**Additional Association Rules Topics**

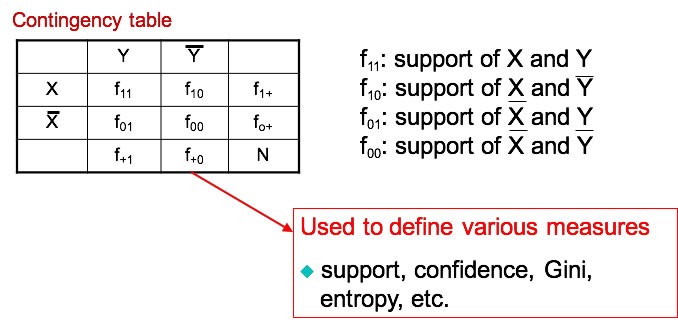
# Rule Pattern Evaluation

Association rule algorithms can produce large number of rules

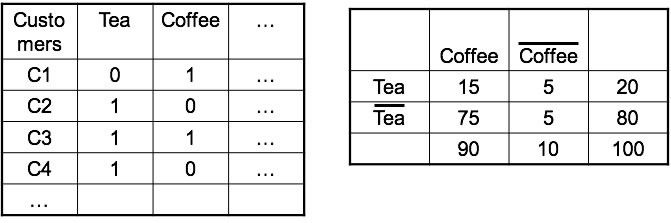
**Interestingness** measures can be used to prune/rank the patterns

In the original formulation, **support** and **confidence** are the only measures used

So to compute an interestingness measure, use a contingency table for X →Y or {X,Y}



## Example drawback on confidence



Association Rule: **Tea → Coffee**

Confidence **P(Coffee|Tea) = 15/20 = 0.75**

The threshold is met, confidence > 50%, meaning people who drink tea are more likely to drink coffee than not drink coffee

So rule seems reasonable.

but **P(Coffee) = 0.9**, which means knowing that a person drinks tea reduces the probability that the person drinks coffee!

Note that **P(Coffee | ~Tea) = 75/80 = 0.9375**

# Measures for Association Rules

So, what kind of rules do we really want?

Confidence(X**→** Y) should be sufficiently high

To ensure that people who buy X will more likely buy Y than not buy Y

Confidence(X **→** Y) > support(Y)

Otherwise, rule will be misleading because having item X actually reduces the chance of having item Y in the same transaction

There are many more measures that capture this constraint.

## Statistical Independence

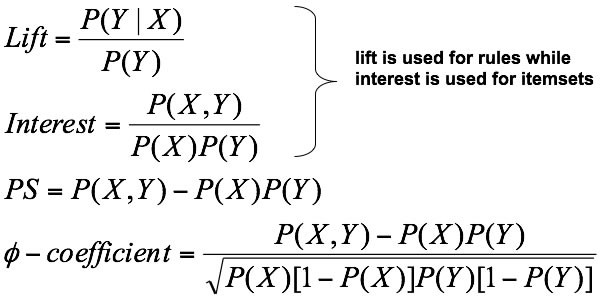
The criterion for statistical independence is that **confidence(X→ Y) = support(Y)** which is is equivalent to: P(Y|X) = P(Y) and P(X,Y) = P(X) **\*** P(Y)

If P(X,Y) > P(X) \* P(Y) : X & Y are positively correlated

If P(X,Y) < P(X)

**\***

P(Y) : X & Y are negatively correlated



All of these measures are reported by Weka and RapidMiner.

## Lift computations

Conf (X→Y) =σ(X,Y) / σ(Y)

|  |  |  |  |
| --- | --- | --- | --- |
| Lift (X→Y) = | P(Y|X)  -------- =  P(Y) | conf(X→Y)  ----------------- =  σ(X) | σ(X,Y)  -------------σ(X)\*σ(Y) |

# Lift/Interest example

Coffee

~Coffee

Tea

15

5

20

~Tea

75

5

80

90

10

100

Association Rule: Tea

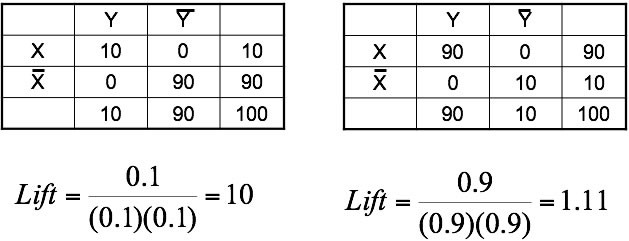
**→**

Coffee

Confidence= P(Coffee | Tea) = 15/20 = 0.75 but P(Coffee) = 0.9

Thus, Lift = **0.75/0.9= 0.8333** = 0.15 / (0.9\*0.2) = 0.8333

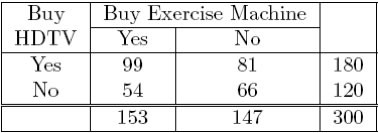
Lift here is < 1, and therefore negatively associated So, is it enough to use confidence and/or lift for pruning?



Set of measures

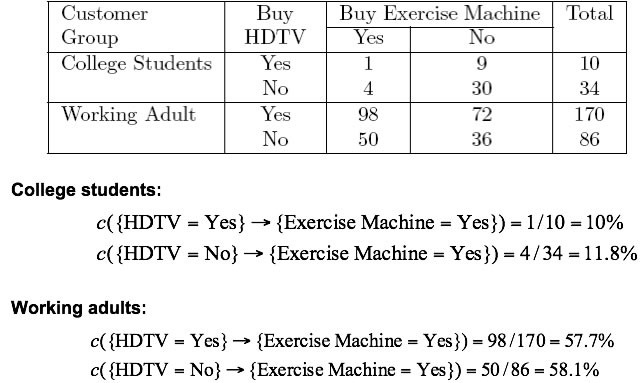


# Simpson Paradox



c({HDTV=Yes} -> {Exercise Machine=Yes} = 99/180 = 55% c({HDTV=No} -> {Exercise Machine=Yes} = 54/120 = 45%

Thus the conclusion is that customers who buy HDTV are more likely to buy exercise machines.



Observed relationship in data may be influenced by the presence of other confounding factors (hidden variables).

Hidden variables may cause the observed relationship to disappear or reverse its direction.

Proper stratification is needed to avoid generating spurious patterns.

## Mathematically

**a/b < c/d** and **p/q < r/s**, where a/b and p/q represent the confidence of the rule A-> B in two different strata, accounting for an additional variable and c/d and r/s represents the confidence of the rule ~A -> B in the same strata.

The paradox occurs when **(a+p)/(b+q) > (c+r)/(d+s),** which can lead to a flawed conclusion.

# Effect of Support Distribution on Association Mining

Many real data sets have skewed support distribution

