Applied Statistical Multivariate Analysis

Association Rules

Instructor: Yaser Zerehsaz

موسسه آموزش عالی آزاد توسعه

برگزار کننده دوره‌های تخصصی علم داده

Association rules

"Association rules attempt to construct simple descriptions that describe regions of high density in the special case of very high-dimensional binary-valued data"

Hastie, Tibshirani and Friedman (2009)

Mathematical definition

- If we have a feature vector $(X_1, X_2, \dots, X_p)$, the goal of association rules is to find the joint values of $X_1, X_2, \dots, X_p$ that appear most frequently in the dataset
- It is most often applied to binary data $X_j \in \{0,1\}$
- In general format, the goal is to find $(X_1 = v_1, X_2 = v_2, \dots, X_p = v_p)$ where $P(X_1 = v_1, X_2 = v_2, \dots, X_p = v_p)$ is relatively large
- This is similar to mode finding, but the problem is

$$\hat{P}(X_1 = v_1, X_2 = v_2, \dots, X_p = v_p) = \frac{N(X_1 = v_1, X_2 = v_2, \dots, X_p = v_p)}{N} \cong 0$$

Mathematical definition

- Change the goal of problem to finding $s_j's$ that maximize

$$P(\bigcap_{j=1}^{p} X_j \in s_j)$$

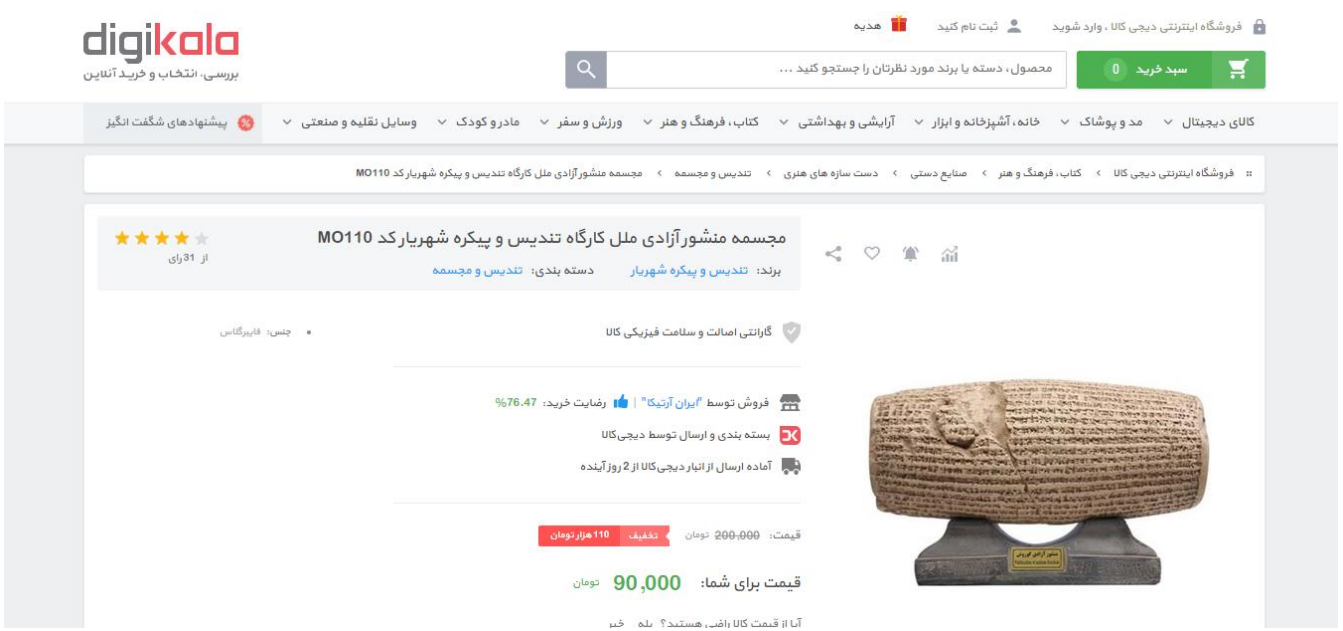
where s_j is a subset of all the values X_j can take.

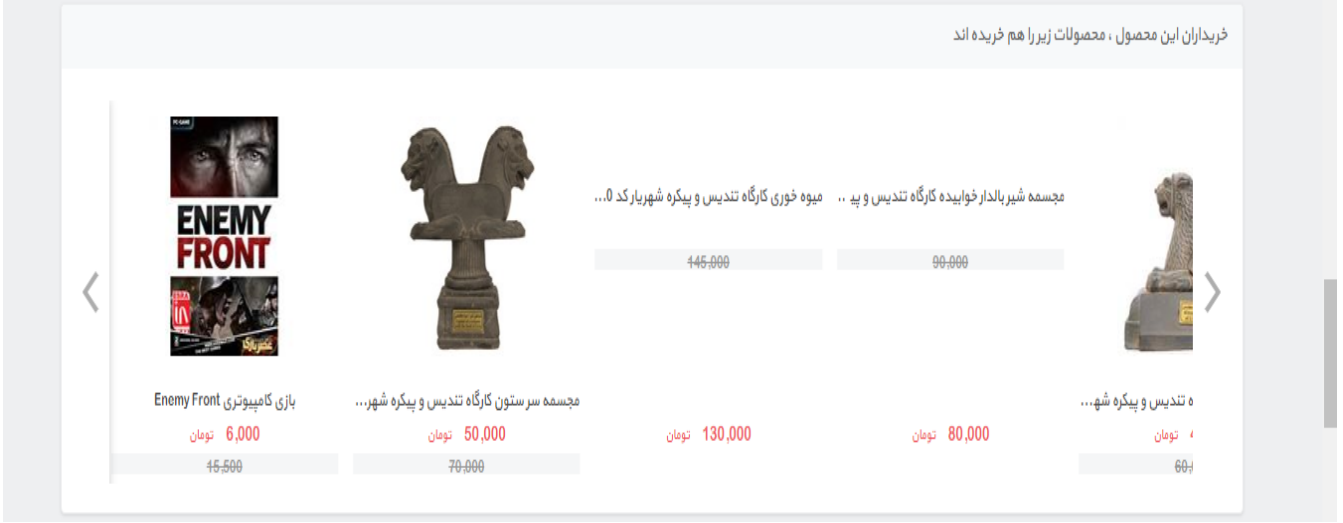
Practical aspect

- Finding frequent patterns, association, correlation or causal structures among sets of items in transaction databases
- Understand customers' buying habits by finding associations and correlations between different items that customers place in their shopping basket
- Basket data analysis, cross-marketing, catalog design, loss-leader analysis, web log analysis, fraud detection

Practical aspect

Practical aspect





Rule format

Antecedent → Consequent [support, confidence]

- Support and confidence measure interestingness

- Buys (movie ticket) → Buys (popcorn)
- Age (28) and Degree (M.Sc. in IE) → Income (3000 $)

Interpretations of Association rules

Let the rule be:

{Spaghetti} → {Ketchup}

For the consequent:

- What can we do to increase the sales of Ketchup?
- Where should we put Ketchup to boost its sales?

For the antecedent:

- If I do not sell Spaghetti, how the other items' sales will be affected?
- If I increase the number of Spaghetti packets, what items will be sold more?

Concept

Inputs:

- A dataset with some transactions
- Each transaction is a list of items purchased by a customer in a visit

Transaction ID	Items
1	{Milk, Cheese, Bread, Diapers}
2	{Milk, Hotdog, ketchup, Tomato, Banana}
3	{Milk, Bread, Yogurt, Cheese, Potato chips}
4	{Yogurt, Gums, Nail clipper, Potato, Potato chips}
5	{Milk, Cheese, Bread, Candy bars, Chocolate bar}

Concept

Outputs:

- all rules that correlate the presence of one set of items (itemset) with that of another set of items
- E.g., 75% of people who purchase milk also get bread and cheese

Transaction ID	Items
1	{Milk, Cheese, Bread, Diapers}
2	{Milk, Hotdog, ketchup, Tomato, Banana}
3	{Milk, Bread, Yogurt, Cheese, Potato chips}
4	{Yogurt, Gums, Nail clipper, Potato, Potato chips}
5	{Milk, Cheese, Bread, Candy bars, Chocolate bar}

$$\{Milk\} \rightarrow \{Cheese, Bread\}$$

Strength of rules

Define:

- Set of all items in a market basket dataset $I = \{i_1, i_2, \ldots, i_d\}$
- The set of all transactions $T = \{t_1, t_2, \ldots, t_N\}$
- Each transaction t_i contains a subset of items in I
- An itemset is a set of zero or more items

- Support: the number of transactions containing an itemset X

$$\sigma(X) = |\{t_i | X \subseteq t_i, t_i \in T\}|$$

where X is an itemset, $\sigma(X)$ is its support and $|\cdot|$ is the number of elements in a set

- The support of itemset {Milk, Cheese, Bread} is 3.

Metrics

- Assume that the association rule is of form $X \rightarrow Y$, where X and Y are disjoint itemsets, i.e., $X \cap Y = \emptyset$.
- The support of the rule is defined as

$$s(X \rightarrow Y) = \frac{\sigma(X \cup Y)}{N}$$

- e.g. $s(\{\text{Milk}\} \rightarrow \{\text{Cheese, Bread}\}) = \frac{\sigma(\{\text{Milk,Cheese, Bread}\})}{N} = \frac{3}{5}$
- It is like an estimate of probability of observing both X and Y
- Confidence of a rule measures how frequently items in Y appear in transactions that contain X

$$c(X \rightarrow Y) = \frac{\sigma(X \cup Y)}{\sigma(X)}$$

- e.g. $c(\{\text{Milk}\} \rightarrow \{\text{Cheese, Bread}\}) = \frac{\sigma(\{\text{Milk,Cheese, Bread}\})}{\sigma(\text{Milk})} = \frac{3}{4}$
- It is an estimate of probability of observing Y given X

Metrics

- Lift of rule $X \rightarrow Y$ is defined as confidence divided by support of Y

$$L(X \rightarrow Y) = \frac{C(X \rightarrow Y)}{\sigma(Y)}$$

- Lift estimates the probability of $\frac{Pr(X\ and\ Y)}{Pr(Y)Pr(X)}$
- The lift metric is commonly used to measure how much more often the antecedent and consequent of a rule $X \rightarrow Y$ occur together than we would expect if they were statistically independent.
- If X and Y are independent, the Lift score will be exactly 1.

Metrics

- Leverage of rule $X \rightarrow Y$ is defined as

$$Lev(X \rightarrow Y) = S(X \rightarrow Y) - S(X)S(Y)$$

- Leverage computes the difference between the observed frequency of X and Y appearing together and the frequency that would be expected if X and Y were independent.
- A leverage value of 0 indicates independence
- Conviction of rule $X \rightarrow Y$ is defined as

$$Conv(X \rightarrow Y) = \frac{1-S(Y)}{1-C(X \rightarrow Y)}$$

- A high conviction value means that the consequent is highly depending on the antecedent.
- In the case of a perfect confidence score, the denominator becomes 0 (due to 1 - 1) for which the conviction score is defined as 'inf'.

Metrics

- If the support of a specific rule is low, the rule might have happened by chance
- Support is used to eliminate uninteresting rules
- Confidence measures the reliability of a rule
- The higher $c(X \rightarrow Y)$ is, the more probable is to observe Y in a basket containing X
- Association analysis does not necessarily imply causality

Association Rule Mining

Definition:

Having a set of transactions T, find all the rules having support $s \geq minsup$ and confidence $c \geq minconf$ where $minsup$ and $minconf$ are the support and confidence thresholds

Common strategy:

Frequent itemset generation, whose objective is to find all the itemsets satisfying $minsup$ threshold. These are frequent itemsets

Rule generation, whose objective is to extract all the high confidence rules from the frequent itemsets obtained in the previous step. These rules are called strong rules

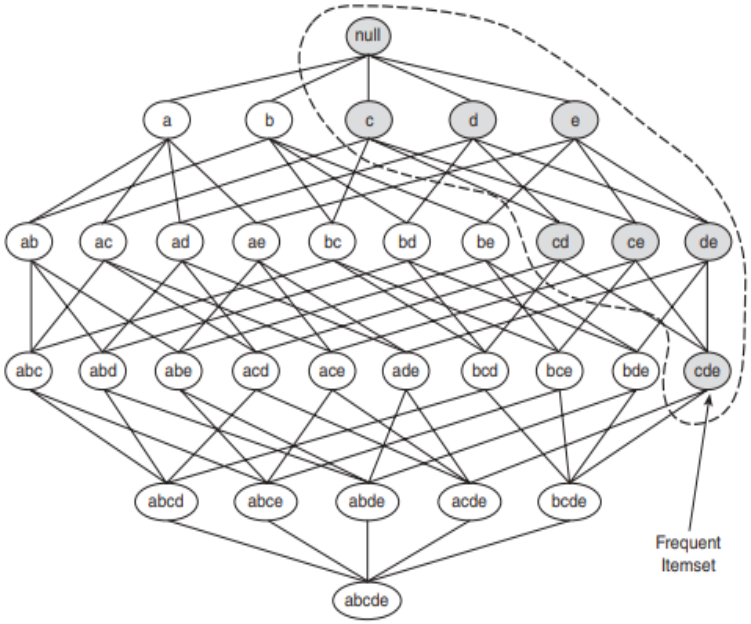
Apriori Algorithm

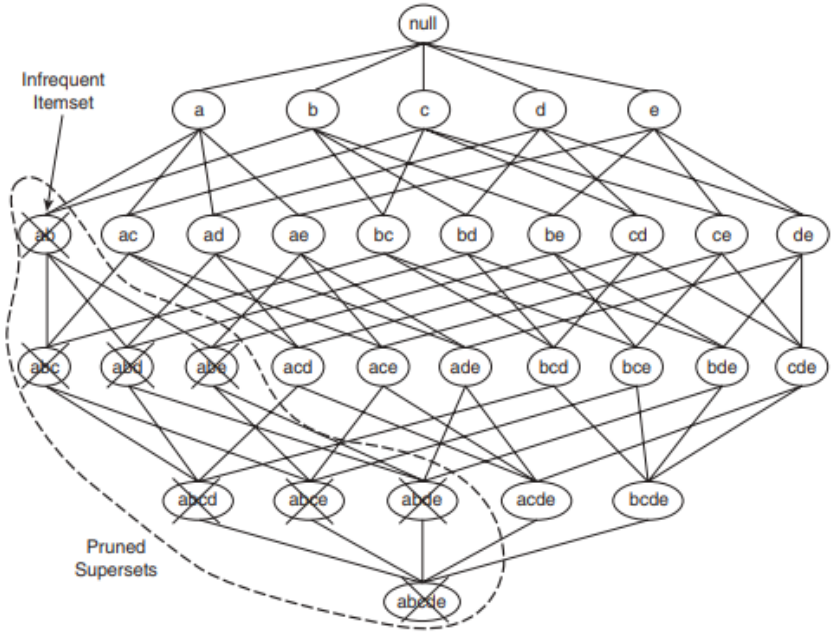
- The algorithm is developed by Agrawal et al. (1995)
- It works based on Apriori Principal

Apriori Principal:

- Any subset of a frequent itemset must be frequent
- A transaction containing {milk, cheese, bread} also contains {cheese, bread}
- No superset of any infrequent itemset should be generated or tested

Apriori Principal





Apriori Algorithm- Example

Assume that $minsup = 50\%$ and $minconf = 60\%$

Transaction ID	Items
1	{Milk, Yogurt, Cheese, Bread, Diapers}
2	{Milk, Hotdog, ketchup}
3	{Milk, Bread, Yogurt, Cheese, Potato chips}
4	{Yogurt, Potato chips}
5	{Milk, Cheese, Bread, Banana}
6	{Yogurt, Potato chips, , ketchup}

Apriori Algorithm- Single Itemsets

- Scan the main dataset and compute support for all single itemsets
- Check whether or not the support is larger than or equal to $minsup = 50\%$

Single itemset	Support
{Milk}	4
{Cheese}	3
{Bread}	3
{Diapers}	1
{Hotdog}	1
{Ketchup}	2
{Yogurt}	4
{Potato chips}	3
Banana	1

Frequent itemsets	Support
{Milk}	4
{Cheese}	3
{Bread}	3
{Yogurt}	4
{Potato chips}	3

Apriori Algorithm- Double Itemsets

- Scan the main dataset and compute support for all double itemsets
- Check whether or not the support is larger than or equal to $minsup = 50\%$

Double itemsets	Support
{Milk, Cheese}	3
{Milk, Bread}	2
{Milk, Yogurt}	2
{Milk, Potato chips}	1
{Cheese, Bread}	3
{Cheese, Yogurt}	2
{Cheese, Potato chips}	1
{Bread, Yogurt}	2
{Bread, Potato chips}	1

Frequent itemsets	Support
{Milk, Cheese}	3
{Cheese, Bread}	3

Apriori Algorithm- Triple Itemsets

- The candidate triple itemset is:

  {Milk, Cheese, Bread}

- But, we are not going to proceed with this itemset
- The reason is that the itemset {Milk, Bread} is not frequent. So, any superset of this itemset will not be frequent

Frequent itemset	Support
{Milk, Cheese}	3
{Cheese, Bread}	3

Apriori Algorithm- Generate Rules

Frequent itemsets	Support
{Milk}	4
{Cheese}	3
{Bread}	3
{Yogurt}	4
{Potato chips}	3

Frequent itemsets	Support
{Milk, Cheese}	3
{Cheese, Bread}	3

- For each itemset, choose all subsets and extract the rules
- Consider the itemset {Milk, Cheese}
- The subsets for this itemset are {Milk}, {Cheese}
- For rule {Milk} → {Cheese}

$$c(\{\text{Milk}\} \rightarrow \{\text{Cheese}\}) = \frac{\sigma(\{\text{Milk,Cheese}\})}{\sigma(\text{Milk})} = \frac{3}{4}$$

- For rule {Cheese} → {Milk}

$$c(\{\text{Cheese}\} \rightarrow \{\text{Milk}\}) = \frac{\sigma(\{\text{Milk,Cheese}\})}{\sigma(\text{Cheese})} = \frac{3}{3}$$

311 Service Requests

www.govtech.com

NYC 311 Service Requests 2014

- 311 is a non-emergency call system taking different types of complaints
- It connects people to different agencies based on the type of complaints
- There is a website providing datasets regarding service requests since 2010
- The datasets contain different types of complaints from five boroughs of New York city

NYC 311 Service Requests 2014

- There are six variables in the dataset

1- Created Month

January, February,…,June

2- Agency

New York Police Department (NYPD), Department of Transportations (DOT), Housing Preservation and Development (HPD), Department of Sanitation (DSNY), …

3- Complaint Type

Heating, Blocked driveway, Noise, Street condition, Street light condition, Paint, Pluming,…

4- Descriptor

Heat, Street light out, Loud Party, ...

NYC 311 Service Requests 2014

5- Location Type

Residential building, street,…

6- Borough

Brooklyn, Queens, Staten Island, Bronx, Manhattan

Lets extract some rules!