Mining Frequent Patterns, Associations, and Correlations – Unit 3 Viva Notes

What Kinds of Patterns Can Be Mined

Descriptive patterns show general properties and relationships in data (e.g. frequent itemsets, association rules). Predictive patterns predict future outcomes based on current data (e.g. sales prediction).

Market Basket Analysis

A method to find frequent itemsets bought together. Based on 'If you buy A, you are likely to buy B'. Example: Car + Accessories → Insurance. Each group of items bought together is called an itemset.

Association Rule Mining

Finds relationships among items in transactions. Rule format: X → Y. Example: {Milk, Bread} → {Eggs}.

Key Terms

1. Itemset – group of items (e.g. {Milk, Bread})
2. k-itemset – itemset with k items
3. Support = count(X)/total transactions
4. Confidence = support(X∪Y)/support(X)

Process

1. Find frequent itemsets with support ≥ min support
2. Generate rules from them with confidence ≥ min confidence → strong rules

Maximal vs Closed Frequent Itemsets

Closed Frequent Itemset: No superset has the same support. Maximal Frequent Itemset: No superset is frequent. Closed ⊂ Maximal (Closed gives more info; Maximal is fewer).

Apriori Algorithm

Used to find frequent itemsets using level-wise search and Apriori property. Apriori property: If an itemset is frequent, all its subsets are also frequent.

1. Scan DB → find L1 (frequent 1-itemsets)
2. Join L1 with itself → make C2 (candidate 2-itemsets)
3. Prune those whose subsets are not frequent → get L2
4. Repeat for L3, L4... until no more frequent sets
5. Generate association rules from these frequent sets

Methods to Improve Apriori Efficiency

1. Hash-based technique – use hash table to reduce candidates
2. Transaction reduction – remove unhelpful transactions
3. Partitioning – divide data, find local frequent sets, merge
4. Sampling – use random small sample
5. Dynamic itemset counting – update counts as new data comes

FP-Growth Algorithm

Finds frequent patterns without candidate generation. Faster than Apriori.

1. Scan DB → get frequent items and counts

2. Sort items by frequency
3. Build FP-Tree by inserting transactions as paths
4. Mine FP-Tree using conditional pattern base and conditional FP-tree recursively

Pattern Evaluation Measures

Support and Confidence may produce many useless rules. Use interestingness measures to filter.

Correlation Analysis (Lift)

Checks if two items are actually related or just occur together by chance.

$\text{Lift} = P(A \cup B) / (P(A)*P(B))$

Lift = 1 → Independent, Lift > 1 → Positive correlation, Lift < 1 → Negative correlation

Example: Games=6000, Videos=7500, Both=4000 (out of 10000): Lift = 0.4 / (0.6*0.75)=0.89 < 1 → Negatively correlated.