Unsupervised and Unstructured Machine Learning

BA820 – Mohannad Elhamod

Association Rule Mining

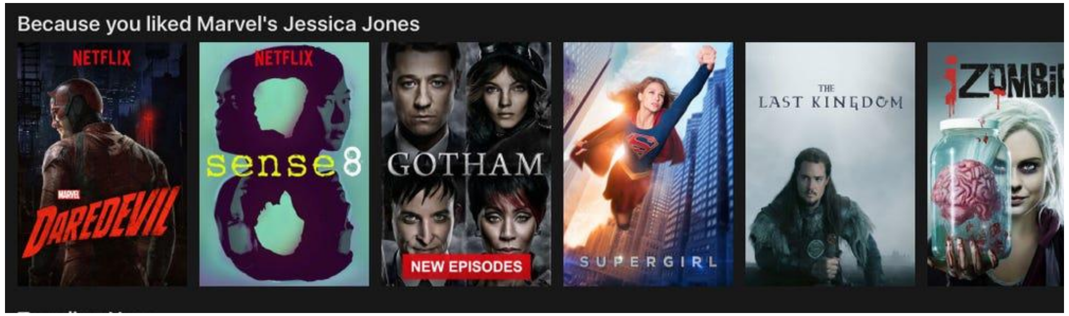
BOSTON UNIVERSITY

Boston University Questrom School of Business

What for?

- A classic application of market basket analysis addresses this question:

<u>Which items are likely to be “purchased” together?</u>



https://www.businessinsider.com/how-netflix-recommendations-work-2016-9

Transactions

- Given a set of transactions, find rules that will predict the occurrence of an item based on the occurrences of other items in the transaction

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} $\rightarrow$ {Beer},
{Milk, Bread} $\rightarrow$ {Eggs,Coke},
{Beer, Bread} $\rightarrow$ {Milk},

- **Questions:**
  - Is this causal?
  - Is this symmetric?

Introduction to Data Mining, 2nd Edition

Terminology

- `Items`: The set of objects or items available to be purchased, or viewed, or streamed.
- `Transaction`: A trip to the store, Netflix viewing history, your most recent Spotify songs streamed. A transaction contains one or more items.
- `Rule`: Can be considered an if *this* then *that*.
  - If purchase bread and eggs, then also purchase milk.
  - {Bread, Eggs} => {Milk}
- `LHS`, or `antecedent`: Left-hand side of the rule
  - The known set of objects. {Bread, Eggs}
- `RHS`, or `consequent`: The items that are associated, or co-occur with the `LHS`
  - Above, this would be {Milk}.
- `Itemset`: A collection of one or more items.

Metrics Rule Mining

- **Count (σ)**

  Frequency of occurrence of an itemset

  E.g. σ({Milk, Bread,Diaper}) = 2

- **Support**

  Fraction of transactions that contain an itemset

  E.g. s({Milk, Bread, Diaper}) = 2/5

$$Support(X) = \frac{Frequency(X)}{N}$$

$$Support(X \rightarrow Y) = \frac{Frequency(X\&Y)}{N}$$

- **Frequent Itemset**

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

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

Introduction to Data Mining, 2nd Edition

Definition: Association Rule

- **Confidence (c)**

Measures how often items in Y appear in transactions that contain X

$$Confidence(X \rightarrow Y) = \frac{Support(X \rightarrow Y)}{Support(X)}$$

- **Lift**

Measures how independent the LHS and RHS are.

$$Lift(X \rightarrow Y) = \frac{Support(X \rightarrow Y)}{Support(X)Support(Y)}$$

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**

Introduction to Data Mining, 2nd Edition

Definition: Association Rule

Example:

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

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

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

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**

But, what are these terms really?!

Rule	Support	Confidence	Lift
$A \Rightarrow D$	2/5	2/3	10/9
$C \Rightarrow A$	2/5	2/4	5/6
$A \Rightarrow C$	2/5	2/3	5/6
$B \& C \Rightarrow D$	1/5	1/3	5/9

Source: UofT

$$Rule: \ X \Rightarrow Y$$

$$Support = \frac{frq(X,Y)}{N}$$

$$Confidence = \frac{frq(X,Y)}{frq(X)}$$

$$Lift = \frac{Support}{Supp(X) \times Supp(Y)}$$

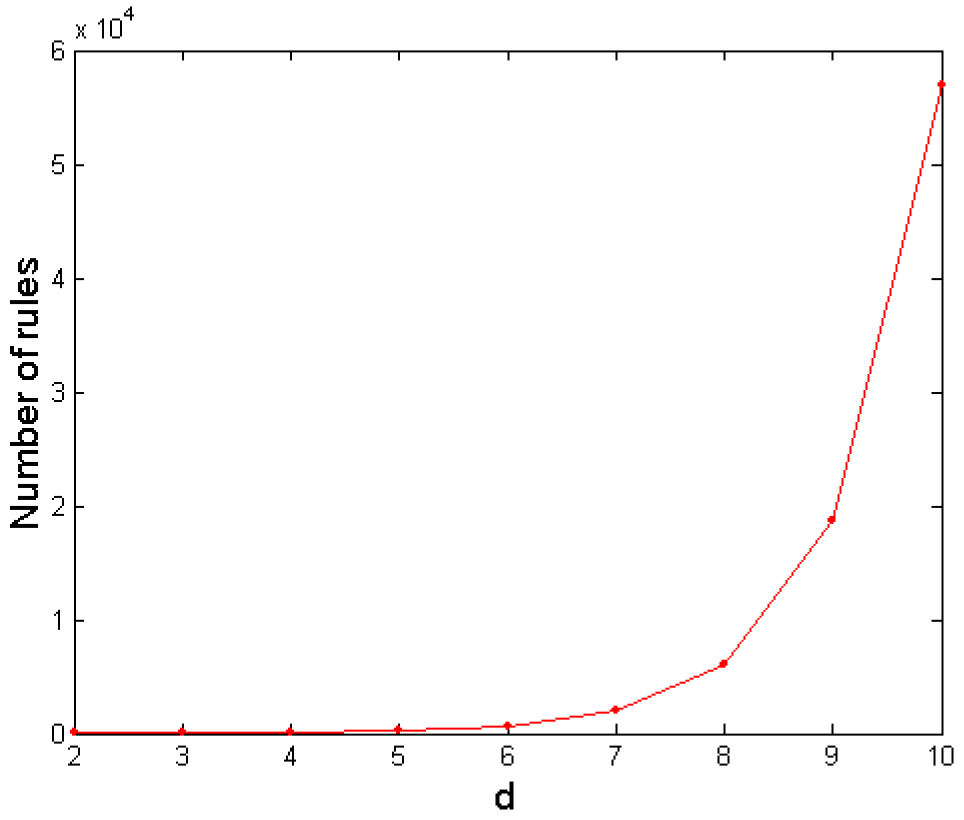
Association Rule Mining Task

- Given a set of transactions T, the goal of association rule mining is to find all rules having
  - support ≥ *minsup* threshold
  - confidence ≥ *minconf* threshold

- 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
  - ⇒ Computationally prohibitive!

Computational Complexity

- Given d unique items, total number of possible association rules as a function of number of items:



Introduction to Data Mining, 2nd Edition

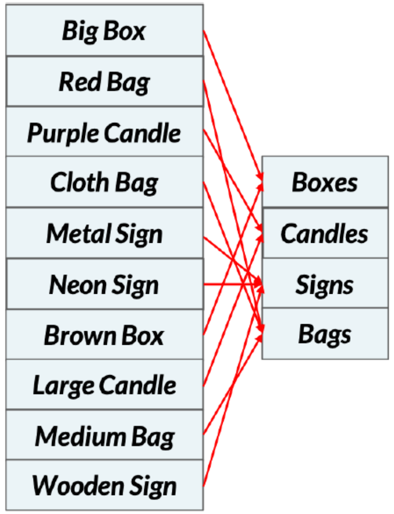
Reducing Number of Candidates

Pruning and aggregation

Pruning

- ~~Big Box~~
- Red Bag
- ~~Purple Candle~~
- ~~Cloth Bag~~
- Metal Sign
- ~~Neon Sign~~
- Brown Box
- ~~Large Candle~~
- ~~Medium Bag~~
- Wooden Sign

Aggregation



Itemset Pruning

- **Apriori principle:**
  - If an itemset is frequent, then all of its subsets must also be frequent
    - Then, if an itemset is infrequent, then all its supersets are infrequent
  - Support of an itemset never exceeds the support of its subsets

Itemset Pruning

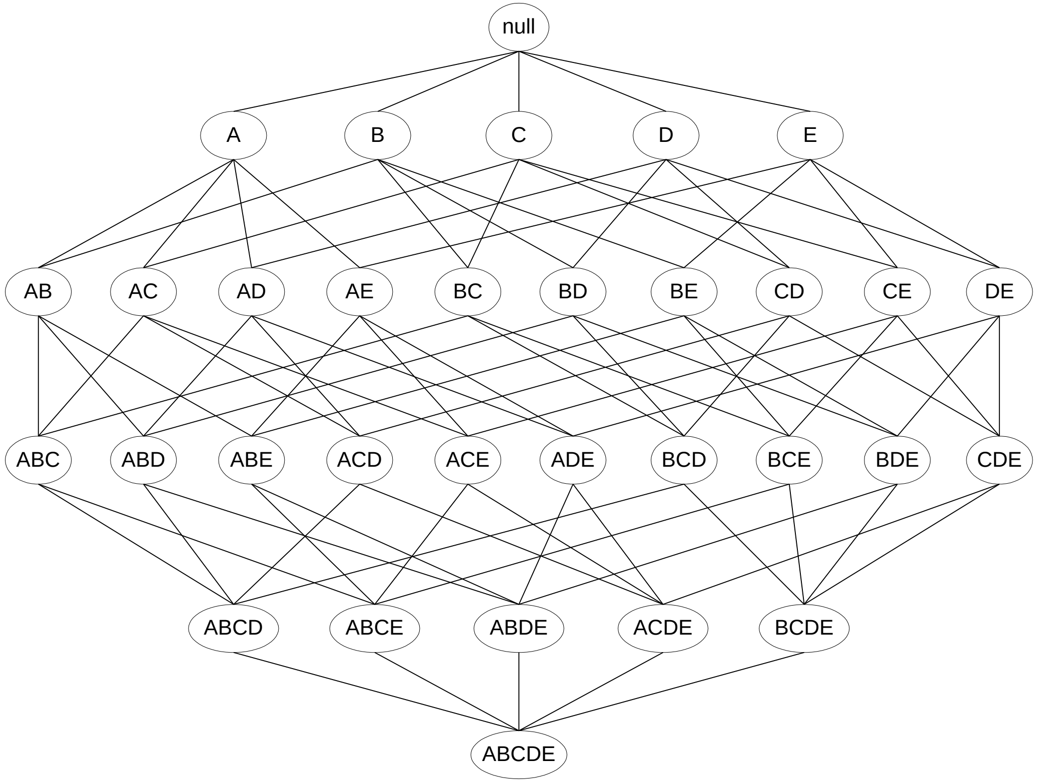
The Apriori principle

- **Apriori principle.**
  - Subsets of frequent sets are frequent.
  - Retain sets known to be frequent.
  - Prune sets not known to be frequent.

- **Candles = Infrequent**
  - -> {Candles, Signs} = Infrequent
- **{Candles, Signs} = Infrequent**
  - -> {Candles, Signs Boxes} = Infrequent
- **{Candles, Signs, Boxes} = Infrequent**
  - -> {Candles, Signs, Boxes, Bags} = Infrequent

Itemset Pruning

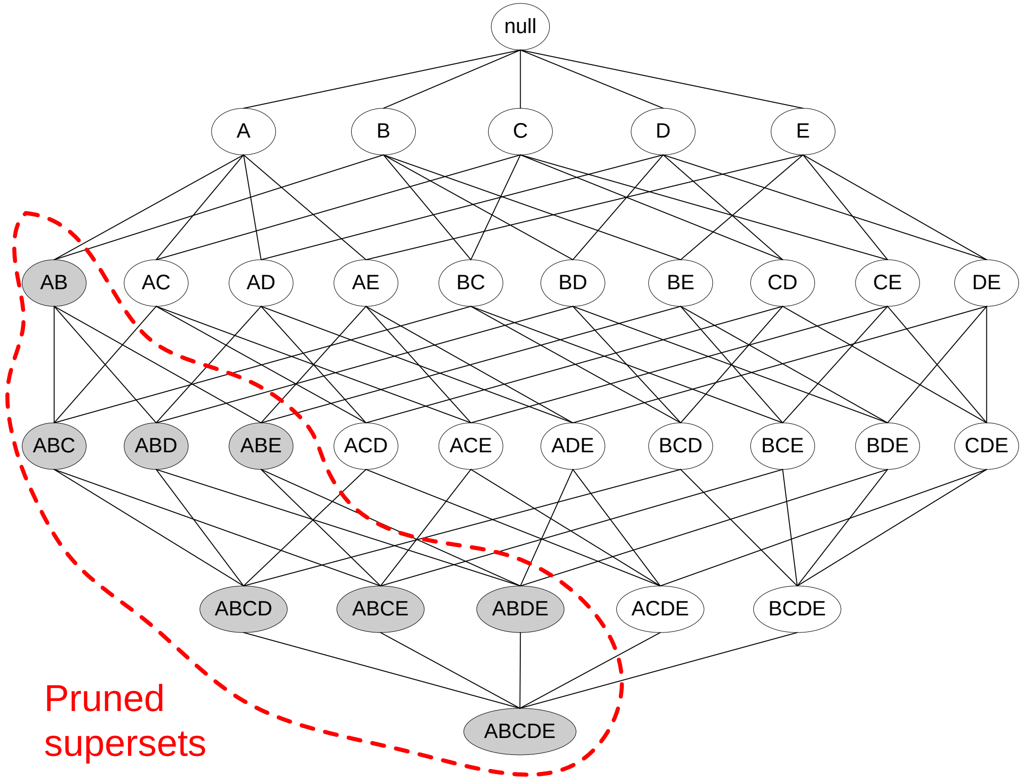
- Given d items, there are 2^d possible candidate itemsets.



Introduction to Data Mining, 2nd Edition

Itemset Pruning

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



Introduction to Data Mining, 2nd Edition

How is transactional data usually represented?

The transactional datasets can vary in how they are stored.

trans_id	item
1	b
1	c
2	c
2	a
3	c

Single format

trans_id	items
1	b,c
2	c,a
3	c

Basket format

Luckily, python/pandas make it really easy to modify the origin source data to fit the libraries expected format

How is transactional data usually represented?

trans_id	item
1	b
1	c
2	c
2	a
3	c

→

	a	b	c
1	False	True	True
2	True	False	True
3	False	False	True

mlxtend
format