BITS Pilani

BITS Pilani
Hyderabad Campus

Dr.Aruna Malapati
Asst Professor
Department of CSIS

innovate achieve lead

BITS Pilani
Hyderabad Campus

Association Rule Mining

Today's Learning objective

- Define the problem of Association rule Mining
- Applications of Association rule Mining
- Define and apply support count ,support and confidence

Association Rule Mining

- Given a set of transactions, find combinations of items (itemsets) that occur frequently

Market-Basket transactions

TID	*Items*
1	Bread, Milk
2	Bread, Diaper, Beer, Eggs
3	Milk, Diaper, Beer, Coke
4	Bread, Milk, Diaper, Beer
5	Bread, Milk, Diaper, Coke

Example of Association Rules

{Diaper} → {Beer},
{Milk, Bread} → {Eggs,Coke},
{Beer, Bread} → {Milk},

Implication means co-occurrence, not causality!

Applications – (1)

- Items = products; baskets = sets of products someone bought in one trip to the store.
- Example application: given that many people buy beer and diapers together:
  - Run a sale on diapers; raise price of beer.
- Only useful if many buy diapers & beer.

Applications – (2)

- Baskets = Web pages; items = words.

- Example application: Unusual words appearing together in a large number of documents, e.g., “Brad” and “Angelina,” may indicate an interesting relationship.

The Market Basket Model

- A (large) set of binary attributes, called items: $I=\{i_1,i_2,..i_n\}$.
  - E.g List of all items sold by a store
- A transaction T consists of a (small) subset of I.
  - , e.g., the list of items (bill) bought by one customer at once.
- The database D is a (large) set of transactions:

$D=\{T_1,T_2,.....T_n\}$

Definition: Frequent Itemset

Itemset

- A collection of one or more items
  - Example: {Milk, Bread, Diaper}
- k-itemset
  - An itemset that contains k items

TID	*Items*
1	Bread, Milk
2	Bread, Diaper, Beer, Eggs
3	Milk, Diaper, Beer, Coke
4	Bread, Milk, Diaper, Beer
5	Bread, Milk, Diaper, Coke

Support count (σ)

- No of transactions containing the itemset
- E.g. σ({Milk, Bread,Diaper}) = 2

An Association rule is interesting only if its support count is at least a few hundred out of a thousand transactions

Support

- Fraction of transactions that contain an itemset
- E.g. s({Milk, Bread, Diaper}) = 2/5

Frequent Itemset

- An itemset whose support is greater than or equal to a *minsup* threshold

Association Rule

- An Association Rule is an implication of the form $X \Rightarrow Y$ where $X, Y \subset I$, and $X \cap Y = \varnothing$.

$$\text{Support}(X \rightarrow Y) = \frac{\sigma(X \cup Y)}{|T|}$$

- Fraction of transactions containing all items of both X and Y.

Definition: Association Rule

- **Association Rule**
  - An implication expression of the form X → Y, where X and Y are itemsets
  - Example:
    {Milk, Diaper} → {Beer}

TID	Items
1	Bread, Milk
2	Bread, Diaper, Beer, Eggs
3	Milk, Diaper, Beer, Coke
4	Bread, Milk, Diaper, Beer
5	Bread, Milk, Diaper, Coke

- **Rule Evaluation Metrics**
  - Support (s)
    - Fraction of transactions that contain both X and Y
  - Confidence (c)
    - Measures how often items in Y appear in transactions that contain X

Example:

$$\{\text{Milk}, \text{Diaper}\} \Rightarrow \text{Beer}$$

$$s = \frac{\sigma(\text{Milk}, \text{Diaper}, \text{Beer})}{|\text{T}|} = \frac{2}{5} = 0.4$$

$$c = \frac{\sigma(\text{Milk}, \text{Diaper}, \text{Beer})}{\sigma(\text{Milk}, \text{Diaper})} = \frac{2}{3} = 0.67$$

Association Rule Mining Task

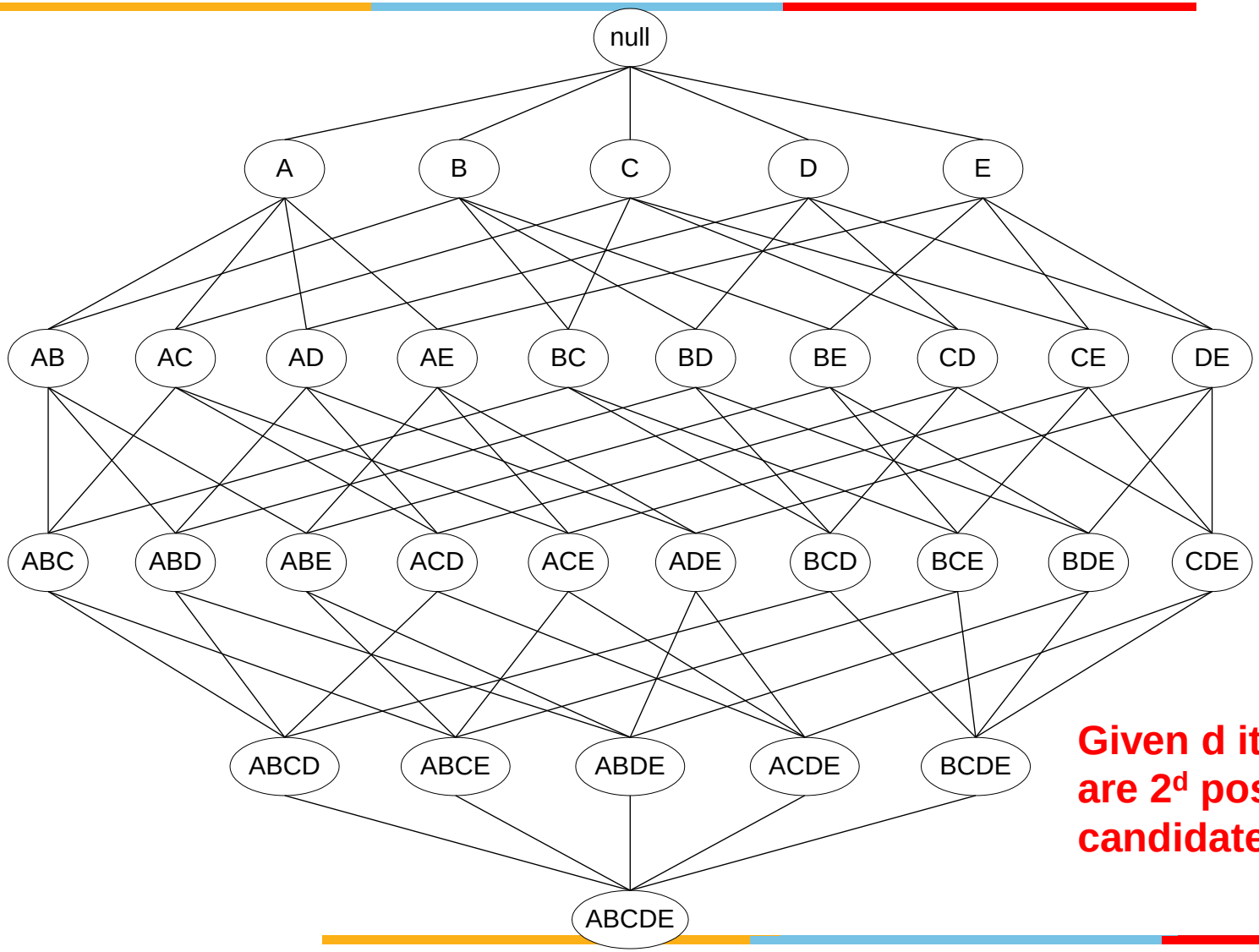
- Input: Given a set of Items I, a set of transactions T, a support threshold minsup, and a minimum confidence minconf.
- Output:Find all rules R such that support(R)>=minsup and confidence(R)>=minconf.
- Brute-force approach:
  - List all possible association rules
  - Compute the support and confidence for each rule
  - Prune rules that fail the *minsup* and *minconf* thresholds

  $\Rightarrow$ Computationally prohibitive!

Mining Association Rules

- Two-step approach:
  - Frequent Itemset Generation
    - Generate all itemsets whose support ≥ minsup
  - Rule Generation
    - Generate high confidence rules from each frequent itemset, where each rule is a binary partitioning of a frequent itemset
- Frequent itemset generation is still computationally expensive

Frequent Itemset Generation



Given d items, there are 2^d possible candidate itemsets

Take home message

- Association rule mining is traditionally called Market Basket analysis.
- Support and confidence are used to find interesting rules.
- Generating Association Rules is a combinatorial problem and hence need heretics.

Take home message (contd..)

- Association rule mining is traditionally called Market Basket analysis.
- Support and confidence are used to find interesting rules.
- Generating Association Rules is a combinatorial problem and hence need heretics.