Slide-1

ASSOCIATION RULES:

. Widely used among Market basket analysis, Relationship mining, Affinity rules.

Slide-2

Market basket analysis:

- Large number of transaction records through data collected using bar-code scanners.

- Each record = All items purchased on a single purchase transaction.

Slide-3

- What item goes with what.

- Are certain groups of items consistently purchased together.

- What business strategies will you device with this knowledge.

Slide-4

- Products shelf placement – a specific product beside another

- Selling of prominent shelves – Slotting Fees

- Stocking – Supply Chain Management

- Price Bundling – Combo offers. How?

Slide-5

Association Rules – Cell phone faceplates:

- A store sells accessories for cellular phones runs a promotion on faceplates.
- *OFFER!*
  Buy multiple faceplates from a choice of 6 different colors & get discount.
- How would you help store managers device strategy to become more profitable.

Slide-6

LIST FORMAT

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			

BINARY MATRIX FORMAT

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	0	1	0	1	0	0
8	1	1	1	0	1	0
9	1	1	1	0	0	0
10	0	0	0	0	0	1

Association Rules are probabilistic "if-then" statements

2 Main Ideas:

Examine all possible "if-then" rule formats

Select rules, which indicates true dependence

Slide-7

Rules for {Red, White, Green}

1. If {Red, White}   then {Green}
2. If {Red, Green}   then {White}
3. If {White, Green}   then {Red}
4. If {Red}   then {White, Green}

5. If {White}   then {Red, Green}
6. If {Green}   then {Red, White}

Below problems might arise

- Many rules are possible

- How to select the TRUE/GOOD rules from all generated rules?

Slide-8

Association Rules – Terminology

"IF" part = Antecedent = A

"THEN" part = Consequent = C

- If {Red, White} then {Green}

- If Red & White phone faceplates are purchased, then Green faceplate is purchased

    . Antecedent: Red & White

    . Consequent: Green

Slide-9

Association Rules – Performance Measures

1) Support
2) Confidence
3) Lift

Slide-10

Association Rules – Support

- Consider only combinations that occur with higher frequency in the database

- Support is the criterion based on frequency

Formula:

Percentage / Number of transactions in which IF/Antecedent & THEN / Consequent appear in the data

Mathematically:

$$= \frac{\text{\# Transactions in which A \& C appear together}}{\text{Total no. of transactions}}$$

Slide-11-   Support – Calculation

ransaction #	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			

1) What is the support for "if white then blue"?

   1) 4   2) 40%   3) 2   4) 90%

2) What is the *support* for "if blue then White"?

   1) 4   2) 40%   3) 2   4) 90%

Slide-12

Support – Problem

- Generating all possible rules is exponential in the number of distinct items

- Solution:

  Frequent item sets using Apriori Algorithm.

Slide-13

Apriori Algorithm for K Products:

- Set minimum support criteria
- Generate list of one-item sets that meet the support criterion
- Use list of one-item sets to generate list of two-item sets that meet support criterion
- Use list of two-item sets to generate list of three-item sets that meet support criterion
- Continue up through k-item sets

Slide-14

Support – Criterion = 2

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			

Create rules from

Frequent item sets only

Item set	Support (Count)
{Red}	5
{White}	8
{Blue}	5
{Orange}	3
{Green}	2
{Red, White}	4
{Red, Blue}	3
{Red, Green}	2
{White, Blue}	4
{White, Orange}	3
{White, Green}	2
{Red, White, Blue}	2
{Red, White, Green}	2

Slide-15

Support Criterion Example

Rules for {Red, White, Green}

1. If {Red, White} then {Green}
2. If {Red, Green} then {White}   How good are these rules
3. If {White, Green} then {Red}   beyond the point
4. If {Red} then {White, Green}   that they have high
5. If {White} then {Red, Green}   support?
6. If {Green} then {Red, White}

Slide-16
Association Rules – Confidence
Confidence:

- Percentage of If/Antecedent transactions that also have the Then/Consequent item set

  Mathematically:

  P (Consequent | Antecedent) = P(C & A) / P (A)

  $$= \frac{\textit{\# Transactions in which A \& C appear together}}{\textit{\# Transactions with A}}$$

Slide-17
Confidence – Calculation:

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			

1) What is the confidence for "if White then blue"?
   a) 4/5   b) 5/8   c) 5/4   d) 4/8
2) What is the *confidence* for "if blue then White"?

a) 4/5   b) 5/8   c) 5/4   d) 4/8

Slide-18

Confidence – Weakness

- If antecedent and consequent have:

    High Support => High / Biased Confidence

    Ex : Antecedent = milk & consequent = banana

Slide-19

Association Rules – Lift Ratio

Lift Ratio

- Confidence / Benchmark confidence
- Benchmark assumes independence between antecedent & consequent

Benchmark confidence

$P (C|A) = P(C \& A) / P (A) = P(C) \times P (A) /P (A) = P(C)$

 # Transactions with consequent item sets

_____

 # Transactions in database

Slide-20

Interpreting Lift:

- Lift > 1 indicates a rule that is useful in finding consequent item sets
- The rule above is much better than selecting random transactions

Slide-21

Lift – Calculation:

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 *Lift* for "if White then blue"?

    a) 4/5   b) 5/10   c) 4/5   d) 1

Slide-22

Rules selection process:

Generate all rules that meet specified Support & Confidence

- Find frequent item sets based on Support specified by applying minimum support cutoff.
- From these item sets, generate rules with defined Confidence. By filtering remaining rules, select only those with high Confidence.

Slide-23

Rules:

Inputs

Data	
# Transactions in Input Data	10
# Columns in Input Data	6
# Items in Input Data	6
# Association Rules	8
Minimum Support	2
Minimum Confidence	70.00%

List of Rules: Rule: If all Antecedent items are purchased, then with Confidence percentage Consequent items will also be purchased.

Row ID	Confidence %	Antecedent (A)	Consequent (C)	Support for A	Support for C	Support for A & C	Lift Ratio
8	100	green	red & white	2	4	2	2.5
4	100	green	red	2	5	2	2
6	100	white & green	red	2	5	2	2
3	100	orange	white	3	8	3	1.25
5	100	green	white	2	8	2	1.25
7	100	red & green	white	2	8	2	1.25
1	80	red	white	5	8	4	1
2	80	blue	white	5	8	4	1

SLIDE-24

Alarming!

- Random data can generate apparently interesting association rules
- More the rules you produce, greater the danger
- Rules based on large numbers of records are less subject to this danger

SLIDE-25

ChildBks	YouthBks	CookBks	DoItYBks	RefBks	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
0	0	1	0	0	0	1	0	0	0	0

Rule #	Conf. %	Antecedent (a)	Consequent (c)	Support(a)	Support(c)	Support(a U c)	Lift Ratio
1	100	ItalCook=>	CookBks	227	862	227	2.320186
2	62.77	ArtBks, ChildBks=>	GeogBks	325	552	204	2.274247
3	54.13	CookBks, DoItYBks=>	ArtBks	375	482	203	2.246196
4	61.98	ArtBks, CookBks=>	GeogBks	334	552	207	2.245509
5	53.77	CookBks, GeogBks=>	ArtBks	385	482	207	2.230964
6	57.11	RefBks=>	ChildBks, CookBks	429	512	245	2.230842
7	52.31	ChildBks, GeogBks=>	ArtBks	390	482	204	2.170444
8	60.78	ArtBks, CookBks=>	DoItYBks	334	564	203	2.155264
9	58.4	ChildBks, CookBks=>	GeogBks	512	552	299	2.115885
10	54.17	GeogBks=>	ChildBks, CookBks	552	512	299	2.115885

SLIDE-26

Applications:

- What if Product & Stores are selected as a tuple for analysis?

- What if crimes in different geographies for each week is known?

SLIDE-27

Recap with an example:

- How can you use the information if you know about the purchase history of customers in a specific geography?

- Supermarket database has 100,000 POS transactions
- 2000 transactions include both Strepsils & Orange Juice
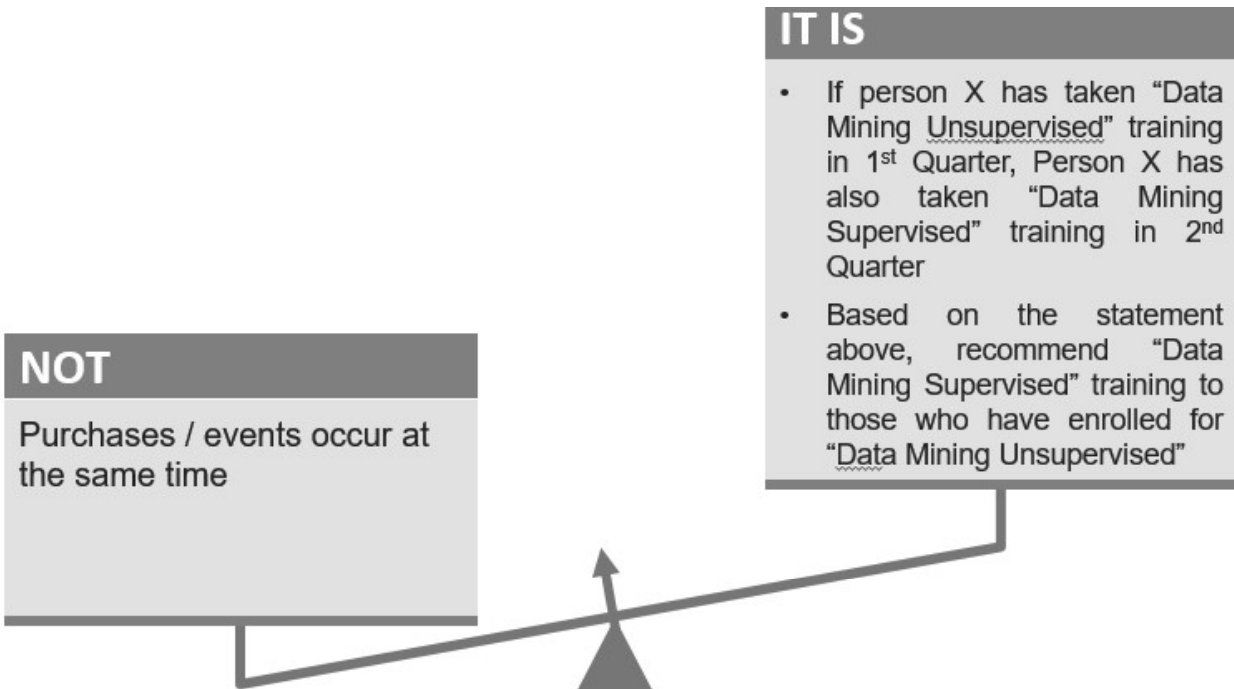- 800 of the above 2000 include Soup purchases

Slide-28

1) What is the *support* for rule "IF (Orange Juice & Strepsils) are purchased THEN (Soup) is purchased on the same trip"?
    a. 0.8 %
    b. 2 %
    c. 40 %

2) What is the *confidence* for rule "IF (Orange Juice & Strepsils) are purchased THEN (Soup) is purchased on the same trip"?

    a. 0.8 %

    b. 2 %

    c. 40 %

Slide-29

What is the *lift ratio* for rule "IF (Orange Juice & Strepsils) are purchased THEN (Soup) is purchased on the same trip"?

SLIDE-30

Sequential Pattern Mining

IT IS

- If person X has taken "Data Mining Unsupervised" training in 1st Quarter, Person X has also taken "Data Mining Supervised" training in 2nd Quarter

- Based on the statement above, recommend "Data Mining Supervised" training to those who have enrolled for "Data Mining Unsupervised"

NOT

Purchases / events occur at the same time

SLIDE-31

Association Rules vs. Sequential Pattern Mining:

- Look for temporal patterns

- Order/sequence of a & b matters for a rule "b follows a"

- However, what happens in between a & b doesn't matter

- In phone faceplates dataset:

  ➢ Association among items, which were bought within the same week were discovered

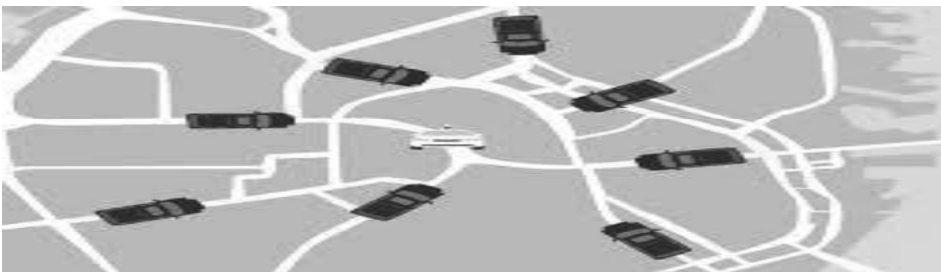➢ How about finding what they would buy next week or the week after, if they had bought 'x' in this week?

SLIDE-32

Applications

- Identify the appropriate Basket
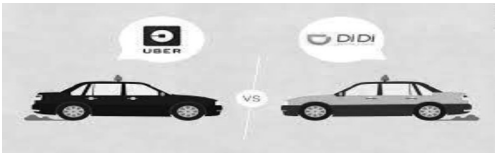
- Identify popular taxi routes

  - Sequential pattern from GPS tracks; spatiotemporal records of taxi trajectories

  - First cluster collocated customers

SLIDE-33

- **Cable TV challenges**
  - Customers are cutting the cord due to cheaper replacements

Challenge 1

> *The fat, pricey cable bundle of 200 channels is fast becoming antiquated as slimmer streaming options emerge. Now two tech giants, Amazon and YouTube (owned by Google), as well as Hulu, a video-streaming service that is jointly owned by Disney, Fox and NBC Universal, are negotiating to offer live television over the internet by the end of the year or early next year. They would offer America's major broadcast networks and many popular sports and entertainment channels, at a price that would cut the typical monthly bill almost in half, to $40 or $50."*
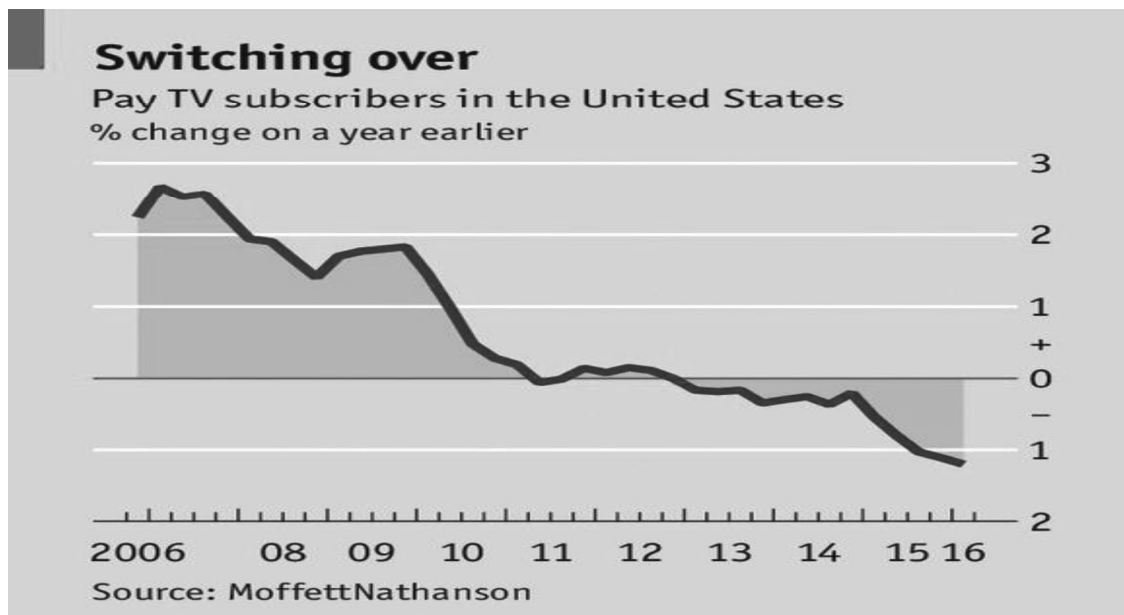
- **Challenge 2**

> *Those that do chop the cord almost never come back, joining the ranks of millennials who avoid signing up for cable in the first place, dubbed "cord-nevers" by media executives. They are lost to the world of subscription video-on-demand: Netflix, Amazon Prime video, Hulu, HBO Now and the like, services that cost around $10 to $15 a month each.*

➢ *To stanch this flow, cable operators can offer "triple-play" packages that combine broadband, television and telephone service, which gives them a pricing advantage. They can also rely on older Americans. Older viewers watch more television than any other group, they watch more of it than they used to, and more are tuning in; and they are not going anywhere.*

■ **ASSOCIATION RULES**

➢ *Hulu is now testing channel combinations at various prices, including around $40 to $50 a month, close to similar packages from Sling TV and Sony PlayStation Vue.*

■ **Bundling Price**

Switching over
Pay TV subscribers in the United States
% change on a year earlier

2006   08   09   10   11   12   13   14   15 16
Source: MoffettNathanson

Economist.com

THANK YOU