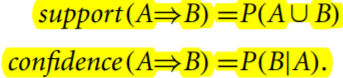
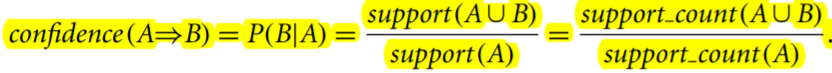
Chapter-6

Mining Frequent Patterns, Associations, and Correlations: Basic Concepts and Methods :

* 1. Frequent Patterns:
     1. patterns (eg itemsets, subsequences, or substructures) that appear frequently in a data set.
     2. Frequent Itemset: bought together
     3. Frequent sequential patterns: buy this then this then this.
     4. Frequent structured pattern: if a substructure like subgraph, sublattice, subtrees occurs frequently.
     5. FP mining helps in classification, clustering
  2. Basic Concepts:
     1. Searching for recurrence relation, associations, correlations.
     2. Market Basket Analysis: A Motivating Example:
        1. If a customer buys these items together then what can be the general itemsets that he can buy.
        2. Freq itemset mining→ leads to associations, correlations.
        3. **Real life applications:**
           1. Catalogue design.
           2. Cross-marketing
           3. Customer shopping behaviour analysis.
        4. Buying patterns using association rules
        5. where support says 2% of the transactions have these items are bought together, confidence says 60% of customers who bought this also bought this.
        6. Interesting association→ if rule supports min support && min conf threshold (set by users or domain experts)
     3. Frequent Itemsets, Closed Itemsets, and Association Rules:
        1. Terminologies:
           1. I = {I\_1, ...I\_m} itemset, T transaction set as D.
           2. Support(relative support) = % of transactions having both A,B in D.
           3. Confidence = % of transactions in D containing A that also contains B.
           4. rules that satisfy both → strong.
           5. k-itemset→ itemset having k items.
           6. Occurrence freq or Freq or count or support count or absolute support→ no. of transactions having those itemsets.
        2. Frequent Itemset:
           1. Itemset having min relative support threshold or min absolute support threshold



* + - 1. Problem statement of mining association rules reduces to finding out freq itemsets(bcz conf can be found out using support).
      2. ARM steps:
         1. Finding freq itemsets: itemsets having min support count.
         2. Generating strong association rules from freq itemsets: rules must have min support && min conf.
         3. Overall performance depends on 1st step.
      3. 1st step often generates multiple frequent itemsets(all subsets of freq itemsets are also freq)
         1. Eg. itemset having 100 items will have 100C1 + 100C2+...100C100 itemsets(C is for combinatorial selection). This is a huge no.
         2. therefore→ closed freq itemsets OR maximal freq itemsets.
      4. Terminology:
         1. Closed itemset:

If no superset of it has the same support count as it has.

* + - * 1. Closed freq itemset:

If it is closed + frequent.

* + - * 1. Maximal freq itemset:

If it has no superset which is freq and it itself is frequent.

* + - * 1. Eg. t1={a1….a100}, t2={a1….a50}

Closed itemsets are

{a1...a100}:1 && {a1,...a50}:2

Maximal itemset is

{a1….a100}.

* + - * 1. We will keep {a1,...a100}
  1. Frequent Itemset Mining Method:
     1. Apriori Algorithm: Finding Frequent Itemsets by Confined Candidate Generation:
        1. uses prior knowledge of frequent itemset properties.
        2. Steps goes like:
           1. 1-itemset is created and then it is filtered for the sets having min\_sup.
           2. To find out L\_k requires full scan of dB once.
        3. To improve the efficiency of the level-wise generation of frequent itemsets, an important property called the Apriori property is used to reduce the search space
        4. **Apriori property:**
           1. If the nonempty subsets of a set is not frequent then the set also cannot be frequent.
        5. 2-step:
           1. Join step
           2. Prune step(remove the sets whose if even a subset is not-frequent).
           3. Refer eg 6.3 form book.
     2. Generating Association Rules from Frequent Itemsets:
        1. Should satisfy both min sup & min conf.
        2. Steps:
           1. Make association rules from freq itemsets.
           2. Output them if they have min conf.
     3. Improving the Efficiency of Apriori:**[NOT COVERED]**
        1. Hash based technique
        2. Transaction reduction
        3. Partitioning
        4. Sampling
        5. Dynamic itemset counting
     4. A Pattern-Growth Approach for Mining Frequent Itemsets :
        1. Disadvantages of apriori:
           1. It may still need to generate a huge number of candidate sets.Eg if there are 10^4 frequent 1-itemsets, it will need to generate more than 10^7 candidate 2-itemsets.
           2. It may need to repeatedly scan the whole DB and check a large set of candidates by pattern matching. It is costly to go over each transaction in the db to get support of the candidate itemsets.
           3. Therefore, costly candidate generation process→ replaced with a method mining all sets of frequent itemset.
        2. FP-Growth:
           1. Uses DAC paradigm→ creates FP-Tree having all frequence itemsets.
           2. Refer example 6.5 .
           3. When DB is large it becomes unrealistic to construct FP Tree and keep in memory.
           4. ALternative to above problem is to partition the DB and construct FP Tree and mine the pattern recursively in each one.
     5. Mining Frequent Itemsets Using the Vertical Data Format:
        1. Previously→ transactions had items→ Tid vs Itemset
        2. In vertical data format→ there are transaction sets having items→ item vs transaction sets→ hence faster bcz it has all of data from the transactions → hence no need to scan through the DB.
        3. It needs less storage space, and can improve the efficiency of data mining.
     6. Mining Closed and Max Patterns:**[NOT COVERED]**
     7. Which Patterns Are Interesting?—Pattern Evaluation Methods :
        1. Using support-confidence framework→ avoids non-interesting patterns but still has some patterns which might not be interested to users.
        2. Additional interestingness→ correlation analysis & other pattern evaluation methods
        3. Strong Rules Are Not Necessarily Interesting:
           1. eg.buys(X,”computer games”) → buys(X,”videos”). Support = 40%, conf = 66%.
           2. Support-conf framework can be deceiving bcz it may be the case that 75% transactions contains “videos” and somewhat similarly for “computer games”.
           3. Hence, we go from association analysis to correlation analysis.
        4. From Association Analysis to Correlation Analysis:
           1. Correlation rule:

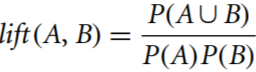
A ⇒ B [support, confidence, correlation].

* + - * 1. Correlation measures:

Lift:

Since P(A U B) here is prob of occurrence of A & B therefore.

P(A U B) = P(A)P(B) if A & B are independent.

 therefore

If lift = 1 means they are independent.

Lift < 1 means they are negatively correlated(means occurrence of one leads to absence of another)

Lift >1 means they are positively correlated(means occurance of one leads to another).

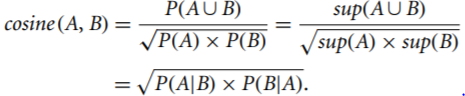
lift(A U B ) = conf(A→ B)/supp(B).

It says the degree to which one “lifts” the occurrence of other.

Chi^2:

DIscussed earlier to be calculated from contingency table.

If chi^2 > 1 & observed val < expected val→ negatively correlated.

* + - 1. A Comparison of Pattern Evaluation Measures:
         1. The above 2 measures lift and chi^2 were not null invariant. Hence other methods introduced.
         2. all confidence, max confidence, Kulczynski, and cosine.
         3. 
         4. 
         5. 
         6. 
         7. A null-transaction is a transaction that does not contain any of the itemsets being examined.
         8. To answer which of the 4 is better use **Imbalance Ratio(IR)**