intelligence”)

Steps in the DM Process:

Business understanding → Data preparation → Model building → Testing & Evaluation → Deployment

Our last lecture!

Date		Lecture contents	Lecturer	Lab topics	Test
Jan-27	1	Intro. to BI+ Data Management	Caron		
Jan-30				SQL-1	1
Jan-28	2	Data warehousing	Caron		
Feb-06				SQL-2	2
Feb-03	3	OLAP business databases & dashboard	Caron		
Feb-13				SQL-3 & 8	
Feb-10	4	Data mining introduction	Ioannou		
	5	Regression models	Ioannou		
Feb-17	6	Naïve Bayes	Ioannou		
	7	k nearest neighbors	Ioannou		
Feb-20				Naïve Bayes	
Feb-27	8	Performance measures	Ioannou		
Mar-02	9	Decision trees	Ioannou		
Mar-05				Dec. trees	5
Mar-09	10	Association rules	Ioannou		
Mar-11,12&13				Ass. Rules	6
Mar-16	11	Clustering (+20 mins exam preparation)	Ioannou		
Mar-19				Clustering	7

exam preparation will be done when the university provides the details

Association Rules

- Identify **item clusters** in event-based or transaction-based databases
- Study of “**what goes with what**”
 - Symptoms related to diagnosis
 - Customers who bought X also bought Y
- Originated with study of customer transactions databases to determine associations among items purchased
- Association Rules also called:
 - Market basket analysis
 - Affinity analysis



Association Rules

- Identify **item clusters** in event-based or transaction-based databases
- **Usage:** display items together, recommendations in online shopping, etc.
- Heavily used in retail for learning about items that are purchased together
- Useful in several other fields:
 - A medical researcher might want to learn what symptoms appear together
 - In law, word combinations that appear too often might indicate plagiarism

Example

Market basket databases

- Consist of a large number of transaction records
- Each record lists all items bought by a customer on a single-purchase transaction
- Detect certain groups of items are consistently purchased together

Information can be used to

- Make decisions on store layouts
- Design the upcoming catalog
- Identify customer segments based on buying patterns
- ...

Example


Market basket databases

- Consist of a large number of transaction records
- Each record lists all items bought by a customer on a single-purchase transaction
- Detect certain groups of items are consistently purchased together

Amazon uses information for

- Recommendations!!!

Cell Phones & Accessories > Cell Phones > Unlocked Cell Phones





Samsung Galaxy S5 SM-G900H 16GB Factory Unlocked International Version - WHITE
by Samsung
★★★★☆ • 926 customer reviews | 731 answered questions

List Price: ~~\$599.99~~
Price: \$396.98
You Save: \$203.01 (34%)

In Stock.
Sold by DeltaMobiles and Fulfilled by Amazon.

This item does not ship to **Taipei City, Taiwan; Republic of China**. Please check other sellers who may ship internationally. Learn more
Color: **White**



 \$409.99  \$399.75  \$396.98

- 5.1" Full HD Super AMOLED(TM) (1080 x 1920)
- Exynos Quad Core; 1.9GHz, 1.3GHz
- 16 MP Camera with LED Flash
- Must be activated with an Americas-region SIM
- 16GB of Internal Memory
- Unlocked cell phones are compatible with GSM carriers but are not compatible with CDMA Carriers.

14 new from \$381.87 33 used from \$315.00
21 refurbished from \$345.99

Roll over image to zoom in

Frequently Bought Together

 +  +  **Price for all three: \$422.96**

[Add all three to Cart](#)
[Add all three to Wish List](#)
[Show availability and shipping details](#)

- ✓ **This item:** Samsung Galaxy S5 SM-G900H 16GB Factory Unlocked International Version - WHITE \$396.98
- ✓ Galaxy S5 Case, Spigen Slim Armor Case for Galaxy S5 - Shimmery White (SGP10755) \$16.99
- ✓ Galaxy S5 Screen Protector, Spigen [Full HD] Samsung Galaxy S5 Screen Protector [Crystal ... \$8.99


Customers Who Bought This Item Also Bought




[Galaxy S5 Screen Protector] Poweradd™ Tempered Glass HD Clear Screen Protector Guard...
★★★★☆ 515
\$7.99 Prime



Samsung Galaxy S5 G900H 16GB Unlocked GSM Octa-Core Android Smartphone - Black
★★★★☆ 411
\$387.00 Prime



Galaxy S5 Screen Protector, Spigen [Full HD] Samsung Galaxy S5 Screen Protector...
★★★★☆ 428
\$8.99 Prime



MPERO Collection 5 Pack of Ultra Clear Screen Protectors for Samsung Galaxy S5 / GS5
★★★★☆ 1,489
\$0.99

Page 1 of 25

Rules

- Represented in an IF-THEN format
 - “IF” part: antecedent
 - “THEN” part: consequent
- Both correspond to sets of items (called itemsets)
- *Itemsets* are
 - Possible combinations of items (e.g., products)
 - Can also be a single item
 - NOT records of what people buy
- Antecedent and consequent are disjoint
 - I.e., have no items in common

Example

itemset {red, white, green}

- Transaction 1 supports several rules:
 - IF red THEN white
 - Meaning: if a red faceplate is purchased then so is a white one
 - IF red AND white THEN green
 - etc.

- ...

itemset {white, blue}

- Transaction 3:
 - IF white THEN blue
 - ...

- ...

Transaction	Faceplate Colors Purchased			
1	red	white	green	
2	white	orange		
3	white	blue		
4	red	white	orange	
5	red	blue		
6	white	blue		
7	white	orange		
8	red	white	blue	green
9	red	white	blue	
10	yellow			

Finding Association Rules

- One itemset has many association rules
 - Every transaction is one itemset
- Supports several rules

Two-stage Process:

1. Generation of frequent itemsets
i.e., Apriory algorithm
2. Selecting the strong rules
i.e., criteria for judging the strength of the rules

Generation of frequent itemsets

Selecting the strong rules

Generation of rules

→ Detect candidates for indicating item associations

Ideal process (check all possible combinations):

- Find all combinations of single items, pairs of items, triplets of items, and so on
 - I.e., {a, b, c} items means: {a}, {b}, {c}, {a, b}, {a, c}, ...
- Requires a long computation time (exponential)

Practical solution:

- Consider only combinations that occur with higher frequency in the transactions, i.e., data set
- Called *frequent itemsets*
- Criterion for frequent is “support”

Frequent Itemsets

IF Antecedent THEN Consequent

- Combinations of items that occur with higher frequency among the transactions
- Criterion for frequent is “support”
- Support: number, or percent, of transactions that include both the antecedent and the consequent
- An itemset that has a support that exceeds a selected minimum support, determined by the user

$$\text{support}(A \cup C) = \text{frequency}(A \cap C)$$

or, as a percentage

$$\text{support}(A \cup C) = \frac{\text{frequency}(A \cap C)}{n}$$

Example

Support: number, or percent, of transactions that include both the antecedent and the consequent

$$\text{support}(A \cap C) \\ = \text{frequency}(A \cap C)$$

Transaction	Faceplate Colors Purchased			
1	red	white	green	
2	white	orange		
3	white	blue		
4	red	white	orange	
5	red	blue		
6	white	blue		
7	white	orange		
8	red	white	blue	green
9	red	white	blue	
10	yellow			



What is the support for the following rules?

- a) IF Red THEN White
- b) IF White THEN Red
- c) If White and Red THEN Green

Support for itemset {red, white} is 4, or as percentage 4 out of 10 transactions, i.e., 40%

Apriori algorithm

Goal: generate the frequent itemsets

For k items:

- User sets a minimum support criterion
- Generate list of one-item sets
- Drop the ones below the support criterion
- Use the list of one-itemsets to generate the two-itemsets
- Drop the ones below the support criterion
- Use the list of two-itemsets to generate the three-itemsets
- Drop the ones below the support criterion
- (continue until k -itemsets)

Assessment of rule strength

- We need to **measure** the strength of the association implied by a rule
- Measures:
 - **Support:**
 - Number of transactions that include all items from the antecedent and consequent
 - (explained in the previous slides)
 - **Confidence**
 - **Lift ratio**

Confidence

IF Antecedent THEN Consequent

Compares the co-occurrence of items in antecedent and consequent to the occurrence of items in antecedent

Shows the percentage in which C appears with A

$$\text{confidence} = \frac{\text{no. transactions with both antecedent and consequent itemsets}}{\text{no. transactions with antecedent itemset}}$$

$$= \frac{\text{frequency}(A \cap C)}{\text{frequency}(A)}$$

Example:

- Supermarket with 100.000 transactions
- 2.000 include both orange juice and flu medication
- 800 from the above include soup purchases
- IF orange juice and flu medication are purchased THEN soup
- Support = 800 transactions (or $800/100.000 = 0.8\%$)
- Confidence = $800 / 2.000 = 40\%$

Relationship of Support with Confidence

IF Antecedent THEN Consequent

Support:

- (Estimated) probability that a transaction selected randomly from the database will contain all items in the antecedent and the consequent
- $\hat{P}(\text{antecedent AND consequent})$

Confidence:

- (Estimated) conditional probability that a transaction selected randomly will include all the items in the consequent given that the transaction includes all the items in the antecedent
- $\hat{P}(\text{antecedent} \mid \text{consequent})$

Relationship of Support with Confidence

- Support: \hat{P} (antecedent AND consequent)
- Confidence: \hat{P} (antecedent | consequent)
- High value of **confidence** suggests a strong association rule, i.e., rule in which we are highly confident
- Can be deceptive when antecedent and consequent are independent, e.g.:
 - Nearly all customers buy bananas and nearly all customers buy ice cream
 - High confidence level of “IF bananas THEN ice-cream”
 - Regardless of whether there is an association between the items

Lift Ratio

- Better way to judge the strength of a rule
- Compares the confidence of the rule with a benchmark value
- **Confidence**: percentage of antecedent transactions that also have the consequent item set
- **Lift**: ratio of confidence with benchmark confidence
- **Benchmark confidence**: transactions with consequent as percentage of all transactions

- Better way to judge the strength of a rule
- Compares the confidence of the rule with a benchmark value
- Assumes independence of the consequent from the antecedent

$$\text{lift} = \frac{\text{confidence}}{\text{benchmark confidence}}$$

$$\text{benchmark confidence} = \frac{\text{no. transactions with consequent itemset}}{\text{no. transactions in data set}}$$

$$= \frac{\text{frequency}(C)}{n}$$

Intuition

IF Antecedent THEN Consequent

- Lift is a value between 0 and infinity
- $\text{Value} > 1$ indicates that antecedent and consequent are dependent on each other, and the degree of which is given by its value
- $\text{Value} < 1$ indicates that the presence of antecedent will have negative effect on consequent
- $\text{Value} \approx 1$ indicates that antecedent and consequent are independent and no rule can be derived from them

Value > 1

- Suggests that there is some usefulness to the rule
- Level of association between the antecedent and consequent itemsets is higher than would be expected if they were independent
- The larger the lift ratio, the greater the strength of the association

Alternative data representation

- Previous slides showed transaction databases:
 - I.e., each row was the list of purchased items
- Binary incidence matrix:
 - Columns are items
 - Rows are transactions
 - Cells indicate the present or absent of items in transactions

Transaction	Faceplate Colors Purchased				Transaction	Red	White	Blue	Orange	Green	Yell
1	red	white	green		1	1	1	0	0	1	0
2	white	orange			2	0	1	0	1	0	0
3	white	blue			3	0	1	1	0	0	0
4	red	white	orange		4	1	1	0	1	0	0
5	red	blue			5	1	0	1	0	0	0
6	white	blue			6	0	1	1	0	0	0
7	white	orange			7	1	0	1	0	0	0
8	red	white	blue	green	8	1	1	1	0	1	0
9	red	white	blue		9	1	1	1	0	0	0
10	yellow				10	0	0	0	0	0	1

Binary incidence matrix

Find all rules with a support count of at least 2

→ equivalent to a percentage 20% (i.e., 2/10)

→ rules with items that were purchased together in at least 20% of the transactions

Transaction	Red	White	Blue	Orange	Green	Yellow
1	1	1	0	0	1	0
2	0	1	0	1	0	0
3	0	1	1	0	0	0
4	1	1	0	1	0	0
5	1	0	1	0	0	0
6	0	1	1	0	0	0
7	1	0	1	0	0	0
8	1	1	1	0	1	0
9	1	1	1	0	0	0
10	0	0	0	0	0	1

Compute **support** of itemset:

- {red}
- {yellow}
- {red, blue}
- {red, white, blue}

Itemset	Support (Count)
{red}	6
{white}	7
{blue}	6
{orange}	2
{green}	2
{red, white}	4
{red, blue}	4
{red, green}	2
{white, blue}	4
{white, orange}	2
{white, green}	2
{red, white, blue}	2
{red, white, green}	2

Example

IF White & Red THEN Green

Support:

- I.e., $\text{frequency}(A \cap C) = \text{frequency}(\{\text{White, Red, Green}\})$
- Answer = 2

Confidence:

- I.e., $\text{frequency}(A \cap C) / \text{frequency}(A)$
- Answer = $2 / 4 = 0.5$

Lift:

- I.e., $\text{confidence} / \text{frequency}(C) / n$
- Answer = $0.5 / (2/10) = 2.5$

Transaction	Red	White	Blue	Orange	Green	Yellow
1	1	1	0	0	1	0
2	0	1	0	1	0	0
3	0	1	1	0	0	0
4	1	1	0	1	0	0
5	1	0	1	0	0	0
6	0	1	1	0	0	0
7	1	0	1	0	0	0
8	1	1	1	0	1	0
9	1	1	1	0	0	0
10	0	0	0	0	0	1

Quiz

Transaction	Red	White	Blue	Orange	Green	Yellow
1	1	1	0	0	1	0
2	0	1	0	1	0	0
3	0	1	1	0	0	0
4	1	1	0	1	0	0
5	1	0	1	0	0	0
6	0	1	1	0	0	0
7	1	0	1	0	0	0
8	1	1	1	0	1	0
9	1	1	1	0	0	0
10	0	0	0	0	0	1



Consider the above transactions and rule
IF Red & Green THEN White. Compute:

- a) Support
- b) Confidence
- c) Lift

Summary

- Association rules produce rules on associations between items from a data sets with transactions
- Widely used in recommender systems
- Most popular method is Apriori algorithm
- To reduce computation, we consider only “frequent” itemsets (i.e., support)
- Performance is measured by confidence and lift
- Can produce a profusion of rules; review is required to identify useful rules and to reduce redundancy

Quiz

Transaction	Red	White	Blue	Orange	Green	Yellow
1	1	1	0	0	1	0
2	0	1	0	1	0	0
3	0	1	1	0	0	0
4	1	1	0	1	0	0
5	1	0	1	0	0	0
6	0	1	1	0	0	0
7	1	0	1	0	0	0
8	1	1	1	0	1	0
9	1	1	1	0	0	0
10	0	0	0	0	0	1

Consider the above transactions and rule
IF Red & Green THEN White. Compute:



- a) Support, number of transactions with {Red, Green, White} $\rightarrow 2$ or $2/10=0.2$
- b) Confidence, co-occurrence / no of antecedent $\rightarrow 2 / 2 = 1$
- c) Lift, confidence / (no consequent / no transactions) $= 1 / 7 / 10 = 1.428$