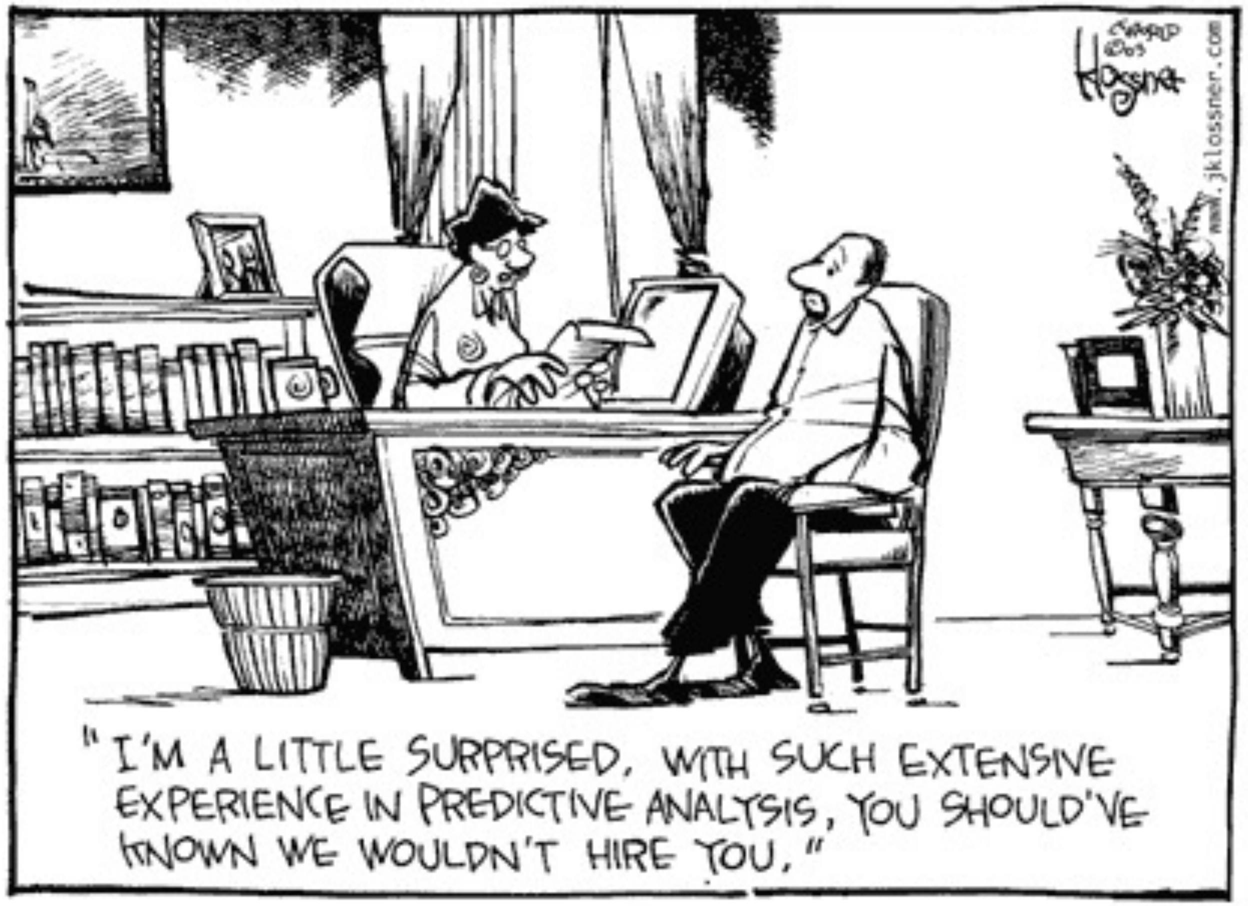

www.jklossner.com
"I'M A LITTLE SURPRISED, WITH SUCH EXTENSIVE EXPERIENCE IN PREDICTIVE ANALYSIS, YOU SHOULD'VE KNOWN WE WOULDN'T HIRE YOU."

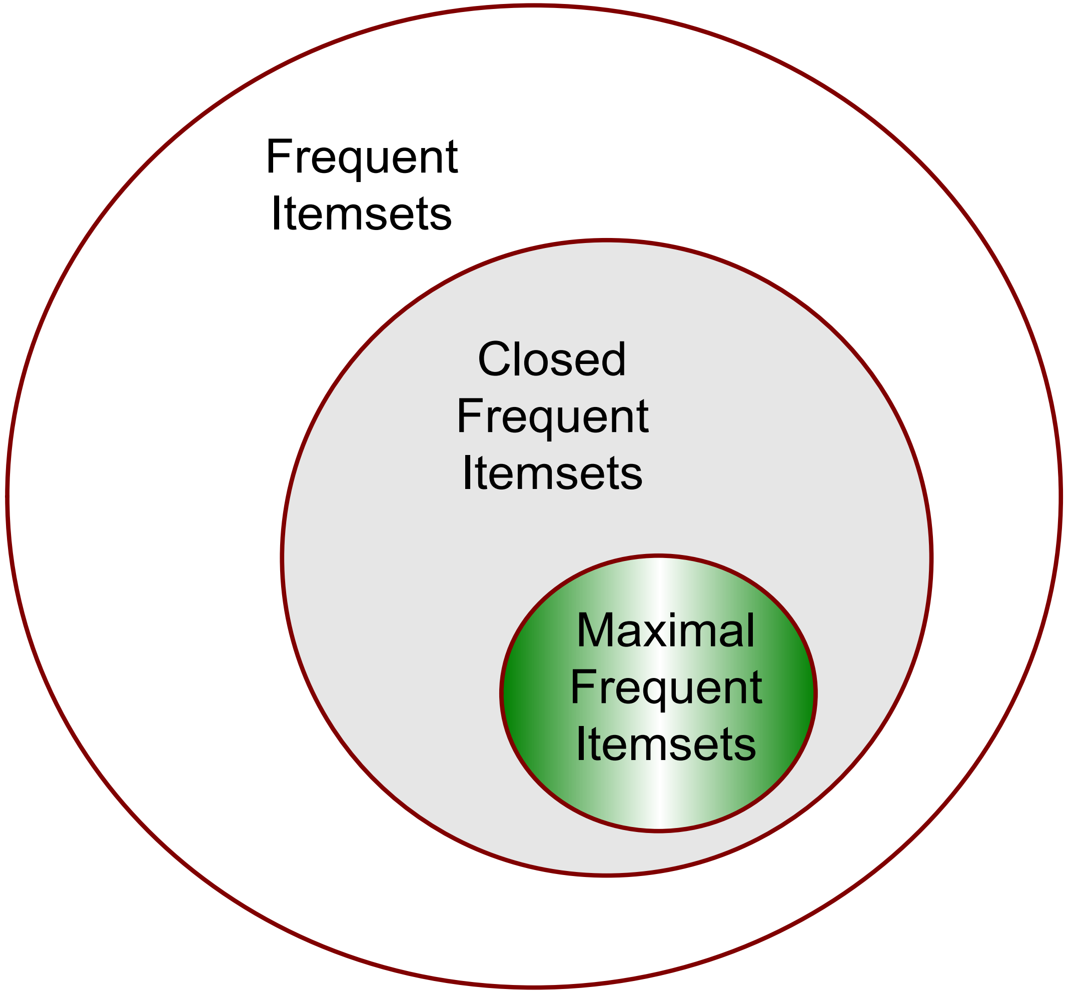

ModernAnalyst.com
COUGH! COUGH! YEARS OF DATA MINING HAVE LEFT ME WITH DATA LUNG. DON'T BE LIKE YOUR OLD MAN – GO INTO MODELING OR VISUALIZATION!

The Kepler-90 solar system. Wendy Stenzel/NASA

“Just as we expected, there are exciting discoveries lurking in our archived Kepler data, waiting for the right tool or technology to unearth them,” said Paul Hertz, director of NASA’s Astrophysics Division in Washington. “This finding shows that our data will be a treasure trove available to innovative researchers for years to come.”

https://www.nasa.gov/press-release/artificial-intelligence-nasa-data-used-to-discover-eighth-planet-circling-distant-star

Recall: Maximal vs Closed Itemsets



Evaluating
Association Analysis

Support and Confidence

- **Support**: Fraction of transactions that contain both X and Y

$$s(X \rightarrow Y) = \frac{\sigma(X \cup Y)}{N} = \mathrm{P}(\mathrm{X},\mathrm{Y})$$

- **Confidence**: Measures how often items in Y appear in transactions that contain X

$$c(X \rightarrow Y) = \frac{\sigma(X \cup Y)}{\sigma(X)} = \frac{\sigma(X \cup Y)/N}{\sigma(X)/N} = \frac{P(X,Y)}{P(X)} = \mathrm{P}(\mathrm{Y}|\mathrm{X})$$

Limitations of Support and Confidence

- There are times when both support and confidence are high, but the rule produced is not good

Ex:

Orange Juice → Milk, 30% support, 75% confidence

Milk, 90% support

Lift

Lift: measures the ratio of the observed frequency of co-occurrence to the expected frequency (also called **surprise**, or **interest**)

lift(X → Y) = $\frac{c(X \rightarrow Y)}{s(Y)} = \frac{s(XY)}{s(X)\, s(Y)} = \frac{P(X,Y)}{P(X)P(Y)}$

- If the two itemsets are statistically independent, then P(X,Y) = P(X)P(Y), corresponding to lift = 1.

Ex: Orange Juice → Milk, 30% support, 75% confidence

Milk, 90% support

Orange Juice, 40% support

Lift(OJ → Milk) = $\frac{0.75}{0.9} = \frac{0.3}{(0.4)(0.9)} = 0.83$

Lift < 1 indicates a negative correlation!

Comparing Support, Confidence, and Lift

TID	Items
1	A B D E
2	B C E
3	A B D E
4	A B C E
5	A B C D E
6	B C D

Rule	sup	conf
E → AC	0.33	0.40
E → AB	0.67	0.80
B → E	0.83	0.83

Contingency Tables

X → Y

	Y	$\overline{Y}$	
X	f_{11}	f_{10}	f_{1+}
$\overline{X}$	f_{01}	f_{00}	f_{0+}
	f_{+1}	f_{+0}	N

Tea → Coffee

	Coffee	$\overline{\textit{Coffee}}$	
Tea	*150*	*50*	*200*
$\overline{\textit{Tea}}$	*650*	*150*	*800*
	800	*200*	*1000*

lift(X → Y) = $\frac{c(X \rightarrow Y)}{s(Y)} = \frac{s(XY)}{s(X)\, s(Y)} = \frac{P(X,Y)}{P(X)P(Y)} = \frac{f_{11}/N}{(f_{1+}/N)(f_{+1}/N)} = \frac{N\, f_{11}}{(f_{1+})(f_{+1})}$

lift(Tea → Coffee) = $\frac{N\, f_{11}}{(f_{1+})(f_{+1})} = \frac{(1000)(150)}{(200)(800)} = 0.94$

Other Interestingness Measures

Table 6.11. Examples of symmetric objective measures for the itemset $\{A, B\}$.

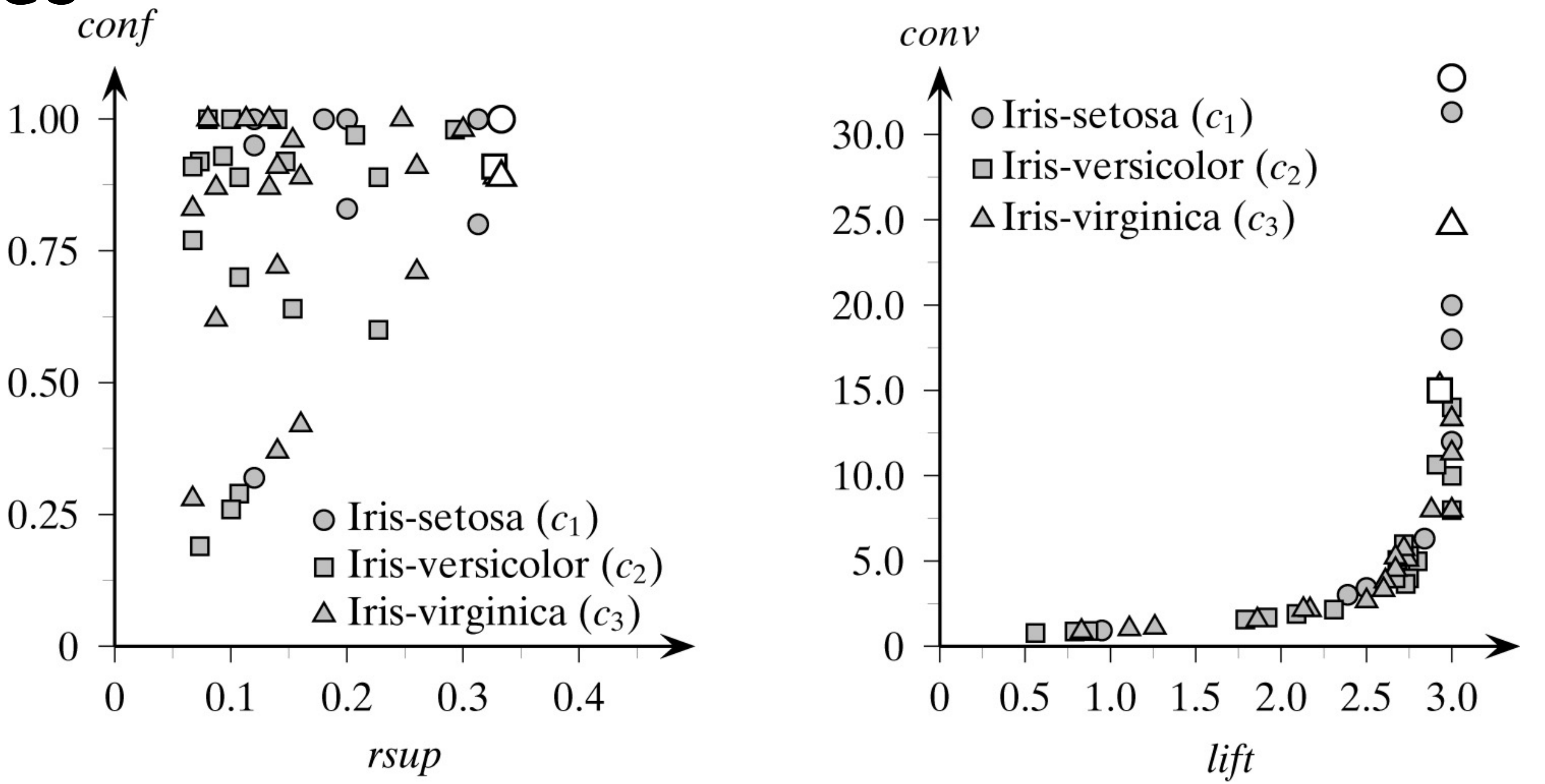
Measure (Symbol)	Definition
Correlation (ϕ)	$\frac{Nf_{11}-f_{1+}f_{+1}}{\sqrt{f_{1+}f_{+1}f_{0+}f_{+0}}}$
Odds ratio (α)	$(f_{11}f_{00})/(f_{10}f_{01})$
Kappa (κ)	$\frac{Nf_{11}+Nf_{00}-f_{1+}f_{+1}-f_{0+}f_{+0}}{N^2-f_{1+}f_{+1}-f_{0+}f_{+0}}$
Interest (I)	$(Nf_{11})/(f_{1+}f_{+1})$
Cosine (IS)	$(f_{11})/(\sqrt{f_{1+}f_{+1}})$
Piatetsky-Shapiro (PS)	$\frac{f_{11}}{N}-\frac{f_{1+}f_{+1}}{N^2}$
Collective strength (S)	$\frac{f_{11}+f_{00}}{f_{1+}f_{+1}+f_{0+}f_{+0}} \times \frac{N-f_{1+}f_{+1}-f_{0+}f_{+0}}{N-f_{11}-f_{00}}$
Jaccard (ζ)	$f_{11}/(f_{1+}+f_{+1}-f_{11})$
All-confidence (h)	$\min\left[\frac{f_{11}}{f_{1+}}, \frac{f_{11}}{f_{+1}}\right]$

Table 6.12. Examples of asymmetric objective measures for the rule $A \longrightarrow B$.

Measure (Symbol)	Definition
Goodman-Kruskal (λ)	$(\sum_j \max_k f_{jk} - max_k f_{+k})/(N - \max_k f_{+k})$
Mutual Information (M)	$(\sum_i \sum_j \frac{f_{ij}}{N} \log \frac{Nf_{ij}}{f_{i+}f_{+j}})/(-\sum_i \frac{f_{i+}}{N} \log \frac{f_{i+}}{N})$
J-Measure (J)	$\frac{f_{11}}{N} \log \frac{Nf_{11}}{f_{1+}f_{+1}} + \frac{f_{10}}{N} \log \frac{Nf_{10}}{f_{1+}f_{+0}}$
Gini index (G)	$\frac{f_{1+}}{N} \times (\frac{f_{11}}{f_{1+}})^2 + (\frac{f_{10}}{f_{1+}})^2] - (\frac{f_{+1}}{N})^2$ $+ \frac{f_{0+}}{N} \times [(\frac{f_{01}}{f_{0+}})^2 + (\frac{f_{00}}{f_{0+}})^2] - (\frac{f_{+0}}{N})^2$
Laplace (L)	$(f_{11}+1)/(f_{1+}+2)$
Conviction (V)	$(f_{1+}f_{+0})/(Nf_{10})$
Certainty factor (F)	$(\frac{f_{11}}{f_{1+}} - \frac{f_{+1}}{N})/(1-\frac{f_{+1}}{N})$
Added Value (AV)	$\frac{f_{11}}{f_{1+}} - \frac{f_{+1}}{N}$

Comparing Rules

Attribute	Range or value	Label
Sepal length	4.30–5.55 5.55–6.15 6.15–7.90	sl_1 sl_2 sl_3
Sepal width	2.00–2.95 2.95–3.35 3.35–4.40	sw_1 sw_2 sw_3
Petal length	1.00–2.45 2.45–4.75 4.75–6.90	pl_1 pl_2 pl_3
Petal width	0.10–0.80 0.80–1.75 1.75–2.50	pw_1 pw_2 pw_3
Class	Iris-setosa Iris-versicolor Iris-virginica	c_1 c_2 c_3



(a) Support vs. confidence

(b) Lift vs. conviction

Best Rules by Support and Confidence

Rule	*rsup*	*conf*	*lift*	*conv*
$\{pl_1, pw_1\} \longrightarrow c_1$	0.333	1.00	3.00	33.33
$pw_2 \longrightarrow c_2$	0.327	0.91	2.72	6.00
$pl_3 \longrightarrow c_3$	0.327	0.89	2.67	5.24

Best Rules by Lift and Conviction

Rule	*rsup*	*conf*	*lift*	*conv*
$\{pl_1, pw_1\} \longrightarrow c_1$	0.33	1.00	3.00	33.33
$\{pl_2, pw_2\} \longrightarrow c_2$	0.29	0.98	2.93	15.00
$\{sl_3, pl_3, pw_3\} \longrightarrow c_3$	0.25	1.00	3.00	24.67

Redundant Rules

- Given two rules that have the same consequent:

  R: X → Y and R': W → Y , such that $W \subset X$

  R: {Diapers, Milk} → {Beer} , R': {Diapers} → {Beer}

- We say that R is *more specific* than R' (or that R' is *more general* than R)
- We say that R is *redundant*, if there exists a more general rule that has the same support.
- If s(R) = s(R') then R is redundant
- If s(R) < s(R') over all generalizations R', then R is non-redundant

Productive Rules

- Given two rules that have the same consequent:

  R : X → Y and R' : W → Y , such that $W \subset X$

  R : {Diapers, Milk} → {Beer} , R' : {Diapers} → {Beer}

- Define the *improvement* of a rule as:

  $\text{imp}(X \rightarrow Y) = c(X \rightarrow Y) - \max_{W \subset X}\{c(W \rightarrow Y)\}$

- A rule is *productive* if its improvement is greater than 0.