Interestingness of Interestingness measures

Simrat Hanspal
Data Scientist
Mad Street Den

What to expect

- What are we mining for?
- Problems mining associations in retail
- How to mine gold in retail domain
- Key takeaways

Interesting associations

- Association mining is of interest in many domains e.g. Bioinformatics, Web Mining, Text Mining, Retail, Fraud detection etc
- Association between item X and Y can defined on multiple actions e.g. Co-Occurrence, Co-Purchase etc
- Depicted as X —> Y
  Where X and Y are disjoint sets

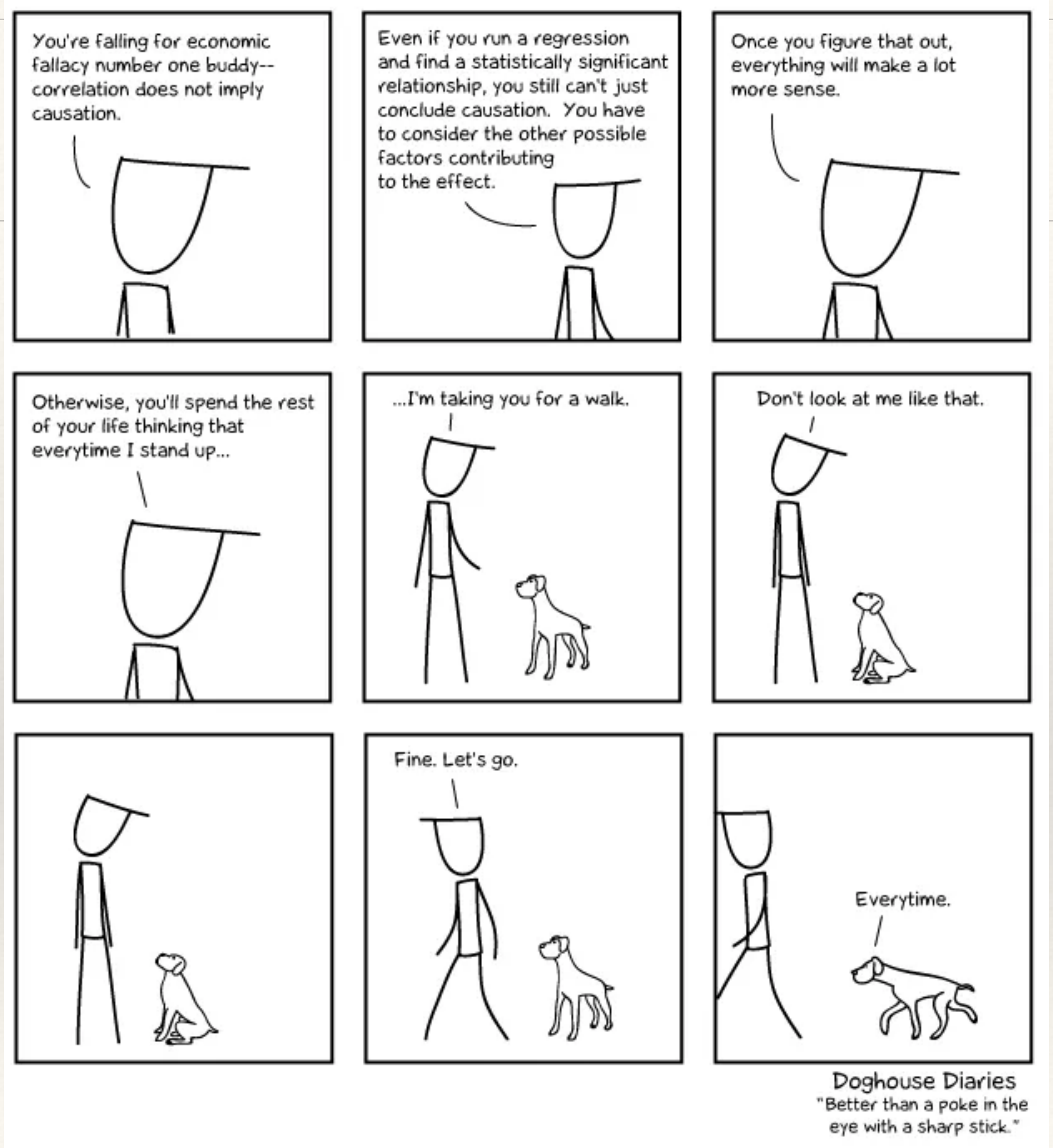
Expectedness	Usefulness	Interestingness	Example
Expected	Useful	Interesting	Bread & Butter
Expected	Not Useful	X	Bread & Morning News
Unexpected	Useful	Very Interesting	Diaper & Beer
Unexpected	Not Useful	X	Diaper & Beer to wrong customer segment

Motivational Example - Market Basket Analysis

- Retail organisations collect huge amount of transactional data
- Mining Unexpected and useful associations provide new opportunity for cross sell
- Popular example -
    - Diapers & Beer

Correlation doesn’t imply Causation

- Strong correlation doesn’t imply causation
- Strongly correlated rules can be used to
  - Grow domain knowledge
  - Increase customer interaction and sales


You're falling for economic fallacy number one buddy-- correlation does not imply causation.
Even if you run a regression and find a statistically significant relationship, you still can't just conclude causation. You have to consider the other possible factors contributing to the effect.
Once you figure that out, everything will make a lot more sense.
Otherwise, you'll spend the rest of your life thinking that everytime I stand up...
...I'm taking you for a walk.
Don't look at me like that.
Fine. Let's go.
Everytime.
Doghouse Diaries
"Better than a poke in the eye with a sharp stick."

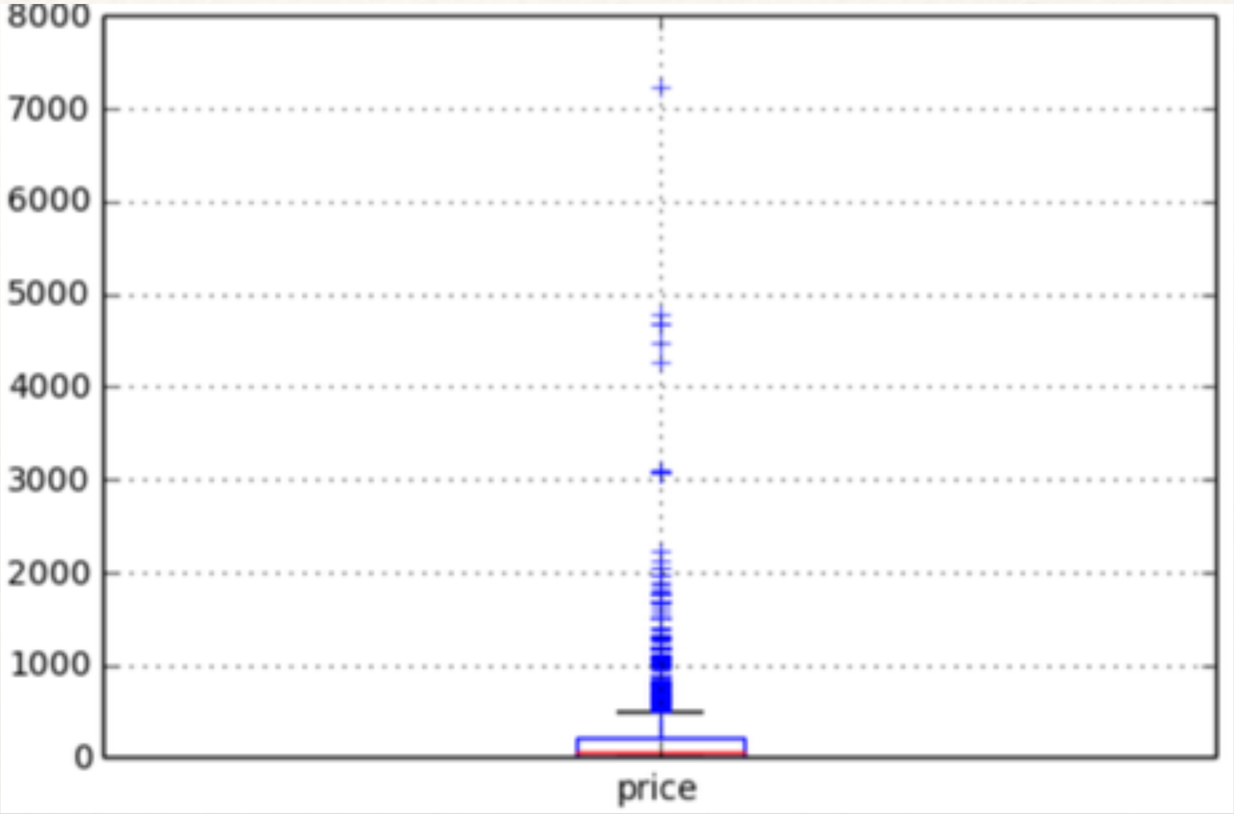
Problems of Associations Mining in Retail domain

Data Sparsity

- Small scale client with 2500+ products
- Total # of possible associations = 2500*2500
  = 6250000 (6.25 Million)
- Number of associations seen in 3 months = 52,000
- 0.83% of the total # of possible associations
- Data sparsity also leads to a lot of associations left undiscovered.

Less frequent but Important transactions

- In market basket analysis, products with low frequency get filtered out.
- But these can be expensive products such as jewellery
- Which makes them rare but interesting associations for mining



Spurious Associations

- Low priced products have high frequency, they can get viewed with many unrelated products too.
- Even though these associations may have no or negative correlation, such associations get boosted many association mining algorithms.

Shopping carts can be mix of products of varied functionalities

- Customers sometimes have very focused shopping sessions
- While, at other times it can be mix of different functionalities
  - Like grocery with electronics
- Such transactions are misleading and should be discarded unless positive correlation is observed

Background -
Association rule generation

Generating associations

- Brute force approach
  - Generate all associations
  - Compute the support for every association and filter by minimum threshold
- This approach is very compute expensive
- Note, the support of rule X—>Y depends on the support of the corresponding items.
- So, we filter and consider only those items which have minimum support

Apriori algorithm

- Finds frequent item sets and rules
- Uses the anti-monotone property
  - Support of a rule never exceeds the support of it's item set.
- Support based pruning to eliminate less frequent item set
- Confidence based pruning to generate new rules
  - e.g.: {acd} —> {b} and {abd} —> {c} have high confidence
  - Then we get {ad} —> {bc}

Limitations of Apriori algorithm

- Setting support and confidence requires domain experience
- Support filter can eliminate interesting associations with low frequency
- Confidence is not a measure of correlation, hence it can be misleading.

Generating rules for retail industry

- Retail industry follows seasonal trends
- Not all product associations are important at all times
- Apriori generates rules on the whole of data which is expensive
- Association rules can be generated for products over a shorter window of interest
- Cost of rule generation over a window is much smaller

We have rules …

Now
let’s evaluate them

Basic measure of strength

- Recall, association rule $x- > y$
- **Support**
  - Measures frequency of rule $\frac{n(x,y)}{N}$
- **Confidence**
  - Measures strength of the rule
  - Conditional probability $P(y/x)$

Two Way Contingency Matrix

	Coffee	~Coffee	
Tea	150	50	200
~Tea	650	150	800
	800	200	1000

Calculating

$$Support = \frac{n(Tea, Coffee)}{N}$$

$$Confidence = \frac{support(Tea, Coffee)}{support(Tea)}$$

	Coffee	~Coffee	
Tea	150	50	200
~Tea	650	150	800
	800	200	1000

$$Support = \frac{150}{1000} = 0.15$$

$$Confidence = \frac{150}{200} = 0.75$$

- Support is used for pruning less frequency associations.
- Confidence is used for pruning weaker associations

Limitations from confidence value

- $x->y$ looks like a good rule with high support and confidence

	Coffee	~Coffe	
Tea	150	50	200
~Tea	650	150	800
	800	200	1000

$$P(Coffee/Tea) = \frac{150}{200} = 0.75$$

$$P(Coffee) = \frac{800}{1000} = 0.8$$

- Probability of drinking coffee decreases if the person drinks tea.

Lift / Interest Factor

$$Lift = \frac{support(Tea, Coffee)}{support(Tea) * support(Coffee)}$$

$$Lift = \frac{likelihood\ of\ rule}{likelihood\ of\ individual\ probabilities}$$

$$Lift = \frac{0.15}{0.2 * 0.8} = 0.94$$

- Tea and Coffee are Negatively correlated

Score	Correlation
Lift > 1	Positive
Lift == 1	zero
Lift < 1	Negative

Point wise Mutual Information

$$PMI = log2\frac{support(x,y)}{support(x) * support(y)}$$

$$PMI = log2\frac{0.15}{0.2 * 0.8} = -0.09$$

	Coffee	~Coffe	
Tea	150	50	200
~Tea	650	150	800
	800	200	1000

- Similar to Lift
- Log takes care of long decimal tail

Limitations of Lift & PMI

- Recall

$$Lift = \frac{support(x, y)}{support(x) * support(y)}$$

$$PMI = log2\frac{support(x, y)}{support(x) * support(y)}$$

- Scores for low support events get boosted up
  - Causing spurious associations to bubble up

IS Measure

	B	~B	
A	880	50	930
~A	50	20	70
	930	70	1000

	Y	~Y	
X	20	50	70
~X	50	880	930
	70	930	1000

$$IS = \frac{support(x, y)}{\sqrt{support(x) * support(y)}}$$

	Sup	Conf	Lift	PMI	IS
A&B	0.88	0.95	1.02	0.025	0.94
X&Y	0.02	0.29	4.08	2.029	0.2

Normalised PMI

	B	~B	
A	880	50	930
~A	50	20	70
	930	70	1000

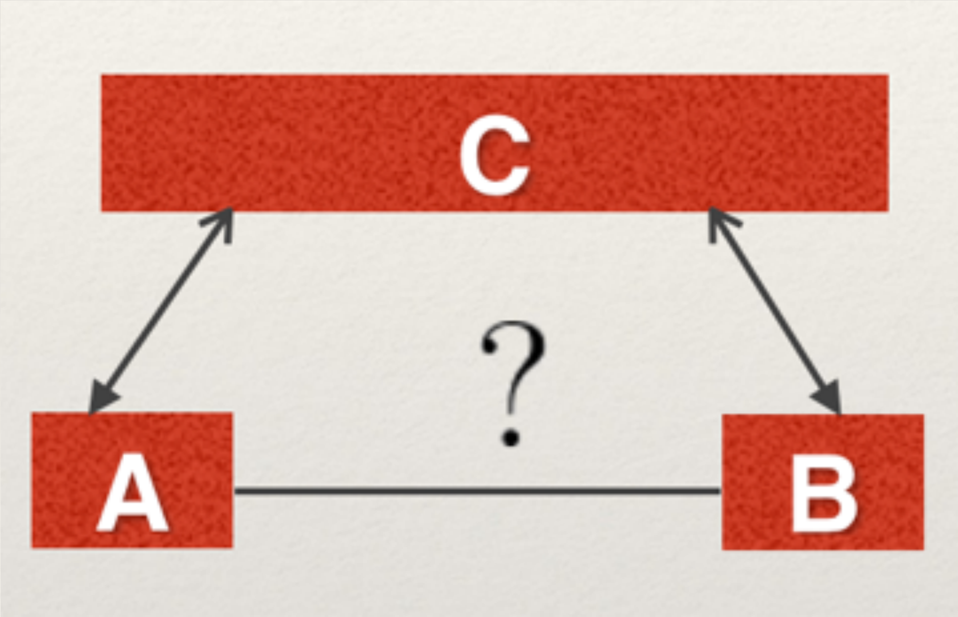
	Y	~Y	
X	20	50	70
~X	50	880	930
	70	930	1000

$$NPMI = \frac{PMI}{log2(support(x, y))}$$

	Sup	Conf	Lift	PMI	IS	NPMI
A&B	0.88	0.95	1.02	0.025	0.94	-0.14
X&Y	0.02	0.29	4.08	2.029	0.2	-0.36

Transitive/Indirect rule mining

- Data sparsity leads to a lot of important associations left undiscovered
- Can we mine rare/undiscovered associations ?



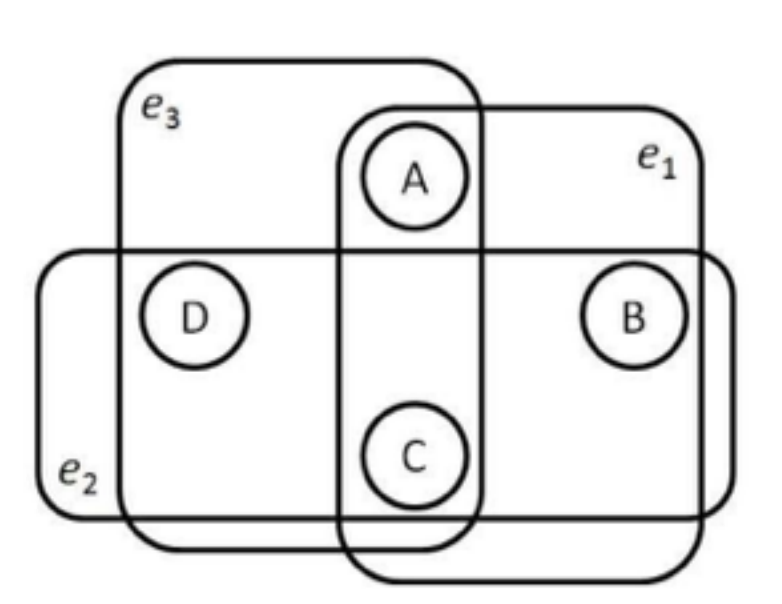
Indirect Association Mining

- Proposed by Tan et al.
- Non existant or rare pair {M,X}
- High dependence on mediator M
- 0 => independence
- 1 => complete dependence
- (0,1] => positively correlated

$$\mu = \frac{P(M,X) - P(M)P(X)}{P(M,X) - (1 - P(X))}$$

Semantic Association Mining

- Hyper Graph by Liu et al.
- Hyper edge: edge connecting to any number of vertices
- Two items are semantically associated if similarity measure > threshold
- Find all similar k - item sets
- Rank similar k - item sets



(c)

Key takeaway from evaluating interestingness measures

- No one measure that works for all
- Main problem of retail data
  - Data sparsity / high undiscovered associations
  - Spurious associations
  - Mixed purchase intend
- What works best is the combination of measures

Thank You!