UCIrvine | The Paul Merage School of Business

2022.11.24

BANA 273 Session 7

Prof. Vibs

The Paul Merage School of Business

University of California, Irvine

Agenda

- Assignment due dates on Canvas
- Please work on your projects
  - Gather Data
  - Refer to proj
- Market Basket Analysis
- Generating and identifying good Association Rules

Weka Memory

- To increase Java virtual memory for WEKA
  - On your computer go to "Start"
  - In "Search "cmd"
  - Use "cd … eka folder (under Program Files)
  - On command prompt type:
  - "Java –Xmx512m –jar weka.jar"

Why mine association rules?

- The goal may be fuzzy or unstructured

- More than on

- Interesting patterns (previously unknown) may emerge that can be used within a business

What can we do with this data?

A retailer (e.g. Target) has the following data sources:

1. Shopping transactions, 2. Shopper information, 3. Census data with information for each zip code

Transaction data set:

Shopper ID	Date and Time of transaction	Items included in the transaction	Store ID	Trans ID
Shopper111	09/10/20 12:09:01			Tran321

Shopper Information:

Shopper ID	Address	Most purchased category	R (days)	uency	Total $ (year)
Shopper111	95616	food	12	4/month	$4000

Census Data:

Zip Code	Median family Income	Median house value	Median age	Population	Population density
95616	70,000	500,000	25	60,000	5700/mile^2

Market Basket Analysis (MBA)

- MBA in retail setting
  - Find out what items are bought together
  - Cross-selling
  - Optimize shelf layout
  - Product bundling
  - Timing prom
  - Discount pla
  - Product selection under limit
  - Targeted advertisement, Pers pons, item recommendations
- Usage beyond Market Basket
  - Medical (associated symptoms)

Association Rules

Rule format:

If {set of items} → Then {set of items}

LHS impli

An association rule is valid if it satisfies some evaluation measures

Association Rule Discovery: Definition

- Given a set of records each of which contain some number of items from a given collection;
  - Produce dependency rules which will predict occurrence of an item based on

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

ered:

{Coke}

{Diaper, Milk} --> {Beer}

Association Rule Discovery: Application 1

- Marketing and Sales Promotion:
  - Let the rule discovered be

    {Bagels, … } --> {Potato Chips}

  - Potato Chips as consequent => Can be used to determine what should be done to boost its sales.
  - Bagels in the ant… which products would be affected if the … is.
  - Bagels in antecedent *and* Potato chi… nt => Can be used to see what products should be sold wi… remote sale of Potato chips.

Association Rule Discovery: Application 2

- Supermarket shelf management.
  - Goal: To identify items that are bought together by sufficiently
  - Approach: ata collected with barcode scanners to find de among items.
  - A classic rule --
    - If a customer buys diaper and milk, then he is very likely to buy beer.
    - So, don't be surprised if you find six-packs stacked next to diapers.

Association Rule Discovery: Application 3

- Inventory Management:
  - Goal: A consumer appliance repair company wants to anticipate the nature of repairs on its consumer products and keep the service vehicles equi on number of visits to consumer ho
  - Approach: Process the data on to equired in previous repairs at different consumer loc over the co-occurrence patterns.

Association Rule Mining

- 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 trans

of Association Rules

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

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

Implication means co-occurrence, not causality!

Definition: Frequent Itemset

- **Itemset**
  - A collection of one or more items
    - Example: {Milk, Bread, Diaper}
  - k-itemset
    - An itemset that contains k items
- **Support count (σ)**
  - Frequency of occur
  - E.g. σ({Milk, Bre
- **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

TID	*Items*
1	Bread, Milk
2	Bread, Diaper, Beer, Eggs
	Milk, Diaper, Beer, Coke
	Bread, Milk, Diaper, Beer
	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 transac contain X

$$\textbf{Support} = \frac{\text{No. of transactions containing items in LHS and RHS}}{\text{Total No. of transactions in the dataset}}$$

$$\textbf{Confidence} = \frac{\text{No. of transactions co}}{\text{LHS}} \quad \frac{\text{LHS and RHS}}{\text{tions containing}}$$

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:

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

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

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

Rule Evaluation - Lift

Transaction No.	Item 1	Item 2	Item 3	Item 4	…
100	Beer	Diaper	Chocolate		
101	Milk	Chocolate	Shampoo		
102	Beer	Milk	Vodka	Chocolate	
103	Beer	Milk	Diaper	Chocolate	
104	Milk	Diaper	Beer		

What's the support a te}→{Milk}?

Support = 3/5 Confidence = 3/4

Very high support and confidence.
Is Chocolate a good predictor of Milk purchase?

No! Because Milk occurs in 4 out of 5 transactions. Chocolate is even decreasing the chance of Milk purchase
3/4 < 4/5, i.e. P(Milk|Chocolate)<P(Milk)

Lift = (3/4)/(4/5) = 0.9375 < 1

Rule Evaluation – Lift (cont.)

- Measures how much more likely is the RHS given the LHS than merely the RHS
- Lift = confidence of the rule / benchmark confidence

  Benchmark Confidence

  = Count of RHS / Number of transactions in database

Example: {Diaper

  - Total numb 0
  - No. of customers buying **Diaper**
  - No. of customers buying **beer**:
  - No. of customers buying **Diaper** & **beer**: 20

- Benchmark confidence of Beer = 50/1000 (5%)
- Confidence = 20/200 (10%)
- Lift = 10%/5% = 2
- Higher Lift indicates better rule

Mining Association Rules

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

Example of Rules:

{Milk,Diaper} → {Beer} (s=0.4, c=0.67)
{Milk,Beer} → {Diaper} (s=0.4, c=1.0)
{Diaper,Beer} → {Milk} (s=0.4, c=0.67)
{Beer} → {Milk,Diaper} (s=0.4, c=0.67)
lk,Beer} (s=0.4, c=0.5)
r,Beer} (s=0.4, c=0.5)

Observations:

- All the above rules are binary partitions of the same itemset:
  {Milk, Diaper, Beer}
- Rules originating from the same itemset have identical support but can have different confidence
- Thus, we may decouple the support and confidence requirements

Association Rule Mining Task

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

- 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!

Exercise 1

Compute the support for subsets {a}, {b, d}, and {a,b,d} by treating each transaction ID as a market basket.

Exercise 2

Use the results in the previous problem to compute the confidence for the association rules {b, d} → {a} and {a} → {b, d}. State what these values mean in plain English.

Exercise 3

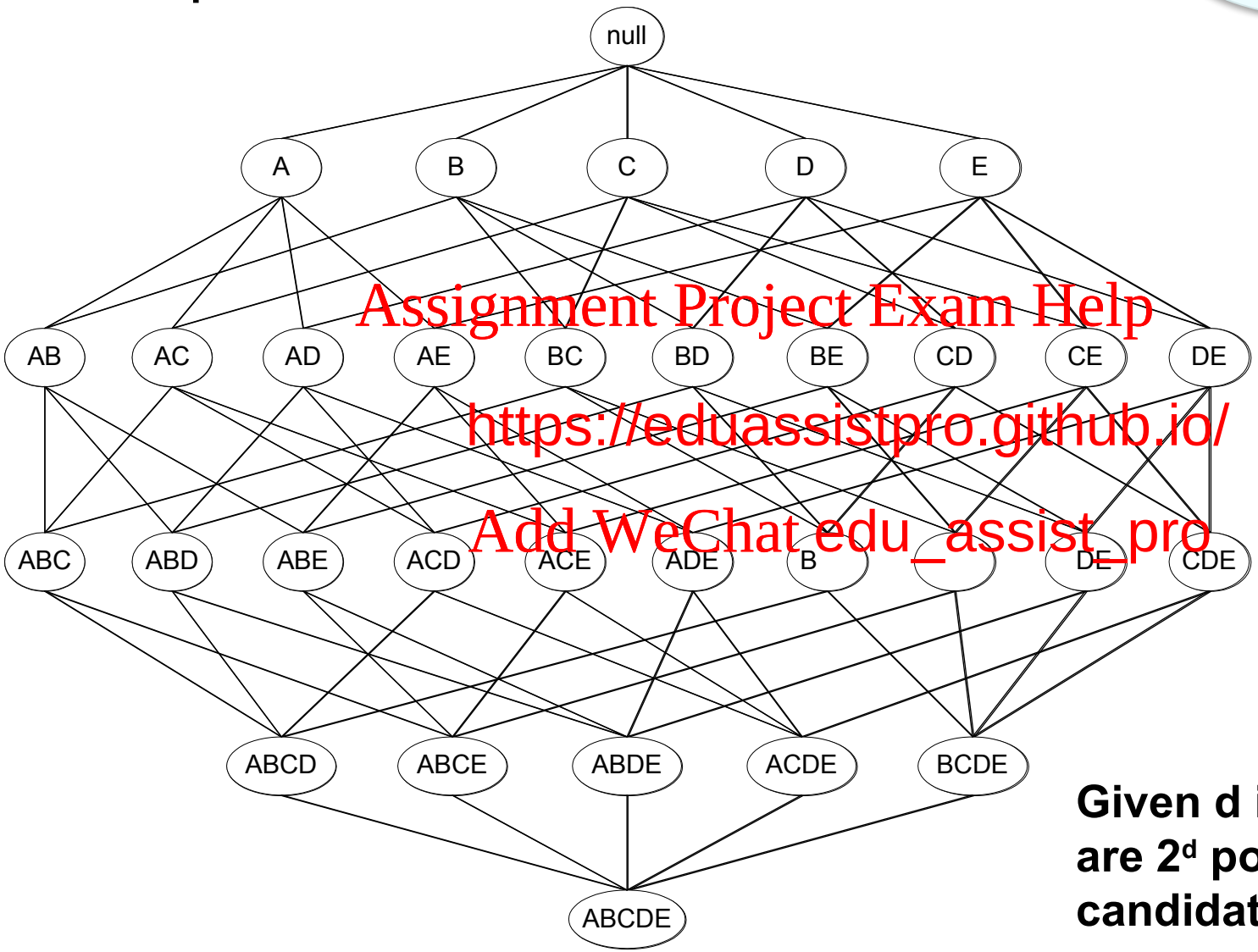
Compute the support for itemsets {a}, {b, d}, and {a,b,d} by treating each customer ID as a market basket.

Exercise 4

Use the results in the previous problem to compute the confidence for the association rules {b, d} → {a} and {a} → {b, d}.

Frequent Itemset Generation

Optional

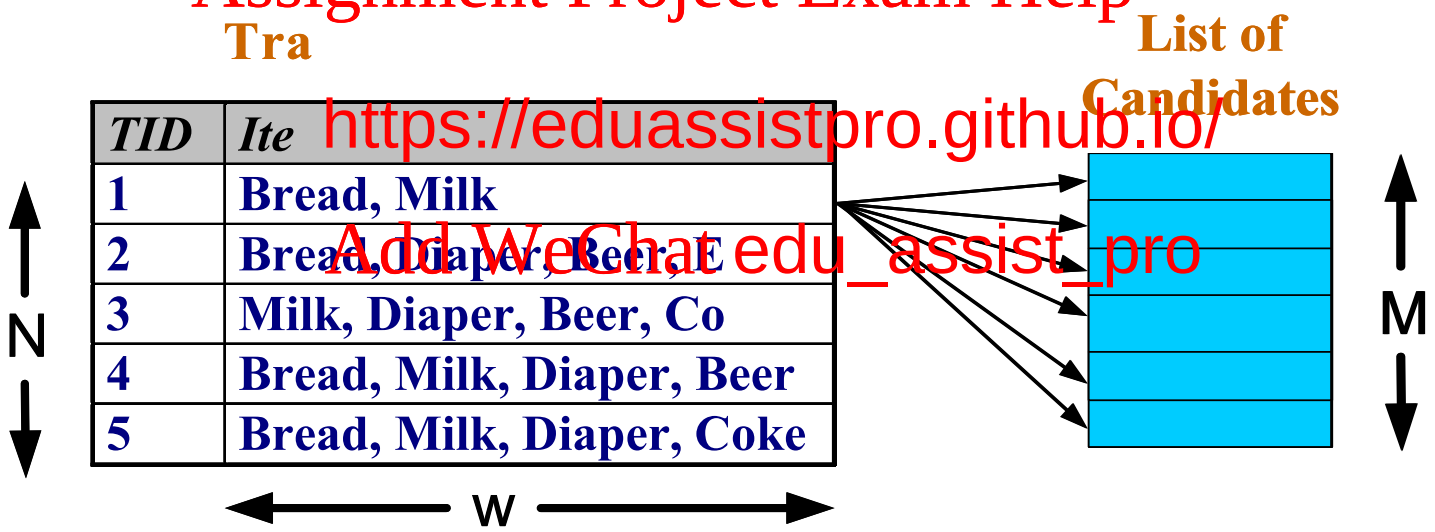


Given d items, there are 2^d possible candidate itemsets

Frequent Itemset Generation

Optional

- Brute-force approach:
  - Each itemset in the lattice is a candidate frequent itemset
  - Count the support of each candidate by scanning the database



  - Match each transaction against every candidate
  - Complexity ~ O(NMw) => Expensive since $M = 2^d$!!!

Computational Complexity

Optional

- Given d unique items:
  - Total number of itemsets = 2^d
  - Total number of possible association rules:

$$R = \sum_{k=1}^{d-1} \left[\binom{d}{k} \times \sum_{j=1}^{d-k} \binom{d-k}{j} \right] = 3^d - 2^{d+1} + 1$$

If d=6, R = 602 rules

Identifying Association Rules

- Two-step approach:
  1. Frequent Itemset Generation
     - Generat insup
  2. Rule Generation
     - Generate high confidence r frequent itemset, where each rule is a binary partitioning of a frequent itemset

- Frequent itemset generation is still computationally expensive

Phase 1: Finding all frequent itemsets

How to perform an efficient search of all frequent itemsets?

If **{diaper, beer}** is frequent then **{diaper}** and **{beer}** are each frequent as well

This means that…

- If an itemset is not frequent (e.g., **{wine}**) then no itemset that includes wine can be frequent either, such as **{wine, beer}** .
- We therefore first find all frequent itemsets of size 1.
  Then try to "expand" all itemsets of size 2 that include frequent itemsets of size 1.

  Example:

  If **{wine}** is not frequent we need not try to find out whether **{wine, beer}** is frequent. But if both **{wine}** & **{beer}** were frequent then it is possible (though not guaranteed) that **{wine, beer}** is also frequent.
- Then take only itemsets of size 2 that are frequent, and try to expand those, etc.

Phase 2: Generating Association Rules

Assume **{Milk, Bread, Butter}** is a frequent itemset.

- Using items contained in the itemset, list all possible rules
  - {Milk} → {Bread, Butter}
  - {Bread} → {Milk, Butter}
  - {Butter} → {Milk, Bread}
  - {Milk, Bread} → {Butter}
  - {Milk, Butt
  - {Bread, But

- Calculate the confidence of each r
- Pick the rules with confidence above the minimum confidence

Confidence of {Milk} → {Bread, Butter}:

$$\frac{\text{No. of transaction that support \{Milk, Bread, Butter\}}}{\text{No. of transaction that support \{Milk\}}} = \frac{\text{Support \{Milk, Bread, Butter\}}}{\text{Support \{Milk\}}}$$

Algorithm Apriori-Gen to Generate Frequent Itemsets

(Agrawal and Srikant 1994)

Input: two itemsets, *I and J, of size (k-1)*

Output: a supported itemset, *L of size k*

L=Null

IF (the first (*k-2) items of I and J match)*

{ copy all items of

copy the

FOR (every subset (*l* of *L*)

IF (*l is not supported) discard L and exit;*

Calculate, from data, support(*L);*

IF (support(*L*)< target) discard *L and exit;*

return *L;*

}

ELSE exit;

Agrawal (94)'s Apriori Algorithm—An Example

Transactions

T-ID	Items
10	A, C, D
20	B, C, E
30	A, B, C, E
40	B, E

An itemset must have been purchased at least twice in order to be considered frequent or supported

Agrawal (94)'s Apriori Algorithm—An Example

Transactions

T-ID	Items
10	A, C, D
20	B, C, E
30	A, B, C, E
40	B, E

C_1

Itemset	sup
{A}	2
{B}	3
{C}	3
{D}	1
{E}	3

1st scan

L_1

Itemset	sup
{A}	2
{B}	3
{C}	3
{E}	3

2

Itemset
{A, B}
{A, C}
{A, E}
{B, C}
{B, E}
{C, E}

2

Itemset	sup
{A, B}	1
{A, C}	2
{A, E}	1
{B, C}	2
{B, E}	3
{C, E}	2

L_2

Itemset	sup
{A, C}	2
{B, C}	2
{B, E}	3
{C, E}	2

C_3

Itemset
{B, C, E}

3rd scan

L_3

Itemset	sup
{B, C, E}	2

{A,B,C}, {A, C, E}?

Exercise 5

Transaction No.	Item 1	Item 2	Item 3	Item 4
100	Beer	Diaper	Chocolate	
101	Milk	Chocolate	Shampoo	
102	Beer	Soap	Vodka	
103	Beer	Cheese	Wine	
104				Chocolate

Given the above list of transactions owing:

1) Find all the frequent itemsets (m port 40%)
2) Find all the association rules (minimum confidence 70%)
3) For the discovered association rules, calculate the lift

Given table below, the confidence of the rule Bread → Coke is

A: 1/4

B: 2/4

C: 3/4

D: 4/4

E: None of the

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

What is the objective of Apriori?

A: Identify good association rules

B: Identify all frequent itemsets

C: Identify all rules from frequent itemsets

D: Determine t plexity of finding asso

E: None of the above

Other Applications of Association Rules

- Recommendations: Determines which books are frequently purchased together and recommends associated books or products to people who express interest in an item.
- Healthcare: Studying the side-effects in patients with multiple presc reviously unknown interactions an .
- Fraud detection: Finding in ins that a certain doctor often works with a cert ay indicate potential fraudulent activity. (virtual items)
- Sequence Discovery: looks for associations between items bought over time. E.g., we may notice that people who buy chili tend to buy antacid within a month. Knowledge like this can be used to plan inventory levels.

WEKA

- Find association rules
  - Apriori
- http://facweb.cs.depaul.edu/mobasher/classes/ect584/WEKA/associate.html

Nest Week

- Clustering using K-Means