Statistics Analysis in Data Mining

By
Dr. Siddique Ibrahim
Assistant Professor
VIT-AP University
Amaravati

- In today’s informative world, data are produce enormously every now and then.
- It could be in scenarios like marketing, manufacturing, transactional, defence, telecom, actuarial, service and so forth.
- The process of extracting useful information from large sets of raw data is known as Data Mining.
- It uses descriptive and inferential statistical analysis for analyzing the extracted data. It would be helpful in business decision making processes.

Association

- It helps to find the relationship between two or more data variables.
- We can solve some of the questions like – “What is the association between the climate and sales of a cloth?”
- “How strong is the relationship between the investment and sales of a company?”

- Correlation analysis is used to find the association between the variables in data mining. Correlation methods are Pearson's product-moment correlation coefficient, Kendall and Spearman rank correlations, etc.

Statistical methods used in Data Mining

1. Sampling – It is a process of taking a small set of observations (sample) from a large population. It is a common tool used in any type of data analysis.

2. Correlation Analysis -It is used to study the closeness of the relationship between two or more variables i.e. the degree to which the variables are associated with each other.

3. Regression Analysis-It is a commendable statistical technique used in data mining. It helps to predict the value of future outcomes by using the past data.

- Linear Regression
- Multiple Regression
- Logistic Regression
- Poisson Regression

4. Graphical Analysis-data are presented in the form of graphs or diagrams

- Histogram Bar chart Pareto chart Scatter plot…

Correlation Analysis

- Correlation Analysis is just an extension of Association Rules.

- Sometimes the support and confidence parameters may still yield uninteresting patterns to the users.

Correlation Measure

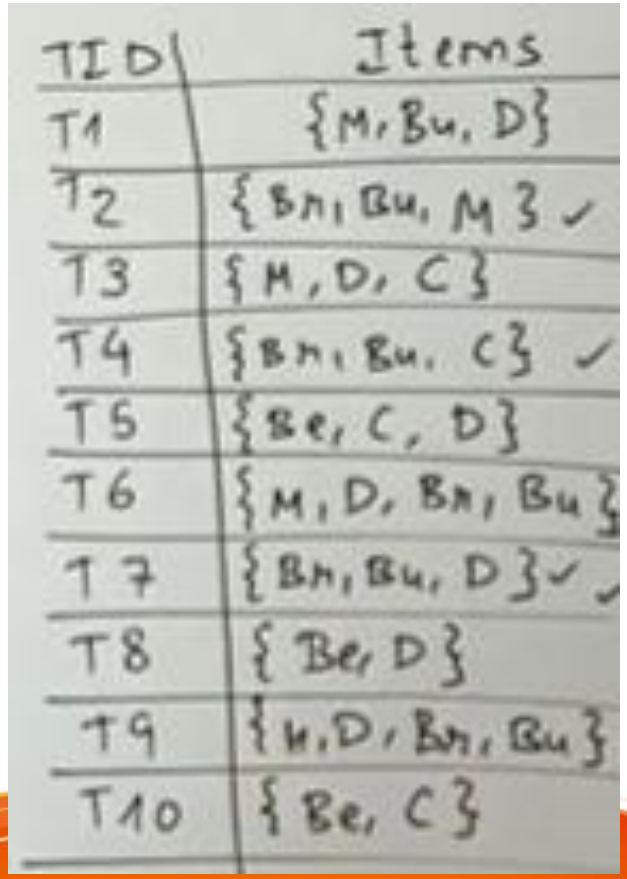
- Correlation rule is measured by support, confidence and correlation between itemsets A and B.
- Correlation is measured by Lift and Chi-Square.
- (i) Lift: As the word itself says, Lift represents the degree to which the presence of one itemset lifts the occurrence of other itemsets.

What is Lift?

- Lift is a measure of how much more likely the consequent of a rule is to occur when the antecedent is present, compared to when it is absent.

- A high lift means that the rule is significant and interesting, and that there is a strong association between the antecedent and the consequent.

- Identify misleading rules -> satisfy both min support and confidence



TID	Items
T1	{M, Bu, D}
T2	{Br, Bu, M} ✓
T3	{M, D, C}
T4	{Br, Bu, C} ✓
T5	{Be, C, D}
T6	{M, D, Br, Bu}
T7	{Br, Bu, D} ✓ ✓
T8	{Be, D}
T9	{M, D, Br, Bu}
T10	{Be, C}

An example

- Out of 1000 transactions analyzed,
- 600 contained only bread,
- while 750 contained butter and
- 400 contained both bread and butter.

- Suppose the min support for association rule run is 30% and the minimum confidence is 60%.

- The support value of 400/1000=40% and confidence value= 400/600= 66% meets the threshold.
- However, we see that the probability of purchasing butter is 75% which is more than 66%. This means that bread and butter are negatively correlated as the purchase of one would lead to a decrease in the purchase of the other. The results are deceiving.

- From the above example, the support and confidence are supplemented with another interestingness measure i.e. correlation analysis which will help in mining interesting patterns.

Topics of Discussion

- Lift (A, B) = P (A U B) / P (A). P (B).
- If it is < 1, then A and B are negatively correlated.
- If it is >1. Then A and B are positively correlated which means that the occurrence of one implies the occurrence of the other.
- If it is = 1, then there is no correlation between them.

Lift measure Formula

The **lift** of a rule X-->Y is calculated as lift(X-->Y) = ((sup(X U Y)/ N) / (sup(X)/ N*sup(Y)/ N), where

- N is the number of transactions in the transaction database,
- sup(XUY) is the number of transactions containing X and Y,
- sup(X) is the number of transactions containing X
- sup(Y) is the number of transactions containing Y.

- The formulas for these are
- confidence = support / antecedent support

- and

- Lift = confidence / consequent support.

- (ii) Chi-Square: This is another correlation measure. It measures the squared difference between the observed and expected value for a slot (A and B pair) divided by the expected value.

$$X^2 = \sum (\text{observed} - \text{expected})^2 / \text{expected}$$

- If it is >1, then it is negatively correlated.

- To illustrate,
- if you have a dataset of 100 transactions with
- 10 of them having bread, butter, and jam;
  20 of them having bread and butter; and
  15 of them having jam,

- then
- the support would be 0.1 (10 / 100), antecedent support would be 0.2 (20 / 100),
- consequent support would be 0.15 (15 / 100),
- confidence would be 0.5 (0.1 / 0.2), and lift would be 3.33 (0.5 / 0.15).

Obtain Association Rule

Transaction id	Items
t1	{1, 2, 4, 5}
t2	{2, 3, 5}
t3	{1, 2, 4, 5}
t4	{1, 2, 3, 5}
t5	{1, 2, 3, 4, 5}
t6	{2, 3, 4}

minsup = 0.5, minconf= 0.9 and minlift = 1

```
rule 0:    4   ==> 2
rule 1:    3   ==> 2
rule 2:    1   ==> 5
rule 3:    1   ==> 2
rule 4:    5   ==> 2
```

```
rule 5:   4 5  ==> 2
rule 6:   1 4  ==> 5
rule 7:   4 5  ==> 1
rule 8:   1 4  ==> 2
rule 9:   3 5  ==> 2
rule 10:  1 5  ==> 2
rule 11:  1 2  ==> 5
rule 12:  1  ==> 2 5
rule 13:  1 4 5  ==> 2
rule 14:  1 2 4  ==> 5
rule 15:  2 4 5  ==> 1
rule 16:  4 5  ==> 1 2
rule 17:  1 4  ==> 2 5
```

```
rule 0:   4  ==> 2      support :  0.66 (4/6) confidence :  1.0  lift :  1.0
rule 1:   3  ==> 2      support :  0.66 (4/6) confidence :  1.0  lift :  1.0
rule 2:   1  ==> 5      support :  0.66 (4/6) confidence :  1.0  lift :  1.2
rule 3:   1  ==> 2      support :  0.66 (4/6) confidence :  1.0  lift :  1.0
rule 4:   5  ==> 2      support :  0.833(5/6)  confidence :  1.0  lift :  1.0
rule 5:   4 5  ==> 2    support :  0.5 (3/6)   confidence :  1.0  lift :  1.0
rule 6:   1 4  ==> 5    support :  0.5 (3/6)   confidence :  1.0  lift :  1.2
rule 7:   4 5  ==> 1    support :  0.5 (3/6)   confidence :  1.0  lift :  1.5
rule 8:   1 4  ==> 2    support :  0.5 (3/6)   confidence :  1.0  lift :  1.0
rule 9:   3 5  ==> 2    support :  0.5 (3/6)   confidence :  1.0  lift :  1.0
rule 10:  1 5  ==> 2    support :  0.66 (4/6)  confidence :  1.0  lift :  1.0
rule 11:  1 2  ==> 5    support :  0.66 (4/6)  confidence :  1.0  lift :  1.2
rule 12:  1  ==> 2 5    support :  0.66 (4/6)  confidence :  1.0  lift :  1.2
rule 13:  1 4 5  ==> 2  support :  0.5 (3/6)   confidence :  1.0  lift :  1.0
rule 14:  1 2 4  ==> 5  support :  0.5 (3/6)   confidence :  1.0  lift :  1.2
rule 15:  2 4 5  ==> 1  support :  0.5 (3/6)   confidence :  1.0  lift :  1.5
rule 16:  4 5  ==> 1 2  support :  0.5 (3/6)   confidence :  1.0  lift :  1.5
rule 17:  1 4  ==> 2 5  support :  0.5 (3/6)   confidence :  1.0  lift :  1.5
```