Strong Association Rule

- An association rule is strong if it satisfies both a minimum support threshold (`min_sup`) and a minimum confidence threshold (`min_conf`)

Strong Association Rules Could Be Misleading …

- Example:
  - 10,000 transactions
  - 6,000 transactions included computer games
  - 7,500 transactions included videos
  - 4,000 transactions included both
- buys (X , ”Game”) □ buys (X , ”Video”)
  - Support?? Confidence??

Strong Association Rules Could Be Misleading …

- buys (X , ”Game”) □ buys (X , ”Video”)
  - Support (Game & Video) =
    4,000 / 10,000 =40%
  - Confidence (Game => Video) =
    4,000 / 6,000 = 66%
  - Suppose it pass our minimum support and confidence (30% , 60%, respectively)

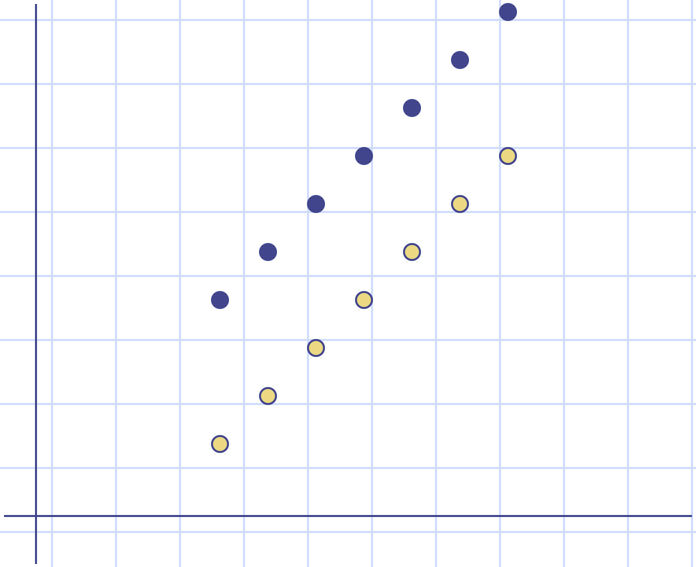
… Strong Association Rules Could Be Misleading

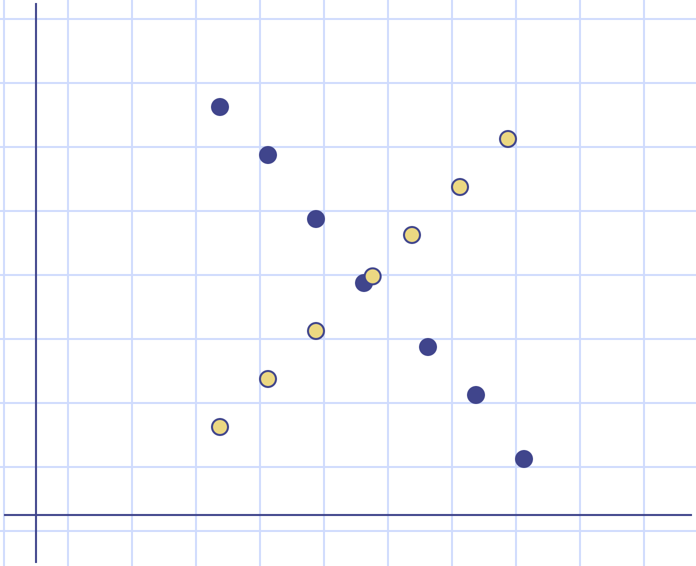
- Does buying game really imply buying video as well??
- Misleading…
  - Because, Purchasing videos is 75%
    - Which is quite greater than 66%

Misleading "Strong" Association Rule

- The truth is : "Computer games and videos are <u>negatively</u> associated"
  - i.e. the purchase of one of these items actually decreases the likelihood of purchasing the other
  - How to get this conclusion??

Correlation





Correlation

- 10,000 transactions
- 6,000 transactions included computer games (G)
- 7,500 transactions included videos (V)
- 4,000 transactions included both (G,V)

ƒP(S ˆ B) = 4000/10000 = 0.40

ƒP(S) × P(B) = 0.60 × 0.75 = 0.45

ƒP(S ˆ B) = P(S) × P(B) => Statistical independence

ƒP(S ˆ B) > P(S) × P(B) => Positively correlated

ƒP(S ˆ B) < P(S) × P(B) => Negatively correlated

Correlation Measures for Association Rules

- Lift
- Cosine
- All_confidence
- χ^2

Lift

- Simple correlation measure
- Considers
  - The occurrence of itemset A is independent of the occurrence of itemset B if $P(A \cup B) = P(A)P(B)$
  - Otherwise, itemset A and B are **dependent** and **correlated** as events
  - i.e. both the antecedent and consequents of the rules are considered here

- $\text{Lift}(A,B) = P(AUB) / P(A)P(B)$

Lift

- **Lift(A,B) = P(AuB) / P(A)P(B)**
  - If the value = 1
    - Then A and B are **independent** and there is no correlation between them
  - If the value > 1
    - Then A and B are **positively correlated**
    - The occurrence of one implies the occurrence of the other
  - If the value < 1
    - The occurrence of A is **negatively correlated** with the occurrence of B

Lift

w.r.t. Association Rules,

If the lift is 1, imply that the probability of occurrence of the antecedent and that of the consequent are independent of each other. When two events are independent of each other, no rule can be drawn involving those two events.

If the lift is > 1, then it will inform the degree to which those two occurrences are dependent on one another, and makes those rules potentially useful for predicting the consequent in future data sets.

The value of lift is that it considers both the confidence of the rule and the overall data set.

Lift

$$lift(A,B) = \frac{P(A \cup B)}{P(A)P(B)}$$

Example:

10,000 transactions

6,000 transactions included computer games

7,500 transactions included videos

4,000 transactions included both

- `lift({game},{video})=??`

  `= 0.40/(0.60 * 0.75)`

  `= 0.89  < 1 -vely correlated`

So buying a computer game actually *reduces* the chance of buying a video!

A useful rule must have lift > 1

Cosine Measure

- Harmonized lift measure
    - Two formulae are similar except that the for cosine, the square root is taken on the product of the probabilities of A and B

$$\text{cosine}(A,B) = \frac{P(A \cup B)}{\sqrt{P(A) \times P(B)}}$$

- `cosine({game},{video})=??`

Cosine vs. Lift

$$lift(A,B)=\frac{P(A\cup B)}{P(A)P(B)}=\frac{\frac{\sup(A\cup B)}{N}}{\frac{\sup(A)}{N}\,\frac{\sup(B)}{N}}=\frac{N\sup(A\cup B)}{\sup(A)\sup(B)}$$

$$\text{cosine}(A,B)=\frac{P(A\cup B)}{\sqrt{P(A)\times P(B)}}=\frac{\frac{\sup(A\cup B)}{N}}{\sqrt{\frac{\sup(A)\sup(B)}{N^2}}}=\frac{\sup(A\cup B)}{\sqrt{\sup(A)\sup(B)}}$$

Cosine value is only
influenced by the supports of A, B and AUB,
and NOT by the total number of transactions (N)

All_confidence

- Variation of the confidence
- All rules which can be generated from itemset **X** have at least a confidence of **all-confidence(X)**

$$all\_conf(\mathbf{X}) = \frac{\sup(\mathbf{X})}{\max\{\sup(i_j) \mid \forall i_j \in \mathbf{X}\}}$$

  - Denominator is the maximum count of transactions that contain any subset of X

- `all_conf({game, video})=??`

All_confidence

- All-confidence possesses the downward-closed closure property
    - It is the smallest confidence of any rule for the set of items, X
        - That is, all rules produced from this item set would have a confidence greater than or equal to its all-confidence value

Misleading “Strong” Association Rule

- Under the normal situation,
  - 60% of customers buy the game
  - 75% of customers buy the video
  - Therefore, it should have 60% * 75% = 45% of people buy both
  - That equals to 4,500 which is more than 4,000 (the actual value)

χ^2 Test

- Test whether A and B are *independent*
- If χ^2
  - 0: A and B are independent
  - >1 then
    - if observed value < expected value
      - Negatively Correlated

Pearson χ^2 Statistic

- Two attributes `A` and `B`
  - `A` has `r` possible values
  - `B` has `c` possible values
- Event $(A=a_i, B=b_j)$
  - Observed frequency: o_{ij}
  - Expected frequency: $e_{ij} = count(A=a_i) * count(B=b_j) / N$

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(o_{ij} - e_{ij})^2}{e_{ij}}$$

Observed Frequencies (Contingency Table)

Example: 10,000 transactions
6,000 transactions included computer games
7,500 transactions included videos
4,000 transactions included both

	game	!game	**total**
video	4000	??	**7500**
!video	??	??	**??**
total	**6000**	**??**	**10000**

Observed Frequencies (Contingency Table)

	game	!game	**total**
video	4000	3500	**7500**
!video	2000	500	**2500**
total	**6000**	**4000**	**10000**

Expected Frequencies

	game	!game	**total**
video	??	??	**7500**
!video	??	??	**2500**
total	**6000**	**4000**	**10000**

Expected frequency:

$\mathbf{e}_{\mathbf{ij}}$=count(A=a_i)*count(B=b_j)/N

Observed (Expected) Frequencies (Contingency Table)

	game	!game	**total**
video	4000 (4500)	3500	**7500**
!video	2000	500	**2500**
total	**6000**	**4000**	**10000**

Expected Frequency=(Count(game)*Count(video))/N

=(6000*7500)/10000=4500

Observed (Expected) Frequencies (Contingency Table)

	game	!game	**total**
video	4000 (4500)	3500 (3000)	**7500**
!video	2000 (1500)	500 (1000)	**2500**
total	**6000**	**4000**	**10000**

Expected Frequency=(Count(!game)*Count(video))/N

=(4000*7500)/10000=3000

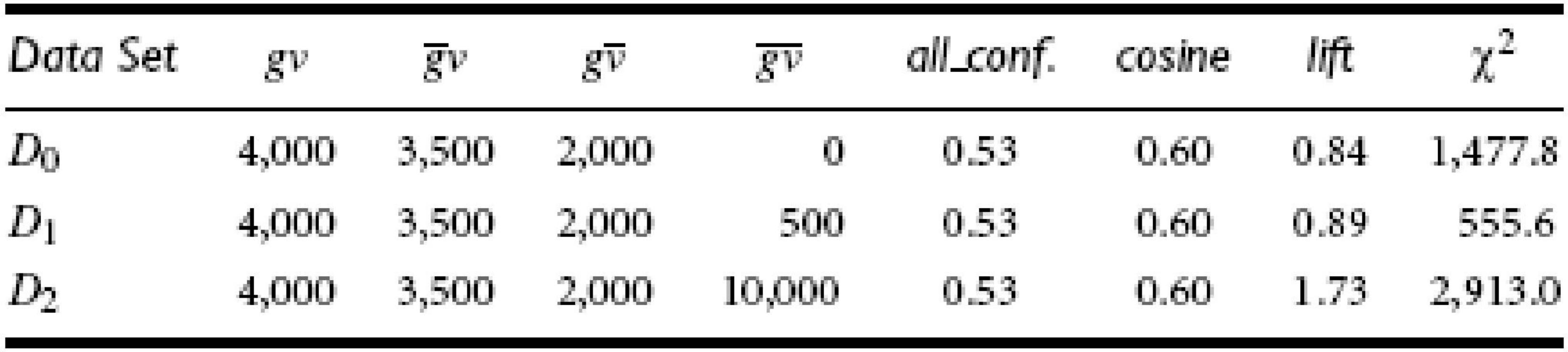
Interpretation of χ^2

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(o_{ij}-e_{ij})^2}{e_{ij}}$$

$$\chi^2 = \frac{(4000-4500)^2}{4500}+\frac{(3500-3000)^2}{3000}+\frac{(2000-1500)^2}{1500}+\frac{(500-1000)^2}{1000} = 555.56$$

- χ^2 is greater than 1
- And the observed value of the slot (game, video) = 4000 which is less than the expected value 4500
- Buying game and buying video are Negatively correlated

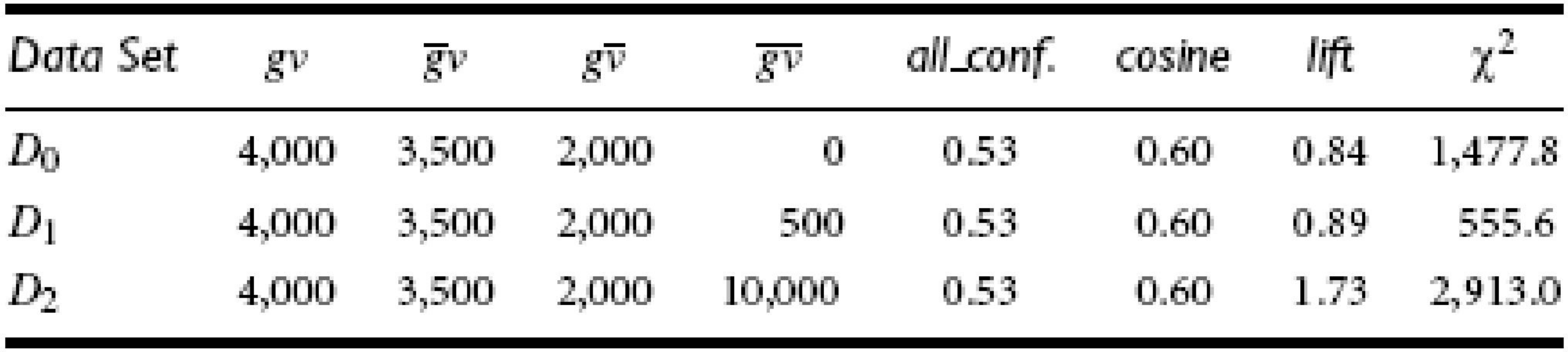
Comparison



Data Set	gv	$\overline{g}v$	$g\overline{v}$	$\overline{gv}$	*all_conf.*	*cosine*	*lift*	χ^2
D_0	4,000	3,500	2,000	0	0.53	0.60	0.84	1,477.8
D_1	4,000	3,500	2,000	500	0.53	0.60	0.89	555.6
D_2	4,000	3,500	2,000	10,000	0.53	0.60	1.73	2,913.0

- Let *D*1 be the original game (*g*) and video (*v*) data set
- Added two more data sets
  - *D*0 has zero null-transactions
  - *D*2 has 10,000 null-transactions (instead of only 500 as in *D*1)
- *gv*, *g'v*, and *gv'* remain the same in *D*0, *D*1, and *D*2

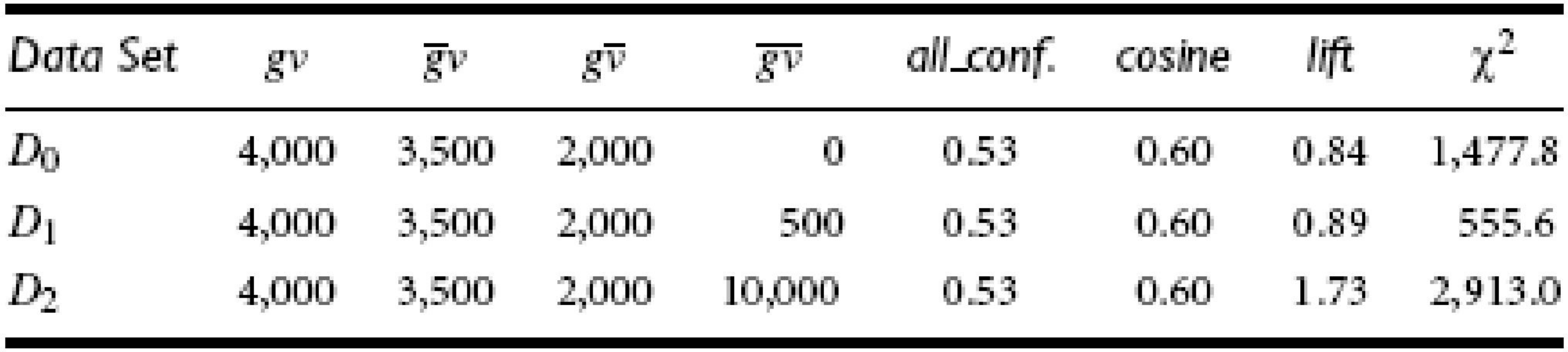
Comparison



Data Set	gv	$\bar{g}v$	$g\bar{v}$	$\overline{gv}$	*all_conf.*	*cosine*	*lift*	χ^2
D_0	4,000	3,500	2,000	0	0.53	0.60	0.84	1,477.8
D_1	4,000	3,500	2,000	500	0.53	0.60	0.89	555.6
D_2	4,000	3,500	2,000	10,000	0.53	0.60	1.73	2,913.0

- lift and χ^2 change from rather negative to rather positive correlations, whereas *all confidence* and *cosine* have the nice null-invariant property, and their values remain the same in all cases
  - Because lift and χ^2 strongly influenced by g'v' (Null Transactions)

Comparison



Data Set	gv	$\bar{g}v$	$g\bar{v}$	$\overline{gv}$	all_conf.	cosine	lift	χ^2
D_0	4,000	3,500	2,000	0	0.53	0.60	0.84	1,477.8
D_1	4,000	3,500	2,000	500	0.53	0.60	0.89	555.6
D_2	4,000	3,500	2,000	10,000	0.53	0.60	1.73	2,913.0

- Unfortunately, one cannot precisely assert that a set of items are positively or negatively correlated when the value of *all confidence* or *cosine* is around 0.5
- Strictly based on whether the value is greater than 0.5,
  - One can claim that *g* and *v* are positively correlated in *D*1
  - But, they are negatively correlated by the lift and χ^2 analysis
- Therefore, a good strategy is to perform the *all confidence* or *cosine* analysis first, and when the result shows that they are *weakly* positively/negatively correlated, other analyses can be performed to assist in obtaining a more complete picture