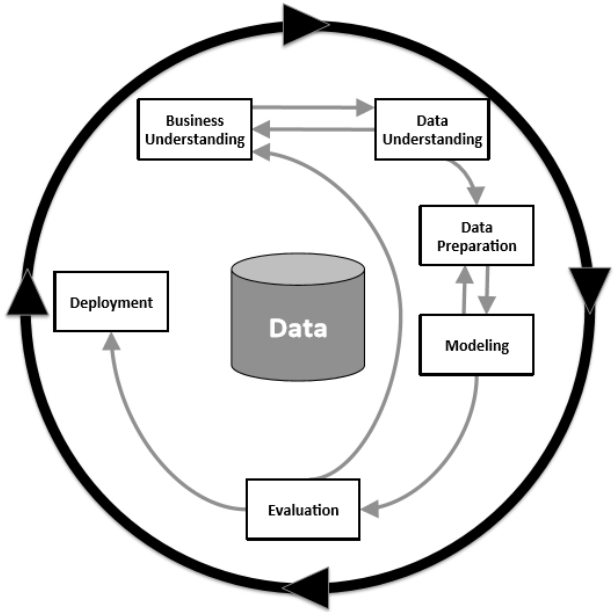
Advanced Data & Network Mining

Modelling - Unsupervised Learning Association Rules

2023-24

terri.hoare@dbs.ie

Recap on Challenges and Methodology
CRISP-DM (Cross Industry Standard Process for Data Mining)



The CRISP-DM process, including the six key phases and the important relationships between them (adapted from Wirth and Hipp, 2000, repr. in Kelleher *et al.*, 2020, p.14)

Data MiningTaxonomy
Matching Problems to Data Mining Algorithms

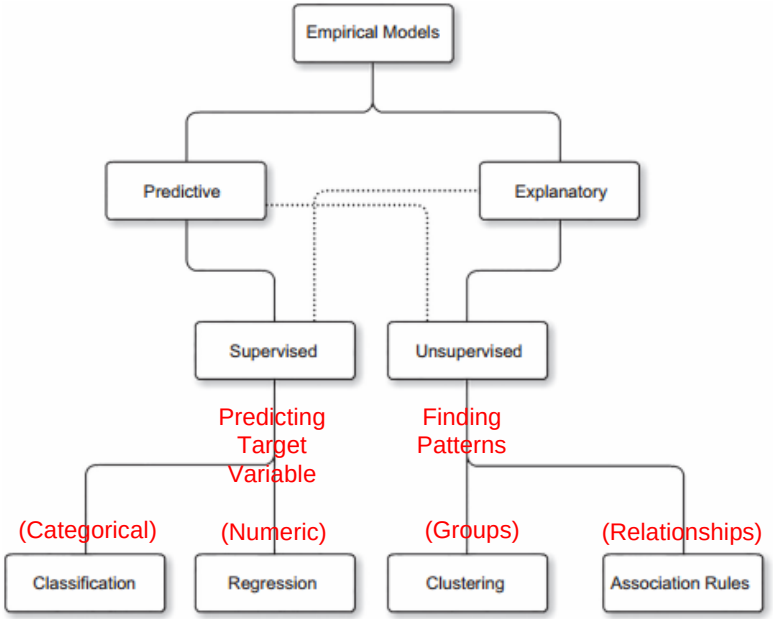
Classification
Predicting a Categorical Target Variable **(supervised)**

Regression
Predicting a Numeric Target Variable **(supervised)**

Association
Unsupervised process for finding relationships between Items

Clustering
Unsupervised process for finding meaningful groups in Data

Data Mining
Taxonomy of Algorithms



K-NN
Decision Trees
Rule Induction
Naïve Bayes
ANN + DeepLearning
SVM + FLM
Ensemble – RF + GBT

Linear Reg.
GLM
Logistic Reg.

K-means
X-Means
K-Medoids
DBSCAN
SOMs
Anomaly-Distance
Anomaly-Density
Anomaly-Local Outlier factor

Apriori
FP Growth
Collaborative Filtering
Content Based Filtering
Graph

Association Rules
Customers who bought this also bought…

- measure the **strength of co-occurrence** between one item with another
- find **usable patterns** in the co-occurrence of the items.
- widely used in retail analysis of transactions, recommendation engines and online clickstream analysis across pages.

Association Rules
Customers who bought this also bought…

A popular application of the technique is **“market basket analysis”** which finds co-occurrences of one retail item with another item within the same retail purchase transaction. A retailer can take advantage of this association for bundle pricing, product placement, and even shelf optimisation within the store layout.

One of the objectives in managing **e-commerce business** is to increase the average order value of the site visit especially when the business has to pay for acquisition traffic through search engine marketing, online advertisements and other marketing. Cross-selling and upselling is important although not at the risk of irritating the customer.

Association Rules
Customers who bought this also bought…

The key input is the **list of past transactions** with product information.

From this we can determine the **most frequent product pairs above a significance threshold**. The result is a rule that says,

“if product A is purchased, there is an increased likelihood that product B will be purchased”…

Association Rules
Key Concepts cont.

Basic association analysis just deals with the occurrence of one item with another. More complicated can take into account quantity of occurrence, price and sequence of occurrence.

Looking at basic association analysis for a media website like BBC or Yahoo news with categories such as news, politics, finance, entertainment and arts. In this case, a transaction or session is one visit for the website where the user can access different categories (products) within a session period. A session period starts after 30 minutes of inactivity.

For association analysis the data must be prepared in Clickstream format for analysis.

Association Rules
Key Concepts cont.

Pivoting the table is required and for simple association the binary format indicating presence or absence of article categories is sufficient.

Table 6.1

Session ID	List of media categories accessed
1	{News, Finance}
2	{News, Finance}
3	{Sports, Finance, News}
4	{Arts}
5	{Sports, News, Finance}
6	{News, Arts, Entertainment}

Clickstream Data Set

Session ID	News	Finance	Entertainment	Sports	Arts
1	1	1	0	0	0
2	1	1	0	0	0
3	1	1	0	1	0
4	0	0	0	0	1
5	1	1	0	1	0
6	1	0	1	0	1

Association Rules
Key Concepts : Item Sets

Association rules are made up of an antecedent and a consequent. Either the antecedent or the consequent may be made up of more than a single disjoint item. Introduction of item sets with more than one item greatly increases the permutations of rules to be considered.

Example : {News, Finance} → {Sports}

Based on historical transactions, (sessions), this rule implies that, if users have accessed news and finance in the same session, there is a high likelihood that they would also access sports articles.

The combination of news and finance is called an **item set.**

The strength of an association rule is commonly quantified by the **support** and **confidence** (Lift and conviction are also sometimes used)

Association Rules
Key Concepts : Support of a Rule

The **support of a rule** is a measure of how all the items in a rule are represented in overall transactions.

Example : {News} → {Sports} = 2/6 = 0.33

Table 6.1

Session ID	List of media categories accessed
1	{News, Finance}
2	{News, Finance}
3	{Sports, Finance, News}
4	{Arts}
5	{Sports, News, Finance}
6	{News, Arts, Entertainment}

Therefore the support measure for a rule indicates whether the rule is worth considering. This is particularly interesting for businesses because leveraging patterns in high volume items leads to more incremental revenue. Rules with low support have either infrequently occurring items or an item relationship that occurs just by chance.

In association analysis, a threshold of support is specified to filter out infrequent rules. Only rules exceeding the threshold are considered for further analysis.

Association Rules
Key Concepts : Confidence of a Rule

The **confidence of a rule** measures the likelihood of occurrence of the consequent of the rule out of all the transactions that contain the antecedent of the rule. of how all the items in a rule are represented in overall transactions. Confidence of the rule (X → Y) is calculated by

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

Example :

Confidence {News, Finance} → {Sport} = $\frac{Support(\{News, Finance, Sports\})}{Support(\{News, Finance\})}$ **=** $\frac{2/6}{4/6}$ **= 0.5**

Confidence provides the reliability measure of the rule. In the case of the rule {News, Finance} → {Sports}, the question that the confidence measure answers is, if a transaction has both News and Finance, what is the likelihood of seeing Sports in it? Half of the transactions that contain News and Finance also contain Sports. This means 50% of users who visit the news and finance pages also visit sports pages.

Association Rules
Key Concepts : Process of Rule Generation

There are two basic tasks in generating meaningful association rules :

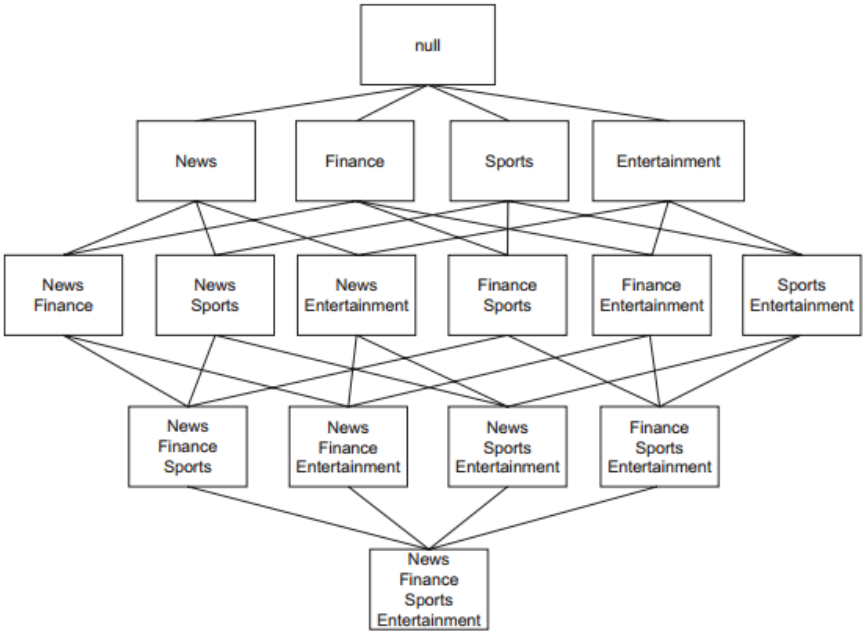
1. Finding all frequent item sets (for n items 2^n - 1 possible excluding null)
2. Extracting rules from frequent item sets (for n items 3^n - 2^{n+1} - 1)

As this two step process generates hundreds of rules even for a small data set with dozens of items, it is computationally less expensive to find all the frequent item sets above a support threshold. It is not uncommon to exclude items (e.g. grocery bag in supermarket example) so that the analysis can focus on a subset of important relevant items.

There are also algorithmic approaches such as Apriori and FP-Growth to efficiently find the frequent item sets from the universe of all the possible item sets.

Association Rules
Process of Rule Generation : Item Set lattice

All possible item sets shown as a lattice for the media example with the item Arts



Item set tree.

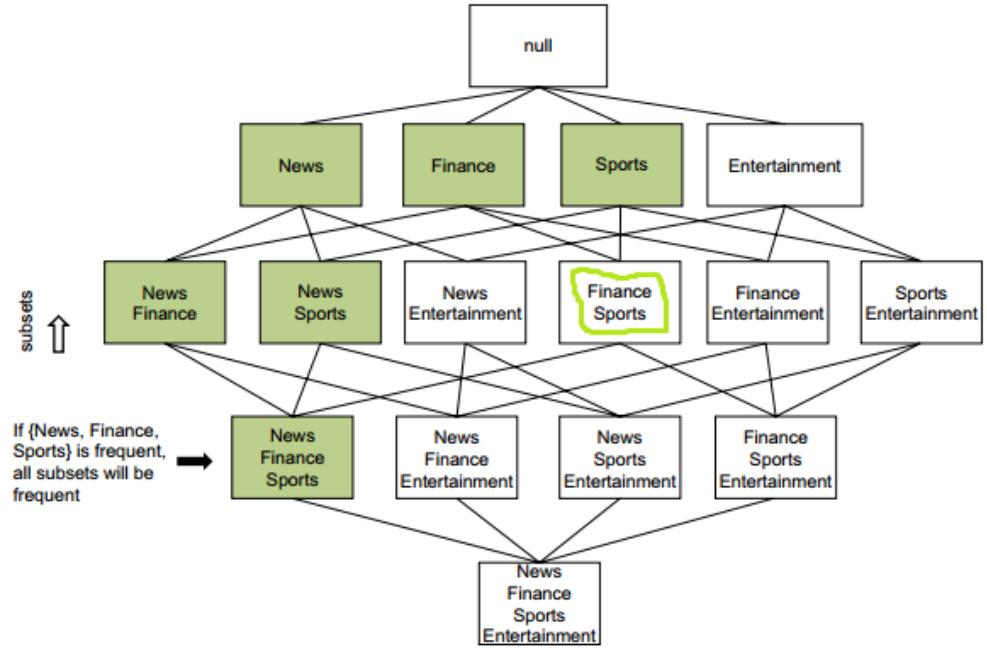
Association Rules
Apriori Algorithm

The Apriori Algorithm leverages some simple logical principles to reduce the number of item sets to be tested for the support measure.

The Apriori principle sates that **"if an item set is frequent, then all its subset items will be frequent". (**Conversely if an item set is infrequent, then all its supersets will be infrequent). The Apriori principle is helpful because not all item sets have to be considered for the support threshold.

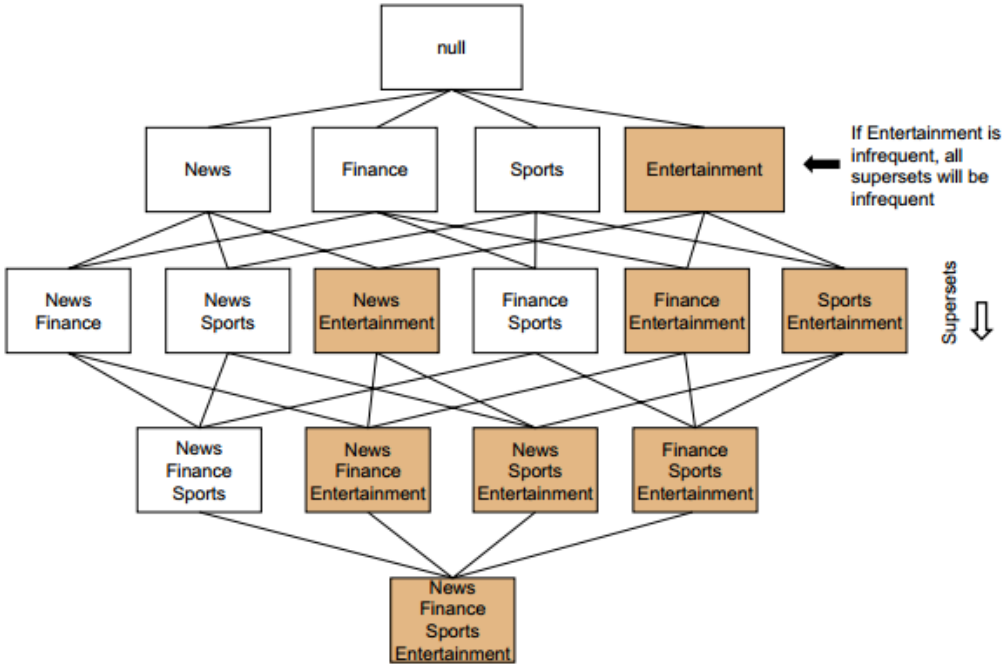
Example (lattice): If {News, Finance, Sports} is a frequent item set that is its support measure is (0.33) which is greater than the support threshold of (0.25), then all of its subset items will be frequent that is,
Support{News, Finance, Sports} = 0.33 (above threshold support)
Support{News, Finance} = 0.66
Support{News, Sports} = 0.33
Support{News} = 0.83
Support{Sports} = 0.33
Support{Finance} = 0.66

Association Rules
Apriori Algorithm Lattice Example 1



Frequent item sets using Apriori principle.

Association Rules
Apriori Algorithm Lattice Example 2



Frequent item sets using Apriori principle: Exclusion.

Association Rules
Apriori Algorithm Frequent Item Set Generation

Support threshold of 0.25 that is items should appear in at least 2 of 6 sessions

Clickstream Data Set: Condensed Version

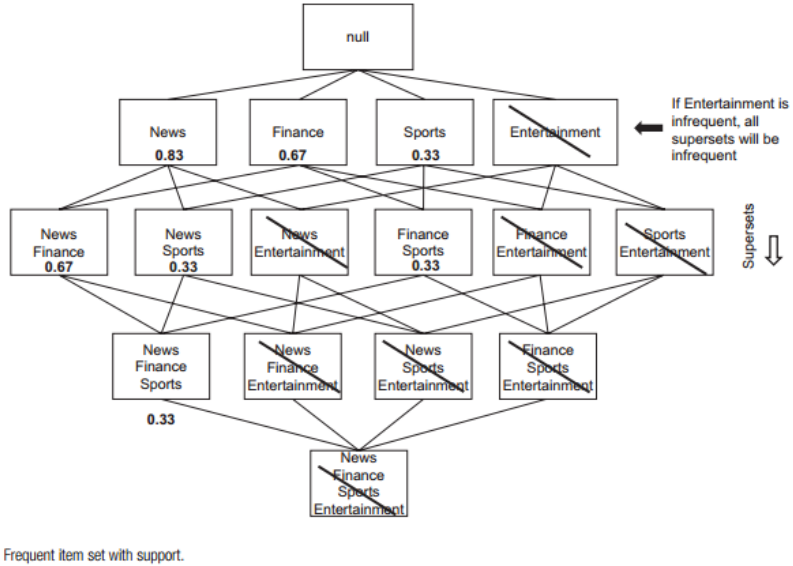
Session	News	Finance	Entertainment	Sports
1	1	1	0	0
2	1	1	0	0
3	1	1	0	1
4	0	0	0	0
5	1	1	0	1
6	1	0	1	0

Table 6.4 Frequent Item Set Support Calculation

Item	Support Count	Support
{News}	5	0.83
{Finance}	4	0.67
{Entertainment}	1	0.17
{Sports}	2	0.33
Two-Item Sets	**Support Count**	**Support**
{News, Finance}	4	0.67
{News, Sports}	2	0.33
{Finance, Sports}	2	0.33
Three-Item Sets	**Support Count**	**Support**
{News, Finance, Sports}	2	0.33

Association Rules
Apriori Algorithm Frequent Item Set Generation cont.

The process continues until all n-item sets have been considered from previous steps. At the end there are 7 out of possible 15 (2^n - 1). By eliminating Entertainment in step 1, 7 item sets did not have to be generated as they would not pass the support threshold (Apriori principle).



Frequent item set with support.

Association Rules
Apriori Algorithm Rule Generation

Now generating rules from frequent item sets with

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

$\{\text{News, Sports}\} \rightarrow \{\text{Finance}\} = 0.33/0.33 = 1$
$\{\text{News, Finance}\} \rightarrow \{\text{Sports}\} = 0.33/0.67 = 0.5$
$\{\text{Sports, Finance}\} \rightarrow \{\text{News}\} = 0.33/0.33 = 1$
$\{\text{News}\} \rightarrow \{\text{Sport, Finance}\} = 0.33/0.83 = 0.4$
$\{\text{Sports}\} \rightarrow \{\text{News, Finance}\} = 0.33/0.33 = 1.0$
$\{\text{Finance}\} \rightarrow \{\text{News, Sports}\} = 0.33/0.67 = 0.5$

Association Rules
Apriori Algorithm Rule Generation

Since all the support scores have already been calculated in the item set generation step, there is no need for another set of computations for calculating confidence.

However, it is possible to prune low confidence scores using the same Apriori principle for example excluding $\{\text{News, Finance}\} \rightarrow \{\text{Sports}\} = 0.33/0.67 = 0.5$ will mean exclusion of $\{\text{News}\} \rightarrow \{\text{Sport, Finance}\}$ and $\{\text{Finance}\} \rightarrow \{\text{News, Sports}\}$.

All the rules passing a particular confidence threshold are considered for output along with both support and confidence measures. These rules should be further evaluated for rational validity to determine if a useful relationship was uncovered, if there was an occurrence by chance, or if the rule confirms a known intuitive relationship.

Association Rules
Summary of Concepts

Frequent Item Set

Frequent patterns are patterns (e.g. item sets, sub-sequences, or substructures) that occur frequently in a dataset. A set of items, such as milk and bread, that appear frequently together in a transaction data set is a Frequent Item Set.

Support of an Item

The relative frequency of an occurrence of an item set in the transaction set.

Support of a Rule

A measure of how all the items in a rule are represented in overall transactions. The support measure for a rule indicates whether a rule is worth considering.

Association Rules
Summary of Concepts

Confidence of a Rule

Measures the likelihood of occurrence of the consequent of the rule out of all the transactions that contain the antecedent of the rule. Confidence provides the reliability measure of the rule.

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

Association Rules
Summary of Concepts cont.

Lift of a Rule

Lift is the ratio of observed support with what is expected if antecedent and consequent were completely independent. Lift values closer to 1 mean the antecedent and consequent of the rules are independent and the rule is not interesting. The higher (above 1) the value of lift, the more interesting the rules are.

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

Association Rules
Summary of Concepts cont.

Conviction of a Rule

The Conviction of the rule $\boldsymbol{X \rightarrow Y}$ is the ratio of the expected frequency of $\boldsymbol{X}$ occurring in spite of $\boldsymbol{Y}$ and the observed frequency of incorrect predictions. Conviction takes into account the direction of the rule. The conviction of $\boldsymbol{(X \rightarrow Y)}$ is not the same as the conviction of $\boldsymbol{(Y \rightarrow X)}$.

$$\boldsymbol{Conviction(X \rightarrow Y) = \frac{1 - Support(Y)}{1 - Confidence(X \rightarrow Y)}}$$

For example {milk, bread} -> {butter} Conviction = 1.2 is interpreted as the rule will be correct 20% more often than if by random chance.

Association Rules
FP-Growth

- The Frequent Pattern (FP)-Growth algorithm uses a special graph data structure called FP-Tree.

- An FP-Tree can be thought of as a transformation of the data set into graph format. Rather than the generate and test approach used in Apriori algorithm, FP-Growth first generates the FP-Tree and uses this compressed tree to generate the frequent item sets.

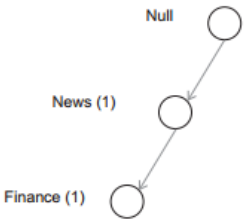
- The efficiency of the FP-Growth algorithm depends on how much compression can be achieved in generating the FP-Tree.

Association Rules
FP Growth : Generating the FP-Tree

Transactions List: Session and Items

Session	Items
1	{News, Finance}
2	{News, Finance}
3	{News, Finance, Sports}
4	{Sports}
5	{News, Finance, Sports}
6	{News, Entertainment}

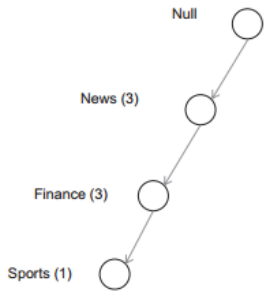
- Sort items in each transaction in descending order of frequency for example {Sports, News, Finance} is re-ordered to {News, Finance, Sports}
- Map the transactions to the FP-Tree starting with a null node



FP-Tree: Transaction 1.

Association Rules
FP Growth : Generating the FP-Tree cont.

- Since the second transaction {News, Finance} is the same as the first, it follows the same path and we can just increment the count
- For the third transaction the tree is extended to Sports and the item path count is incremented

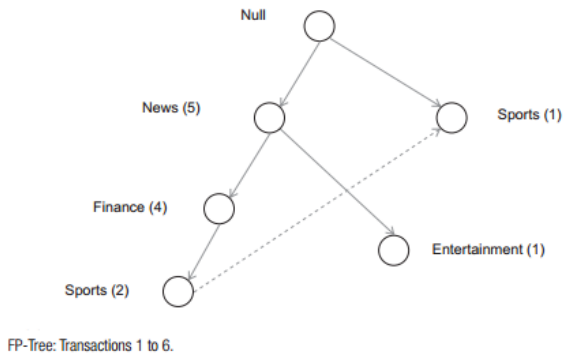


FP-Tree: Transactions 1, 2, and 3

- The fourth transaction only contains the {Sports} item. Since Sports is not preceded by News and Finance, a new path should be created from the null item and the item count noted. It should be joined to the other occurrence of Sports by a dotted line

Association Rules
FP Growth : Generating the FP-Tree cont.

- This process is continued until all the transactions are scanned. All of the transactions can now be represented by a compact FP-Tree



FP-Tree: Transactions 1 to 6.

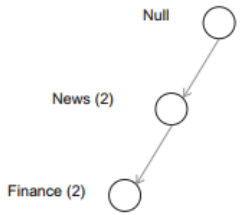
- NOTE : The compression of the FP-Tree depends on how frequently a path occurs within a given transaction set

Association Rules
FP-Tree : Frequent Pattern Set Generation

- Once the transaction set is expressed as a compact FP-Tree, the most frequent item set can be generated from the tree effectively. To generate the frequent item set, the FP-Growth algorithm adopts a bottoms-up approach of generating all the item sets starting with the least frequent items. Since the structure of the tree is ordered by the support count, the least frequent items can be found in the leaves of the tree.

- Finding the entire item set ending with a particular item is actually made possible by generating a prefix path and conditional FP-Tree for an item.

Association Rules
FP-Tree : Frequent Pattern Set Generation



Conditional FP-Tree.

- The prefix path of an item is a subtree with only paths that contain the item of interest. A Based on the conditional FP-Tree, the algorithm repeats the process of finding leaf nodes.

Association Rules
Discussion Points

Advantages

- Only 2 passes over data set
- Compresses the data set
- No candidate generation required
- Much faster than Apriori

Disadvantages

- FP-Tree may not fit in memory
- FP-Tree is expensive to build, however, once built, frequent item can easily be read off from the tree
- Support can only be calculated once entire data set has been added to the FP-Tree

Association Rules Discussion Points

- Association rules have gained popularity over last two decades particularly in retail, online cross-selling, recommendation engines, text analysis, document analysis, and web analysis
- Typically commercial data mining software tools will include association analysis although there may be variations in implementations
- Applications with very large numbers of items and real-time decision making demand new approaches to efficient and scalable association analysis
- Association analysis is one of the prevalent algorithms applied to information stored using big data technologies, data streams, and large databases

Association Rules
FP Growth and Apriori : Summary

Model
Finds simple easy to understand rules like {Milk, Diaper} → {Beer}

Input
Transactions format with items in the columns and transactions in the rows

Output
List of relevant rules developed from the data set

Pros
Unsupervised approach with minimal user inputs. Easy to understand rules

Cons
Requires pre-processing if input is of a different format

Use Cases
Recommendation engines, cross-selling, and content suggestions

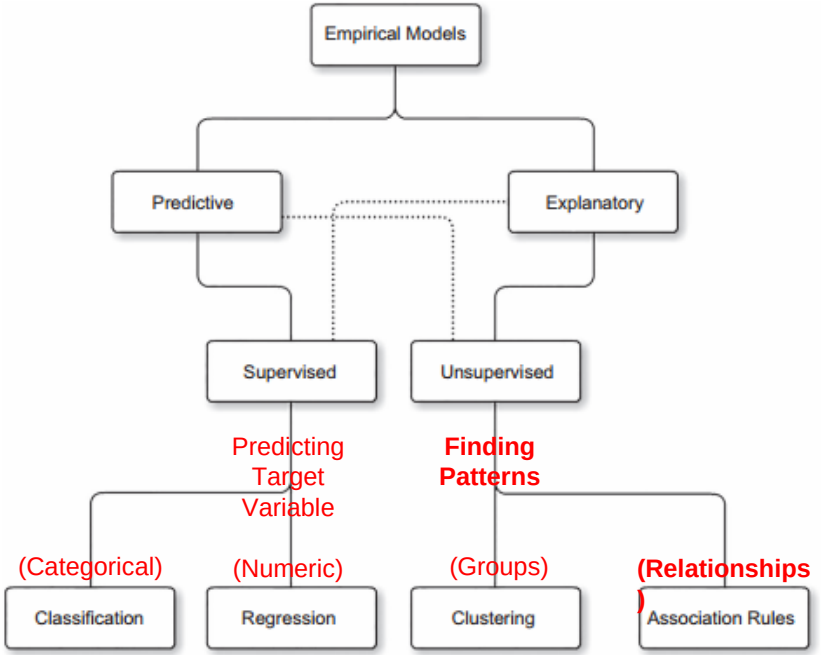
Association Rules
Use Cases : Cross Selling

Consider an e-commerce website that sells a large selection of products online. One of the objectives in managing e-commerce business is to increase the average order value of the visit. Optimizing order size is even more critical when the businesses pay for acquisition traffic through search engine marketing, online advertisements, and affiliate marketing. Businesses attempt to increase average order value by cross-selling and up-selling relevant products to the customer, many times based on what they have purchased or are currently purchasing in the current transaction (a common fast-food equivalent: "Do you want fries with the burger?"). Businesses need to be careful by weighing the benefit of suggesting an extremely relevant product against the risk of irritating a customer who is already making a transaction. In a business where there are limited products (e.g., fast-food industry), cross-selling a product with another product is straightforward and is quite inherent in the business. But, when the number of unique products runs in thousands and millions, determining a set of *affinity products* when customers are looking at a product is quite a tricky problem.

To better learn about product affinity, we turn to purchase history data. The information on how one product creates affinity to another product relies on the fact that both the products appear in the same transaction. If two products are bought together, then we can speculate that the necessity of those products arise in the same time frame for the customer. If the two products are bought together many times, by a large number of customers, then there is definitely an affinity pattern within these products. In a new later transaction, if a customer picks one of those affinity products, then there is an increased likelihood that the other product will be picked by the customer, in the same transaction.

The key input for affinity analysis is a list of past transactions with product information. Based on the analysis of these transactions, we can determine what the most frequent product pairs are. We need to define a threshold for "frequent" because a few appearances of a product pair doesn't qualify as a pattern. The result of the affinity analysis is a rule set that says, "If product A is purchased, there is an increased likelihood that product B will be purchased in the same transaction." This rule set can be leveraged to provide cross sell recommendations on the product page of product A. Affinity analysis is the concept behind the web widgets which state, "Customers who bought this also bought..."

Association Rules
Recap on Taxonomy



References

- Agrawal, R., Imielinski, ´ T., & Swami, A. (1993). *Mining association rules between sets of items in large databases*. In SIGMOD ‘93 proceedings of the 1993 ACM SIGMOD international conference on management of data (pp. 207216).
- Agrawal, R., & Srikant, R. (1994). *Fast algorithms for mining association rules*. The international conference on very large databases, 487499.
- Akbar, M., & Angryk, R. (2008). *Frequent pattern-growth approach for document organization*. In Proceeding of the 2nd international workshop on Ontologies and information systems for the semantic web, ACM (pp. 7782). Available from http://dl.acm.org/citation.cfm?id 5 1458496.
- Bodon, F. (2005). *A tree-based APRIORI implementation for mining frequent item sequences*. In Proceedings of the 1st international workshop on open-source data science frequent pattern mining implementations OSDM ‘05 (pp. 5665). http://dx.doi.org/10.1145/1133905.1133913.

References

- Han, J., Pei, J., & Yin, Y. (2000). *Mining frequent patterns without candidate generation*. In SIGMOD ‘00 proceedings of the 2000 ACM SIGMOD international conference on management of data (pp. 112).
- Shang, X., Sattler, K.U., Geist, I. (2004). *SQL based frequent pattern mining without candidate generation*. In 2004 ACM symposium on applied computing Poster Abstract (pp. 618619).
- Tan, P.-N., Steinbach, M., & Kumar, V. (2005). *Association analysis: Basic concepts and algorithms*. Introduction to data mining (pp. 327404). Boston, MA: Addison Wesley.
- Tanbeer, S.K., Ahmed, C.F., Jeong, B.-S., Lee, Y.-K. (2008). Efficient frequent pattern mining over data streams. In Proceeding of the 17th ACM conference on information and knowledge mining CIKM ‘08 (Vol. 1, pp. 14471448). http://dx.doi.org/10.1145/1458082.1458326.

References

- Witten, I. H., & Frank, E. (2005). *Algorithms: The basic methods: Mining association rules*. Data science: Practical machine learning tools and techniques (pp. 112118). San Francisco, CA: Morgan Kaufmann.
- Zaki, M. Jk (2000). *Scalable algorithms for association mining*. IEEE Transactions on Knowledge and Data Engineering, 12(3), 372390. Available from https://doi.org/10.1109/69.846291.