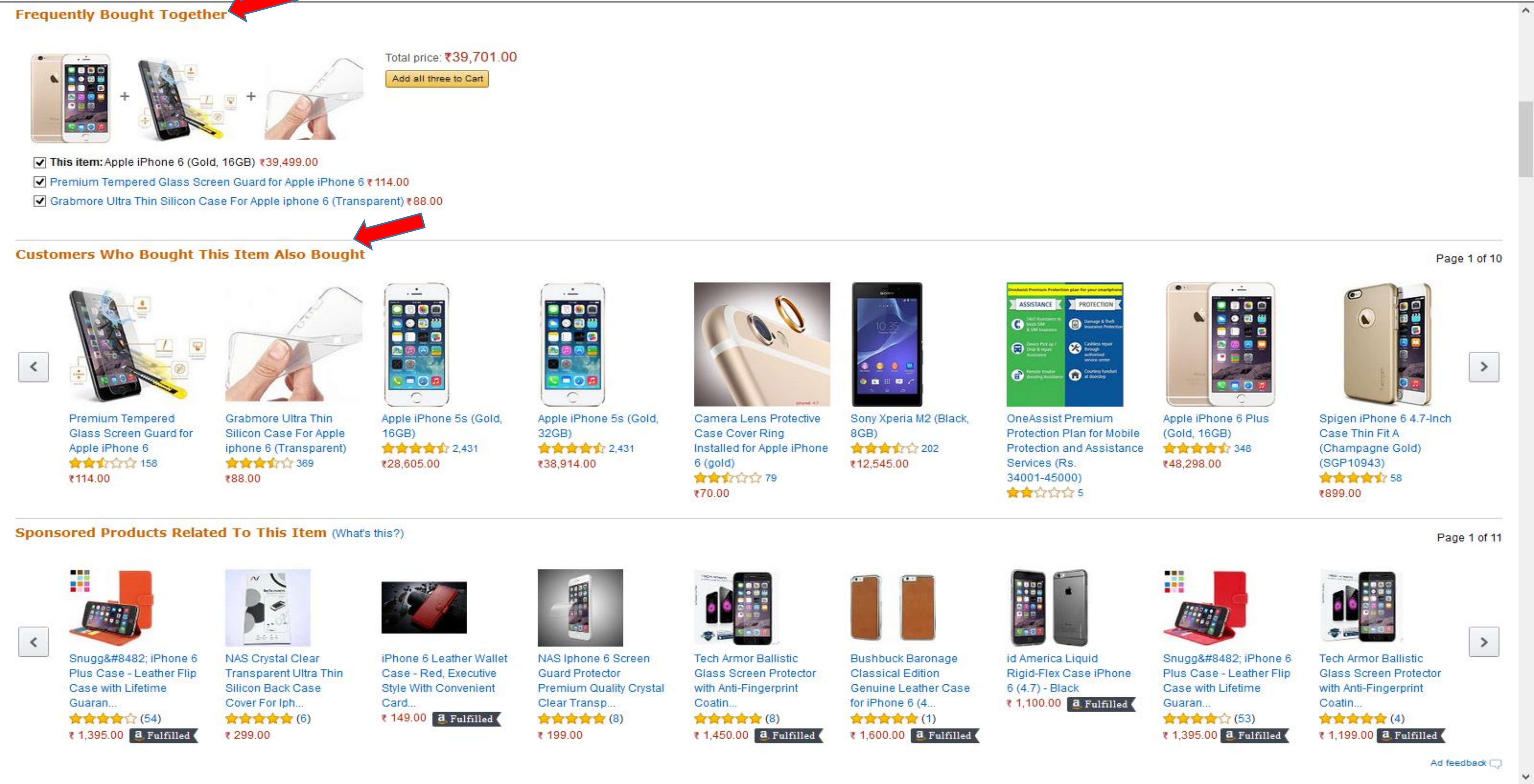
Association Rules: An Introduction

Online Recommendations



Data Mining Methods and Nature of Data

TABLE 1.1 **ORGANIZATION OF DATA MINING METHODS IN THIS BOOK, ACCORDING TO THE NATURE OF THE DATA***

	Supervised: Continuous Response	Supervised: Categorical Response	Unsupervised: No Response
Continuous predictors	Linear regression (6) Neural nets (11) k-Nearest neighbors (7) Ensembles (13)	Logistic regression (10) Neural nets (11) Discriminant analysis (12) k-Nearest neighbors (7) Ensembles (13)	Principal components (4) Cluster analysis (15) Collaborative filtering (14)
Categorical predictors	Linear regression (6) Neural nets (11) Regression trees (9) Ensembles (13)	Neural nets (11) Classification trees (9) Logistic regression (10) Naive Bayes (8) Ensembles (13)	Association rules (14) Collaborative filtering (14)

***Numbers in parentheses indicate chapter number.**

Ref: Data Mining for Business Analytics: Concepts, Techniques and Applications in R, by Galit Shmueli et al., Wiley India, 2018.

What are Association Rules?

- Study of “what goes with what”
  - “Customers who bought X also bought Y”
  - What symptoms go with what diagnosis
- Transaction-based or event-based
- Also called “market basket analysis” and “affinity analysis”
- Originated with study of customer transactions databases to determine associations among items purchased

Generating Rules: Terms

"IF" part = **antecedent**

"THEN" part = **consequent**

"Item set" = the items (e.g., products) comprising the antecedent or consequent

- Antecedent and consequent are *disjoint* (i.e., have no items in common)

Illustrative Example: Phone Faceplates

Transaction	Color(s) purchased
1	red white green
2	white orange
3	white blue
4	red white orange
5	red blue
6	white blue
7	red blue
8	red white blue green
9	red white blue
10	yellow

Ref: Data Mining for Business Analytics: Concepts, Techniques and Applications in R, by Galit Shmueli et al., Wiley India, 2018.

Many Rules are Possible

For example: Transaction 1 supports several rules, such as

- “If red, then white” (“If a red faceplate is purchased, then so is a white one”)
- “If white, then red”
- “If red and white, then green”
- + several more

Transaction	Color(s) purchased
1	red white green
2	white orange
3	white blue
4	red white orange
5	red blue
6	white blue
7	red blue
8	red white blue green
9	red white blue
10	yellow

Frequent Item Sets

- Ideally, we want to create all possible combinations of items
- **Problem:** computation time grows exponentially as # items increases
- **Solution:** consider only “frequent item sets”
- Criterion for frequent: *support*

Support

Support for an itemset = # (or percent) of transactions that include an itemset

- Example: support for the item set {red, white} is 4 out of 10 transactions, or 40%

Support for a rule = # (or percent) of transactions that include both the antecedent and the consequent

Transaction	Color(s) purchased
1	**red white** green
2	white orange
3	white blue
4	**red white** orange
5	red blue
6	white blue
7	red blue
8	**red white** blue green
9	**red white** blue
10	yellow

Confidence

- Confidence of the rule : A measure that expresses the degree of uncertainty about the if-then rule.
- Compares the co-occurrence of the antecedent and consequent item sets in the database to the occurrence of the antecedent item sets.
- Confidence is defined as the ratio of the number of transactions that include all antecedent and consequent item sets (namely, the support) to the number of transactions that include all the antecedent item sets:
- $$\textit{Confidence} = \frac{\text{\# Transactions with both antecedent and consequent item sets}}{\text{\# Transactions with antecedent item set}}$$

Support and Confidence

- For example, suppose a supermarket database has 100,000 point-of-sale transactions.
- Of these transactions, 20,000 include both "Modern Bread" and "Amul Butter", and 8000 of these include "Kissan Mixed-Fruit Jam".
- The association rule

IF "Modern Bread" and "Amul Butter" are purchased THEN "Kissan Jam" is purchased on the same trip has a support of 8000 transactions

(alternatively 8% = 8000/100,000)

and a confidence of 40% (= 8000/20000).

Support for a rule = # (or percent) of transactions that include both the antecedent and the consequent

$$\textit{Confidence} = \frac{\text{\# Transactions with both antecedent and consequent item sets}}{\text{\# Transactions with antecedent item set}}$$

Confidence - Caveats

- A high value of confidence suggests a strong association rule (in which we are highly confident).

- However, this can be deceptive because if the antecedent and/or the consequent has a high level of support, we can have a high value for confidence even when the antecedent and consequent are independent!

- For example, if nearly all customers buy bananas and nearly all customers buy ice cream, the confidence level of a rule such as “IF bananas THEN ice-cream” will be high regardless of whether there is an association between the items.

Lift Ratio

- A better way to judge the strength of an association rule is to compare the confidence of the rule with a benchmark value, where we assume that the occurrence of the consequent item set in a transaction is independent of the occurrence of the antecedent for each rule.

- The estimate of this benchmark from the data, called the *benchmark confidence value* for a rule is computed by
- *Benchmark confidence* = $\dfrac{\text{\# Transactions with consequent item set}}{\text{\# Transactions in database}}$

Lift Ratio

- We compare the confidence to the benchmark confidence by looking at their ratio: this is called the *lift ratio* of a rule.

- The lift ratio is the confidence of the rule divided by the confidence, assuming independence of consequent from antecedent.

  lift ratio = confidence / benchmark confidence

- A lift ratio greater than 1.0 suggests that there is some usefulness to the rule. In other words, the level of association between the antecedent and consequent item sets is higher than would be expected if they were independent.
- The larger the lift ratio, the greater the strength of the association.

Apriori Algorithm

Generating Frequent Item Sets

For *k* products…

1. User sets a minimum support criterion
2. Next, generate list of one-item sets that meet the support criterion
3. Use the list of one-item sets to generate list of two-item sets that meet the support criterion
4. Use list of two-item sets to generate list of three-item sets
5. Continue up through *k*-item sets

Measures of Rule Performance

***Confidence*:** the % of antecedent transactions that also have the consequent item set

Benchmark confidence = transactions with consequent as % of all transactions

Lift = *confidence*/(*benchmark confidence*)

Lift > 1 indicates a rule that is useful in finding consequent items sets (i.e., more useful than just selecting transactions randomly)

Leverage = *P(antecedent AND consequent) - P(antecedent) x P(consequent)*

- Leverage = 0 when the two items are independent. It ranges from -1 (antecedent and consequent are antagonistic) to +1 (antecedent makes consequent more likely). In a sales setting, leverage tells us how much more frequently the items are bought together compared to their independent sales

Conviction = *P(antecedent x **NOT** consequent) / P(antecedent AND **NOT** consequent)*

is similar to confidence and ranges from 0 to ∞. If antecedent and consequent are independent, conviction is equal to 1. If the rule always holds (the items always appear together), its value is infinity

Alternate Data Format: Binary Matrix

Transaction	**Color(s) purchased**
1	red white green
2	white orange
3	white blue
4	red white orange
5	red blue
6	white blue
7	red blue
8	red white blue green
9	red white blue
10	yellow

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

Support for Various Itemsets

Transaction	Color(s) purchased
1	red white green
2	white orange
3	white blue
4	red white orange
5	red blue
6	white blue
7	red blue
8	red white blue green
9	red white blue
10	yellow

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

TABLE 14.3 ITEMSETS WITH SUPPORT COUNT OF AT LEAST TWO

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

Process of Rule Selection

TABLE 14.3 ITEMSETS WITH SUPPORT COUNT OF AT LEAST TWO

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
[illegible]	2

Generate all rules that meet specified support & confidence

- Find frequent item sets (those with sufficient support of 2 here)
- From these item sets, generate rules with sufficient confidence

Confidence = <u># Transactions with both antecedent and consequent item sets</u>

Transactions with antecedent item set

Eg. Rules from the frequent itemset {red, white, green}

Rule	Confidence	Lift
{red, white} → {green}	$\frac{\text{support of \{red, white, green\}}}{\text{support of \{red, white\}}} = 2/4 = 50\%$	$\frac{\text{confidence of rule}}{\text{benchmark confidence}} = \frac{50\%}{20\%} = 2.5$
{green} → {red}	$\frac{\text{support of \{green, red\}}}{\text{support of \{green\}}} = 2/2 = 100\%$	$\frac{\text{confidence of rule}}{\text{benchmark confidence}} = \frac{100\%}{60\%} = 1.67$
{white, green} → {red}	$\frac{\text{support of \{white, green, red\}}}{\text{support of \{white, green\}}} = 2/2 = 100\%$	$\frac{\text{confidence of rule}}{\text{benchmark confidence}} = \frac{100\%}{60\%} = 1.67$

Plus 3 more with confidence of 33%, 29% & 100% , If confidence criterion is 70%, report only rules 2, 3 and 6

Apriori Algorithm Numerical example

Transaction ID	Items
T1	I1, I3, I4
T2	I2, I3, I5, I6
T3	I1, I2, I3, I5
T4	I2, I5
T5	I1, I3, I5

Dataset for Apriori Algorithm

Find association rules from the above dataset. Use the minimum support count of 2 and minimum confidence of 75 percent.

	I1	I2	I3	I4	I5	I6
T1	1	0	1	1	0	0
T2	0	1	1	0	1	1
T3	1	1	1	0	1	0
T4	0	1	0	0	1	0
T5	1	0	1	0	1	0

Transaction Matrix

minimum support count 2

Create Frequent Itemsets With One Item

Itemset	Support Count
{I1}	3
{I2}	3
{I3}	4
{I4}	1
{I5}	4
{I6}	1

Candidate Itemsets with one item

Itemset	Support Count
{I1}	3
{I2}	3
{I3}	4
{I5}	4

Frequent itemsets with one item

Create Frequent Itemsets With Two Items

Itemset	Support Count
{I1,I2}	1
{I1,I3}	3
{I1,I5}	2
{I2,I3}	2
{I2,I5}	3
{I3,I5}	3

Itemset	Support Count
{I1,I3}	3
{I1,I5}	2
{I2,I3}	2
{I2,I5}	3
{I3,I5}	3

Frequent itemsets with two items

Calculate Frequent Itemsets With Three Items

Itemset	Subsets	All the subsets are frequent itemsets?
{I1, I3, I5}	{I1, I3},{I1, I5},{I3, I5}	Yes
{I1, I2, I3}	{I1, I2}, {I1, I3}, {I2, I3}	No
{I1, I2, I5}	{I1, I2},{I1, I5},{I2, I5}	No
{I2, I3, I5}	{I2, I3}, {I2, I5}, {I3, I5}	Yes

Pruning itemsets with three items

Itemset	Support Count
{I2, I3, I5}	2
{I1, I3, I5}	2

Candidate itemsets with three items

Itemset	Support Count
{I2, I3, I5}	2
{I1, I3, I5}	2

Frequent itemsets with three items

Calculate Frequent Itemsets With Four Items

- {I1, I2, I3, I5}
- Itemset has four subsets with three elements i.e. {I2, I3, I5},{I1, I3, I5}, {I1, I2, I5}, {I1, I2, I3}
- {I1, I2, I5} and {I1, I2, I3} are not frequent itemsets.
- Prune the itemset {I1, I2, I3, I5}.
- No candidate set for itemsets with 4 items

Frequent itemsets in the dataset

Itemset	Support Count
{I1}	3
{I2}	3
{I3}	4
{I5}	4
{I1,I3}	3
{I1,I5}	2
{I2,I3}	2
{I2,I5}	3
{I3,I5}	3
{I2, I3, I5}	2
{I1, I3, I5}	2

Generate Association Rules From Frequent Itemsets

Itemset	Association Rules
{I1}	x
{I2}	x
{I3}	x
{I5}	x
{I1,I3}	{I1}->{I3}, {I3}->{I1}
{I1,I5}	{I1}->{I5}, {I5}->{I1}
{I2,I3}	{I2}->{I3}, {I3}->{I2}
{I2,I5}	{I2}->{I5}, {I5}->{I2}
{I3,I5}	{I3}->{I5}, {I5}->{I3}
{I2, I3, I5}	{I2}->{I3, I5}, {I3}->{I2, I5}, {I5}->{I2, I3}, {I2, I3}->{I5}, {I2, I5}->{I3}, {I3, I5}->{I2}
{I1, I3, I5}	{I1}->{I3, I5}, {I3}->{I1, I5}, {I5}->{I1, I3}, {I1, I3}->{I5}, {I1, I5}->{I3}, {I3, I5}->{I1}

Association Rule	Confidence
{I1}->{I3}	100%
{I3}->{I1}	75%
{I2}->{I3}	66.67%
{I3}->{I2}	50%
{I1}->{I5}	66.67%
{I5}->{I1}	50%
{I2}->{I5}	100%
{I5}->{I2}	75%
{I3}->{I5}	75%
{I5}->{I3}	75%
{I2}->{I3, I5}	66.67%
{I3}->{I2, I5}	50%
{I5}->{I2, I3}	50%
{I2, I3}->{I5}	100%
{I2, I5}->{I3}	66.67%
{I3, I5}->{I2}	66.67%
{I1}->{I3, I5},	66.67%
{I3}->{I1, I5},	50%
{I5}->{I1, I3},	50%
{I1, I3}->{I5},	66.67%
{I1, I5}->{I3},	100%
{I3, I5}->{I1},	66.67%

Association rules with confidence

Association Rule	Confidence
{I1}->{I3}	100%
{I3}->{I1}	75%
{I2}->{I5}	100%
{I5}->{I2}	75%
{I3}->{I5}	75%
{I5}->{I3}	75%
{I2, I3}->{I5}	100%
{I1, I5}->{I3},	100%

Interpretation

- *Lift ratio* shows how effective the rule is in finding consequents (useful if finding particular consequents are important)
- *Confidence* shows the rate at which consequents will be found (useful in learning costs of promotion)
- *Support* measures overall impact

Apriori Algorithm - Recap

For *k* products…

1. User sets a minimum support criterion
2. Next, generate list of one-item sets that meet the support criterion
3. Use the list of one-item sets to generate list of two-item sets that meet the support criterion
4. Use list of two-item sets to generate list of three-item sets
5. Continue up through *k*-item sets
6. Generate all rules that meet specified support & confidence

Caution: The Role of Chance

Random data can generate apparently interesting association rules. The more rules you produce, the greater this danger.

Rules based on large numbers of records are less subject to this danger.

Rules from some randomly-generated transactions:

	lhs		rhs	support	confidence	lift
[18]	{item.2}	=>	{item.9}	0.08	0.8	1.481481
[89]	{item.2,item.7}	=>	{item.9}	0.04	1.0	1.851852
[104]	{item.3,item.4}	=>	{item.8}	0.04	1.0	1.851852
[105]	{item.3,item.8}	=>	{item.4}	0.04	1.0	5.000000
[113]	{item.3,item.7}	=>	{item.9}	0.04	1.0	1.851852
[119]	{item.1,item.5}	=>	{item.8}	0.04	1.0	1.851852
[149]	{item.4,item.5}	=>	{item.9}	0.04	1.0	1.851852
[155]	{item.5,item.7}	=>	{item.9}	0.06	1.0	1.851852
[176]	{item.6,item.7}	=>	{item.8}	0.06	1.0	1.851852

Even chance data can produce high lift

Example: Charles Book Club

TABLE 14.7 SUBSET OF BOOK PURCHASE TRANSACTIONS IN BINARY MATRIX FORMAT

ChildBks	YouthBks	CookBks	DoItYBks	cefBks	ArtBks	GeogBks	ItalCook	ItalAtlas	ItalArt	Florence
0	1	0	1	0	0	1	0	0	0	0
1	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0
1	1	1	0	1	0	1	0	0	0	0
0	0	1	0	0	0	1	0	0	0	0
1	0	0	0	0	1	0	0	0	0	1
0	1	0	0	0	0	0	0	0	0	0
0	1	0	0	1	0	0	0	0	0	0
1	0	0	1	0	0	0	0	0	0	0
1	1	1	0	0	0	1	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0

Row 1, e.g., is a transaction in which books were bought in the following categories: Youth, Do it Yourself, Geography

Rules Produced by `apriori`

Output

```
> inspect(sort(rules, by = "lift"))
   lhs                      rhs        support confidence lift
16 {DoItYBks,GeogBks}    => {YouthBks} 0.05450 0.5396040  2.264864
18 {CookBks,GeogBks}     => {YouthBks} 0.08025 0.5136000  2.155719
13 {CookBks,RefBks}      => {DoItYBks} 0.07450 0.5330948  2.092619
14 {YouthBks,GeogBks}    => {DoItYBks} 0.05450 0.5215311  2.047227
20 {YouthBks,CookBks}    => {DoItYBks} 0.08375 0.5201863  2.041948
10 {YouthBks,RefBks}     => {CookBks}  0.06825 0.8400000  2.021661
15 {YouthBks,DoItYBks}   => {GeogBks}  0.05450 0.5278450  1.978801
19 {YouthBks,DoItYBks}   => {CookBks}  0.08375 0.8111380  1.952197
12 {DoItYBks,RefBks}     => {CookBks}  0.07450 0.8054054  1.938400
11 {RefBks,GeogBks}      => {CookBks}  0.06450 0.7889908  1.898895
17 {YouthBks,GeogBks}    => {CookBks}  0.08025 0.7679426  1.848237
21 {DoItYBks,GeogBks}    => {CookBks}  0.07750 0.7673267  1.846755
7  {YouthBks,ArtBks}     => {CookBks}  0.05150 0.7410072  1.783411
9  {DoItYBks,ArtBks}     => {CookBks}  0.05300 0.7114094  1.712177
3  {RefBks}              => {CookBks}  0.13975 0.6825397  1.642695
8  {ArtBks,GeogBks}      => {CookBks}  0.05525 0.6800000  1.636582
4  {YouthBks}            => {CookBks}  0.16100 0.6757608  1.626380
6  {DoItYBks}            => {CookBks}  0.16875 0.6624141  1.594258
1  {ItalCook}            => {CookBks}  0.06875 0.6395349  1.539193
5  {GeogBks}             => {CookBks}  0.15625 0.5857545  1.409758
2  {ArtBks}              => {CookBks}  0.11300 0.5067265  1.219558
```

Cautions:

Duplication (same trio of books)

No useful info!

Summary – Association Rules

- Association rules (or *affinity analysis,* or *market basket analysis*) produce rules on associations between items from a database of transactions
- Widely used in **recommender systems**
- Most popular method is **Apriori algorithm**
- To reduce computation, we consider only “frequent” item sets (=support)
- Performance of rules is measured by *confidence* and *lift*
- Can produce a profusion of rules; review is required to identify useful rules and to reduce redundancy

Slide Contents - References

- The contents of this presentation were sourced and assembled from
  - Data Mining for Business Analytics: Concepts, Techniques and Applications in R, by Galit Shmueli et al., Wiley India, 2018.