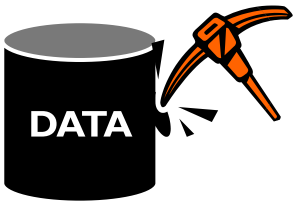
National College of Ireland



Data & Web Mining

6. Association Rules

Dr. Jason Roche
Jason.roche@ncirl.ie

6. Association Rules

6.1 Introduction

- Introduction
  - *"Think back to the last time you made an impulse purchase. Maybe you were waiting in the grocery store checkout lane and bought a pack of chewing gum or a candy bar. Perhaps on a late-night trip to a convenience store for diapers and formula you picked up a caffeinated beverage or a six-pack of beer… In any case, it is no coincidence that gum and candy are located in checkout lanes, convenience stores stock beer in addition to diapers..."*
  - In years past, recommendations were based on the subjective experience of marketing professionals and inventory managers or buyers.
  - More recently, machine learning has been used to learn these patterns of purchasing behavior.
  - Barcode scanners, computerized inventory systems, and online shopping have led to a wealth of transactional data ripe for such data mining

6.1 Introduction

- What we will look at…
  - Methods for finding useful associations in large databases using simple statistical performance measures.
  - How to manage the peculiarities of working with transactional data.
  - The start-to-finish steps needed for using association rules to perform a market basket analysis on real-world data.

6.2 Association Rules

- The result of a market basket analysis is a set of **association rules** that specify patterns of relationships among items. A typical rule might be expressed in the form:

  {peanut butter, jelly} → {bread}

- This association rule states that if peanut butter and jelly are purchased, then bread is also likely to be purchased.
  - In other words, "peanut butter and jelly imply bread."

- Groups of one or more items are surrounded by brackets to indicate that they form a set, or more specifically, an **itemset** that appears in the data with some regularity.

- Association rules are learned from subsets of itemsets:
  - e.g., the preceding rule was identified from the set of {*peanut butter, jelly, bread*}.

6.2 Association Rules

- Developed in the context of Big Data and database science, association rules are not used for prediction, but rather for *unsupervised* knowledge discovery in large databases, unlike the classification algorithms we have seen so far.

- Because association rule learners are unsupervised, there is no need for the algorithm to be trained; data does not need to be labeled ahead of time.

- The program is simply unleashed on a dataset in the hope that interesting associations are found.

- The downside, of course, is that there isn't an easy way to objectively measure the performance of a rule learner:
  - aside from evaluating them for qualitative usefulness—typically an eyeball test of some sort.

6.2 Association Rules

Beyond the Basket

- Although association rules are most often used for market basket analysis, they are helpful for finding patterns in many different types of data.

- Other potential applications include:
  - Searching for interesting and frequently occurring patterns of DNA and protein sequences in an analysis of cancer data.
  - Finding patterns of purchases or medical claims that occur in combination with fraudulent credit card or insurance use
  - Identifying combinations of behavior that proceed customers dropping their cellular phone service or upgrading their cable television package

National College of Ireland

6.2 Association Rules

- Utility
  - Association rule analysis is used to search for interesting connections among a very large number of variables.

  - Human beings are capable of such insight quite intuitively, but it often takes expert-level knowledge or a great deal of experience to do what a rule-learning algorithm can do in minutes or even seconds.
  - Additionally, some data is simply too large and complex for a human being to find the needle in the haystack.
  - We use algorithms that use heuristics to reduce the potential search space.
    - Apriori Algorithm
  - But lets start with a simple example first… before thinking about what processing real-world examples would entail.

6.2 Association Rules

- Example

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

6.2 Association Rules

- Frequent Itemsets

  - Given a set of transactions **D**, find combination of items that occur frequently:

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

6.2 Association Rules

- Definitions

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

- Itemset
  - A set of one or more items
  - e.g.: **{Milk, Bread, Diaper}**
- k-itemset
  - An itemset that contains k items
- Support count ()
  - Frequency of occurrence of an itemset (number of transactions it appears)
  - e.g. **({Milk, Bread,Diaper}) = 2**
- Support
  - Fraction of the transactions in which an itemset appears
  - e.g. **s({Milk, Bread, Diaper}) = 2/5**
- Frequent Itemset
  - An itemset whose support is greater than or equal to a *minsup* threshold
  - Minsup = minimum support level: supplied by user!

6.2 Association Rules

- Frequent Itemsets

  - Why do we want to find frequent itemsets?
    - Find all combinations of items that occur together
    - They might be interesting (e.g., in placement of items in a store…)

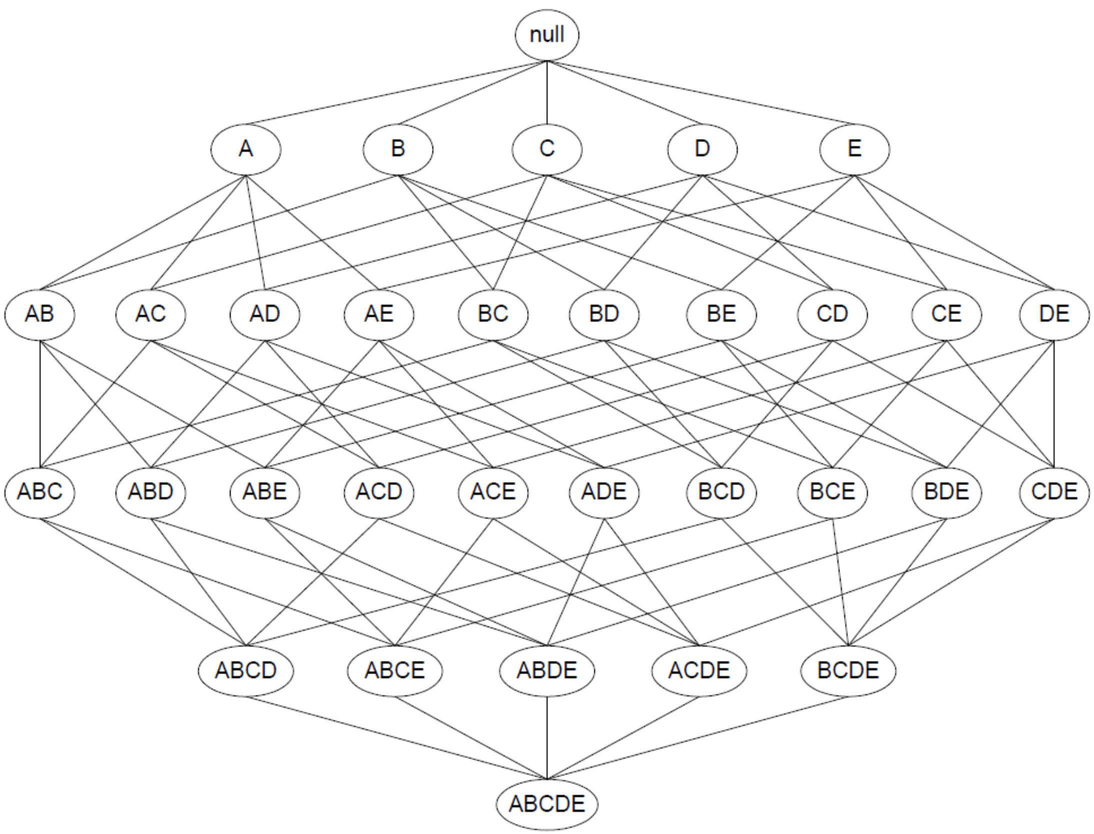
  - Frequent itemsets are only positive combinations (we do not report combinations that do not occur frequently together)

  - Frequent itemsets aims at providing a summary for the data

  - Given a transaction database **D** and a **minsup** threshold to find all frequent itemsets and the frequency of each set in this collection:
    - Count the number of times combinations of attributes occur in the data. If the count of a combination is above minsup report it.

6.2 Association Rules

- Can get very big, very quickly!
  - Given d items, there are 2^d possible itemsets.



6.2 Association Rules

- Frequent Itemsets
  - If **minsup= 0**, then all subsets of ***I*** will be frequent and thus the size of the collection will be very large
    - This summary is very large (maybe larger than the original input) and thus not interesting
  - The task of finding all frequent sets is interesting typically only for relatively large values of **minsup.**
  - **What about association rules?**
    - Association Rule
      - An implication expression of the form X → Y, where X and Y are itemsets
      - Example: {Milk, Diaper} → {Beer}

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

6.2 Association Rules

- {Milk, Diaper} → {Beer}
  - X → Y
- Support = Fraction of transactions that contain both X and Y
  - 2 / 5 = 0.4
- Confidence = Measures how often items in Y appear in transactions that contain X
  - 2/3 = 0.67
- {Milk, Diaper} → {Beer}, s=0.4, c=0.67

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

We don't want really want to create a huge number of candidates that aren't really meaningful…

We have a **minsup** and we also have a **minconf** threshold.

6.2 Association Rules

- Examples
  - Some candidate rules:

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 Rules:

{Milk,Diaper} $\rightarrow$ {Beer} (s=0.4, c=0.67)
{Milk,Beer} $\rightarrow$ {Diaper} (s=0.4, c=1.0)
{Diaper,Beer} $\rightarrow$ {Milk} (s=0.4, c=0.67)
{Beer} $\rightarrow$ {Milk,Diaper} (s=0.4, c=0.67)
{Diaper} $\rightarrow$ {Milk,Beer} (s=0.4, c=0.5)
{Milk} $\rightarrow$ {Diaper,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

6.2 Association Rules

- A Quick Check

  - {Milk,Diaper} → {Beer} (s=0.4, c=0.67)
    - We did this one initially…
  - {Milk,Beer} → {Diaper} (s=0.4, c=1.0)
    - i.e., s = 2/5, c = 2/2
  - Can you see where we get '0.4' and '1.0' from?

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

6.3 Mining Association Rules

Mining Association Rules

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

- How do we process the data?
  - 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
  - Problem: Computationally prohibitive!

- So lets introduce a widely used approach –
  - the **Apriori Algorithm**…

6.3 Mining Association Rules

Strong Rules

- **Strong rules** have both high support and confidence.

- The Apriori Algorithm uses minimum levels of support and confidence with the Apriori principle to quickly find strong rules by reducing the number of rules to a more manageable level.

- Basic principle:
  - *the Apriori principle states that all subsets of a frequent itemset must also be frequent.*
  - *In other words, if {A, B} is frequent, then {A} and {B} both must be frequent.*
  - *Recall also that by definition, the support metric indicates how frequently an itemset appears in the data.*
  - *Therefore, if we know that {A} does not meet a desired support threshold, there is no reason to consider {A, B} or any itemset containing {A}; it cannot possibly be frequent.*

6.3 Mining Association Rules

- Apriori Stages

  - The Apriori Algorithm uses this logic to exclude potential association rules prior to actually evaluating them.
  - The actual process of creating rules occurs in two phases:
    - Identifying all itemsets that meet a minimum support threshold.
    - Creating rules from these itemsets that meet a minimum confidence threshold.

  - The first phase occurs in multiple iterations.
    - Each successive iteration involves evaluating the support of storing a set of increasingly large itemsets. For instance:
      - iteration 1 involves evaluating the set of 1-item itemsets (1-itemsets),
      - iteration 2 evaluates the 2-itemsets, and so on.
      - The result of each iteration *i* is a set of all *i*-itemsets that meet the minimum support threshold.

6.3 Mining Association Rules

Apriori Algorithm

- All the itemsets from iteration i are combined in order to generate candidate itemsets for evaluation in iteration i + 1.
  - The Apriori principle can eliminate some of these before the next round.
  - If {A}, {B}, and {C} are frequent in iteration 1 while {D} is not frequent, then iteration 2 will consider only {A, B}, {A, C}, and {B, C}.
  - Thus, the algorithm needs to evaluate only three itemsets rather than six.

- Suppose in iteration 2 {A, B} and {B, C} are frequent, but not {A, C}.
  - Although iteration 3 would normally begin by evaluating the support for {A, B, C}, this step need not occur at all.
  - Why? The Apriori principle states that {A, B, C} cannot be frequent if the subset {A, C} is not.
  - Having generated no new itemsets, the algorithm may stop.

- At this point, the second phase of the Apriori algorithm may begin.
  - Given the set of frequent itemsets, association rules are generated from all possible subsets.
  - For instance, {A, B} would result in candidate rules for {A} → {B} and {B} → {A}.
  - These are evaluated against a minimum confidence threshold, and any rules that do not meet the desired confidence level are eliminated.

6.4 Example

- Example in R

  - Market Basket Analysis:

…*market basket analysis is used behind the scenes for the recommendation systems used in many brick-and-mortar and online retailers. The learned association rules indicate combinations of items that are often purchased together in a set. The acquired knowledge might provide insight into new ways for a grocery chain to optimize the inventory, advertise promotions, or organize the physical layout of the store. For instance, if shoppers frequently purchase coffee or orange juice with a breakfast pastry, then it may be possible to increase profit by relocating pastries closer to the coffee and juice.*

6.5 Summary

- Introduction
- Association Rules
- Mining Association Rules
- Example