STA380 Aug 8th 2016 Second Session

Apriori Algorithm (also called: Market-Basket analysis or Association Rule mining)

By *Yawen Ye* and *Yiqing Alice Zhu*

Imagine we are in a relational database. we have different TID in one column, and second column contains combination of items corresponding to each TID. We are going to use this table to illustrate the following examples.

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

Association Rules

- An implication expression of the form X --> Y, where X and Y are item sets
  Example: {Diaper} --> {Beer}

  {Milk, Bread} --> {Eggs,Coke}

Note: The arrow (Implication) just means correlation, not causality!

Frequent Items set

Item set

- A collection of one or more items
  Ex: {Milk, Bread, Diaper}

- K-item set
  Ex: An item set that contains K items

Support count (**σ**)

- Frequency of occurrence of an item set
  Ex: **σ** ({Milk, Bread, Diaper]) = 2

Support

- ({items set}) = fraction of transactions that contains an item set
  Ex: **σ** ({Milk, Bread, Diaper}) = 3/5 = 0.4

Rule Evaluation Metrics

Suppose we have {X} --> {Y}

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

Let's look an example: {Milk, Diaper} --> {Beer}

- Support(s) = **σ** (Milk, Diaper, Beer) / (Total TID)
  = 3/5 = 0.6

- Confidence(c) = Probability({Beer} | {Milk, Diaper})
  = (Milk, Diaper, Beer) / **σ** (Milk, Diaper)

  = 3/4 = 0.75

How to calculate TOTAL number of possible association rules in a given dataset?

For d items, there are $3^d - 2^{d+1} + 1$ possible association rules.

For example, if d = 6, there are 206 possible association rules;

if d = 10, there are more than 6*(10^4) possible association rules. (Too many? No problem. We do pruning under apriori principle.

Illustrating Apriori Principle

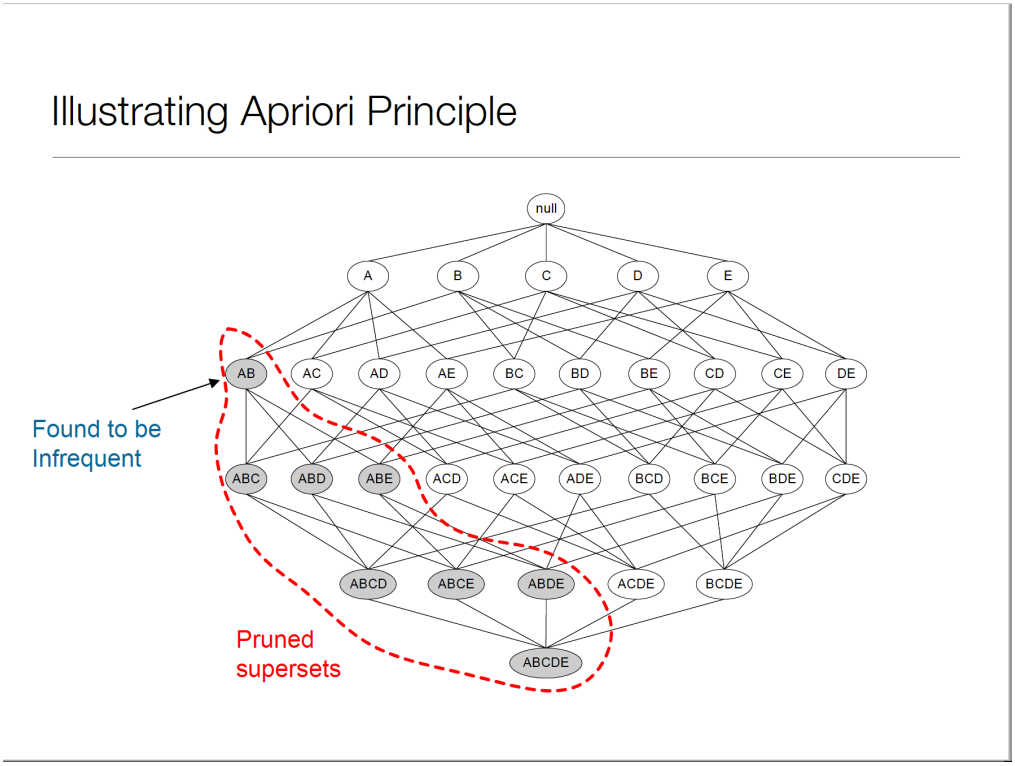

Image1: from Pradeep Ravikumar's notes on association rule mining

The principle basically says that for sets that are found to be infrequent, in other words, with lower support, we pruned the whole supersets to reduce computational burdens. In this case we take set AB for example.

From the picture above, AB fails to meet our support threshold and so do all its subsets. Thus, we pruned AB and it descendants.

R Studio Example

Preparation

- **Package** used in this example: arules

Arules also has a big ecosystem of packages built around it.

- **Datasets** used in this example: playlist.csv
- **R Script** used in this example: playlists.R

Highlights of this example

- Turn user into a factor:

Users were originally integers in the data and we turned them to categorical values.

```
# Turn user into a factor
playlists_raw$user <- factor(playlists_raw$user)
```

- Split data into a list of artists for each user

Apriori algorithm expects a list of baskets in a special format. In this case, one 'basket' of songs were created for each user by spliting the data into a list of artists for each of them.

```
playlists <- split(x=playlists_raw$artist, f=playlists_raw$user)
```

We certainly don't want duplicate items in one basket under apriori analysis. Thus, we drop duplicates in each basket.

```
playlists <- lapply(playlists, unique)
```

- Interpretation of **lift**

After getting the results of the apriori algorithm, let's try to interpret ***lift.***

Take line16 for example:

```
## 16 {snow patrol,the killers} => {coldplay}  0.0104  0.5954198 3.7558
02
```

Among people who listen to snow patrol and the killers, 59.54% of them also listens to coldplay. Yet,we don't know whether this high percentage is due to coldplay's popularity or there's a true pattern in people's music choices between snow patrol, the killers and coldplay.

Lift does a good job in evaluating that. A lift of 3.75 means that people who listen to snow patrol and the killers are 3.75 times more likely than other people to have coldplay on their lists.

There's a great possibility that we may observe high lift but low confidence. Let's see the following example.

If we change our threshold to a support level of .001 and confidence level of .1, and increase the lift level up to above 50, a big chunk of data still showed up when we inspect results. This suggests that it's fairly possible that items appeared less often in a certain basket indeed has a strong correlation with items in that basket.

Q&A Q: How to choose threshold? A: The bigger the data set, the looser the thresold.