Module 6 – Advanced Analytics - Theory and Methods Part I

EMC^2 PROVEN PROFESSIONAL

Module 6: Advanced Analytics – Theory and Methods

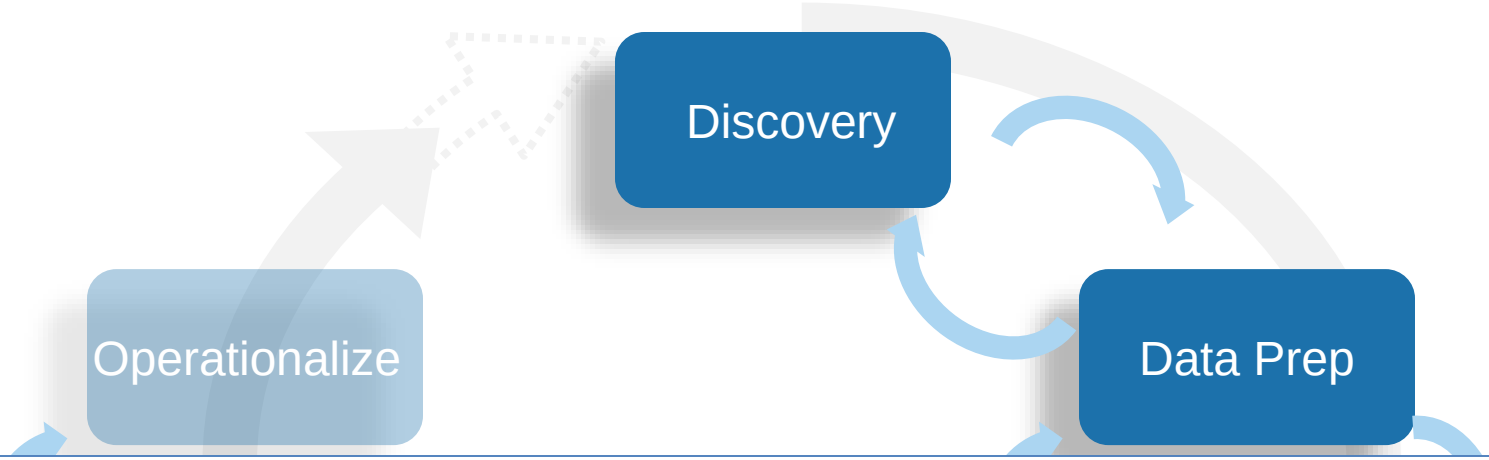
Upon completion of this module, you should be able to:

- Examine analytic needs and select an appropriate technique based on business objectives; initial hypotheses; and the data's structure and volume
- Apply some of the more commonly used methods in Analytics solutions
- Explain the algorithms and the technical foundations for the commonly used methods
- Explain the environment (use case) in which each technique can provide the most value
- Use appropriate diagnostic methods to validate the models created
- Use R and in-database analytical functions to fit, score and evaluate models

Where “R” we?

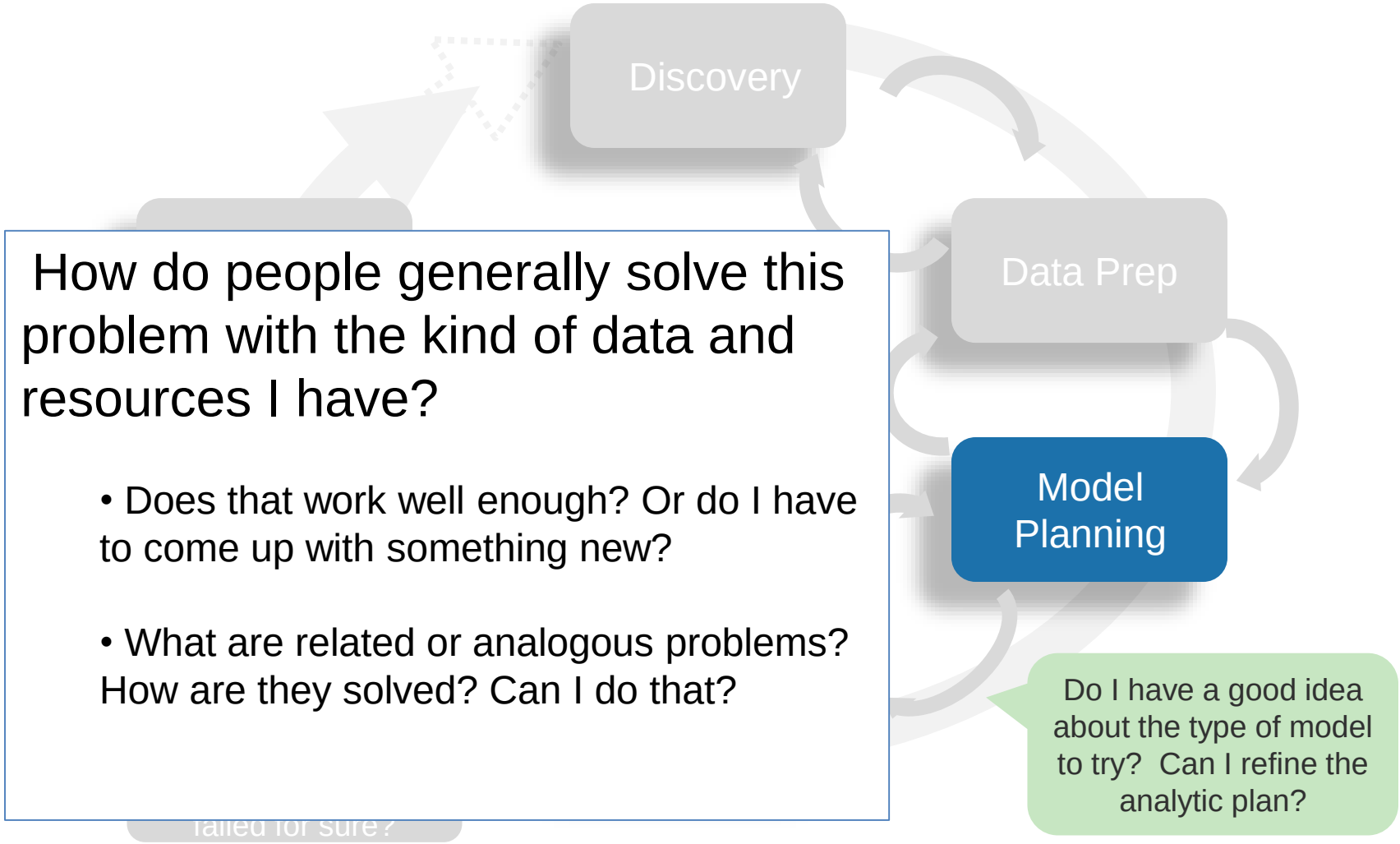
- In Module 3 we reviewed R skills and basic statistics
- You can use R to:
  - Generate summary statistics to investigate a data set
  - Visualize Data
  - Perform statistical tests to analyze data and evaluate models
- Now that you have data, and you can see it, you need to plan the analytic model and determine the analytic method to be used

Applying the Data Analytics Lifecycle



- In a typical Data Analytics Problem - you would have gone through:
  - Phase 1 – Discovery - have the problem framed
  - Phase 2 – Data Preparation - have the data prepared
- Now you need to plan the model and determine the method to be used.

Phase 3 - Model Planning



How do people generally solve this problem with the kind of data and resources I have?

- Does that work well enough? Or do I have to come up with something new?
- What are related or analogous problems? How are they solved? Can I do that?

What Kind of Problem do I Need to Solve? How do I Solve it?

<this module focuses on K-Means and Apriori>

The Problem to Solve	The Category of Techniques	Covered in this Course
I want to group items by similarity. I want to find structure (commonalities) in the data	Clustering	K-means clustering
I want to discover relationships between actions or items	Association Rules	Apriori
I want to determine the relationship between the outcome and the input variables	Regression	Linear Regression Logistic Regression
I want to assign (known) labels to objects	Classification	Naïve Bayes Decision Trees
I want to find the structure in a temporal process I want to forecast the behavior of a temporal process	Time Series Analysis	ACF, PACF, ARIMA
I want to analyze my text data	Text Analysis	Regular expressions, Document representation (Bag of Words), TF-IDF

Why These Example Techniques?

- Most popular, frequently used:
  - Provide the foundation for Data Science skills on which to build
- Relatively easy for new Data Scientists to understand & comprehend
- Applicable to a broad range of problems in several verticals

Module 6: Advanced Analytics – Theory and Methods

Lesson 1: K-means Clustering

During this lesson the following topics are covered:

- Clustering – Unsupervised learning method
- K-means clustering:
  - Use cases
  - The algorithm
  - Determining the optimum value for K
  - Diagnostics to evaluate the effectiveness of the method
  - Reasons to Choose (+) and Cautions (-) of the method

Clustering

How do I group these documents by topic?

How do I group my customers by purchase patterns?

- Sort items into groups by similarity:
  - Items in a cluster are more similar to each other than they are to items in other clusters.
  - Need to detail the properties that characterize “similarity”
    - Or of distance, the "inverse" of similarity
- Not a predictive method; finds similarities, relationships
- Our Example: K-means Clustering

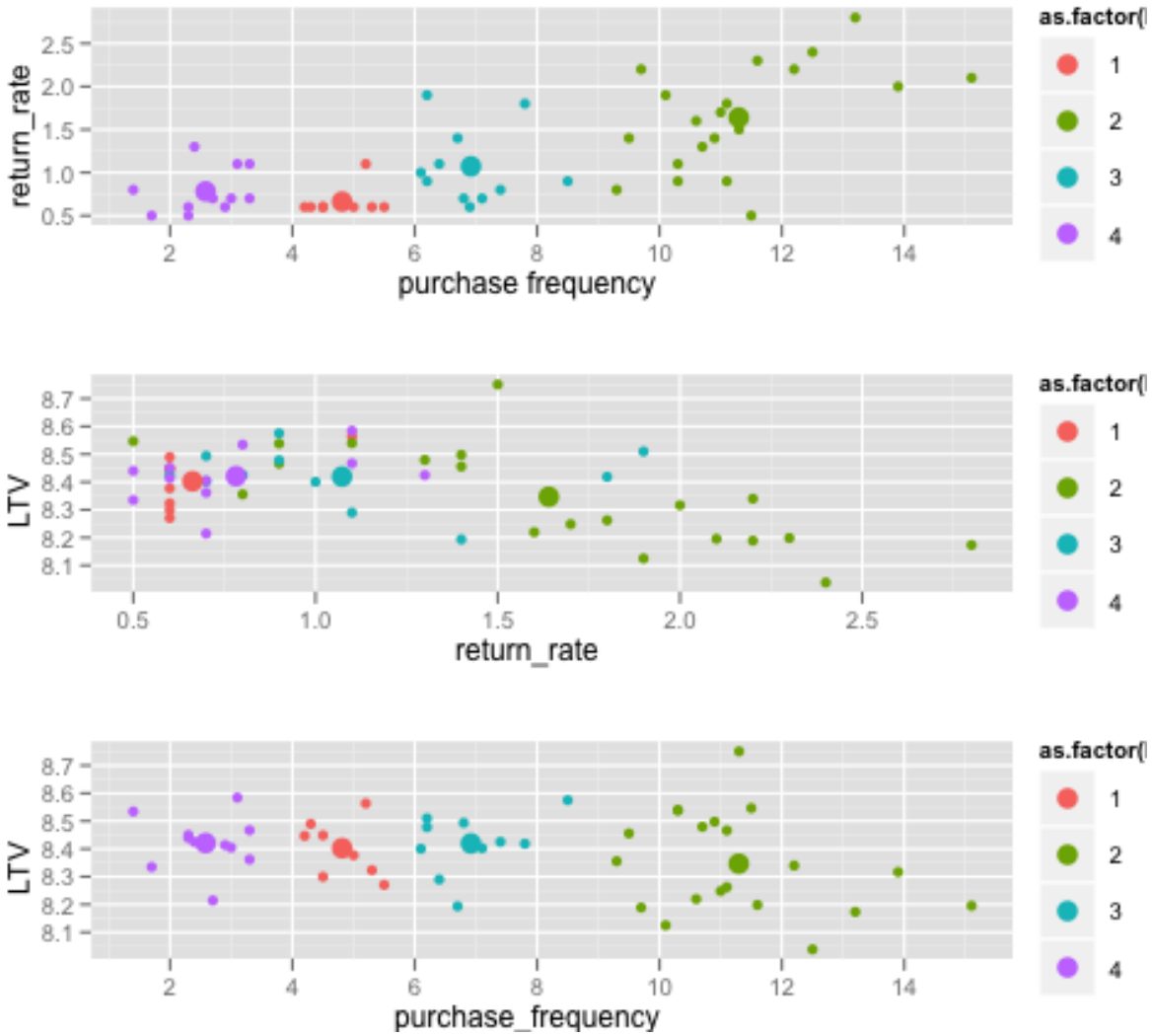
K-Means Clustering - What is it?

- Used for clustering numerical data, usually a set of measurements about objects of interest.
- Input: numerical. There must be a distance metric defined over the variable space.
  - Euclidian distance
- Output: The centers of each discovered cluster, and the assignment of each input datum to a cluster.
  - Centroid

Use Cases

- Often an exploratory technique:
  - Discover structure in the data
  - Summarize the properties of each cluster
- Sometimes a prelude to classification:
  - "Discovering the classes“
- Examples
  - The height, weight and average lifespan of animals
  - Household income, yearly purchase amount in dollars, number of household members of customer households
  - Patient record with measures of BMI, HBA1C, HDL

Use-Case Example – On-line Retailer



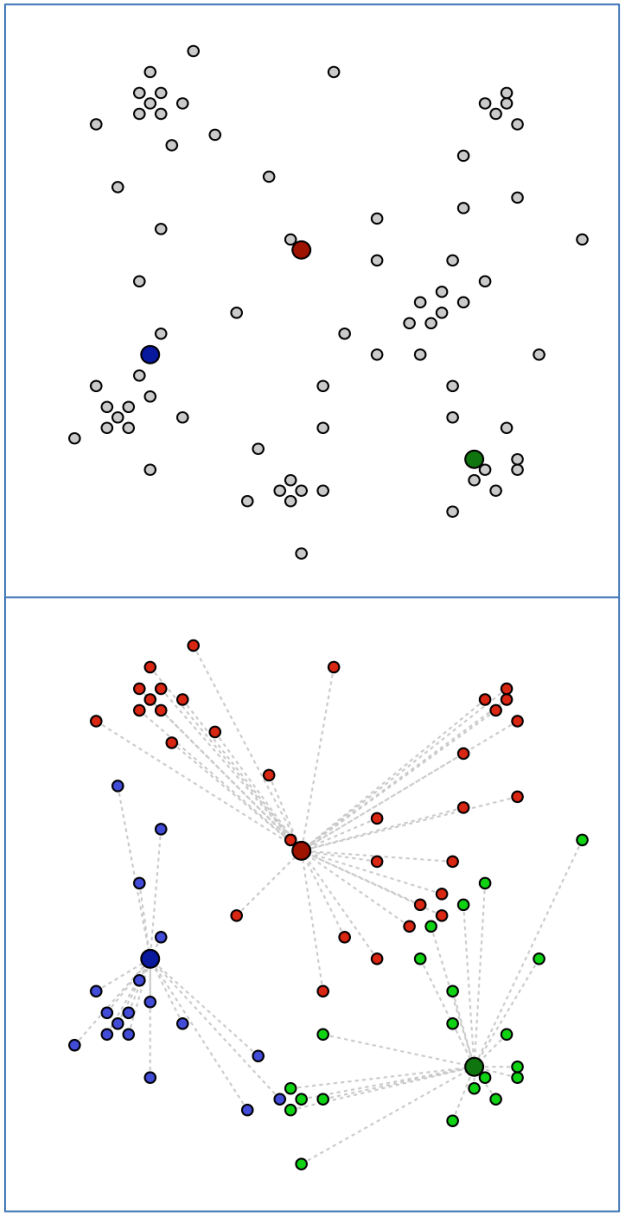
LTV – Lifetime Customer Value

The Algorithm

1. Choose *K*; then select K random "centroids"

   In our example, K=3

2. Assign records to the cluster with the closest centroid



The Algorithm (Continued)

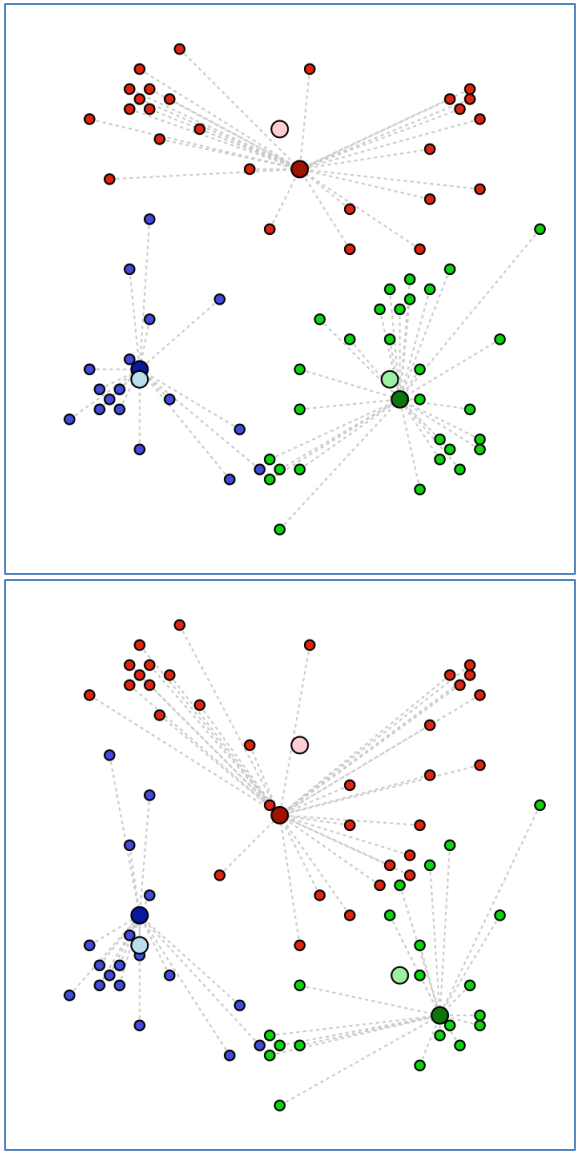
3. Recalculate the resulting centroids

    Centroid: the mean value of all the records in the cluster

4. Repeat steps 2 & 3 until record assignments no longer change

Model Output:

- The final cluster centers
- The final cluster assignments of the training data



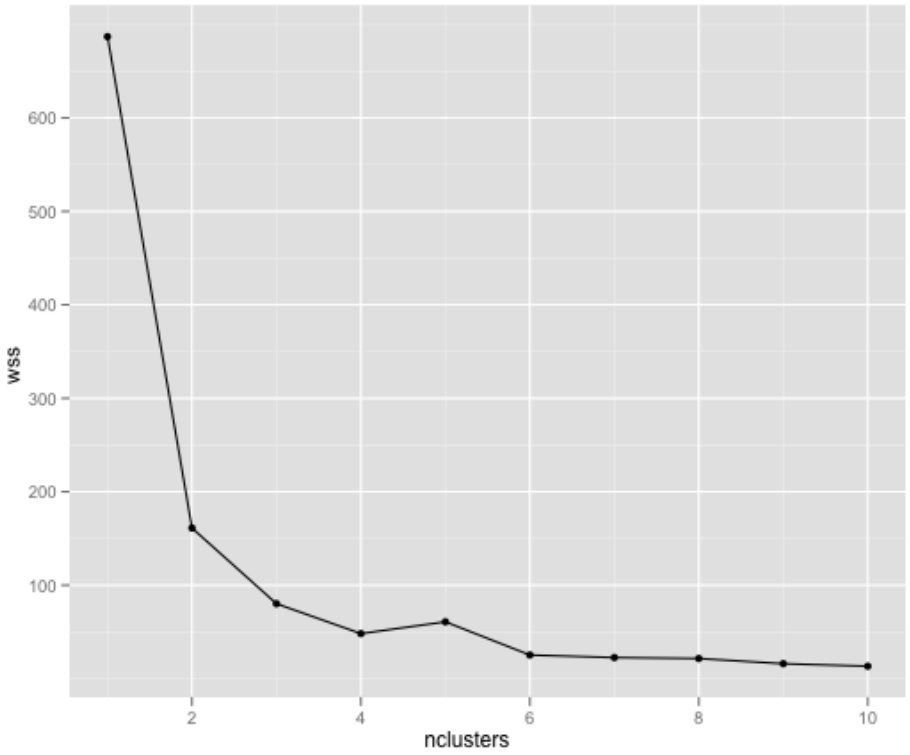
Picking K

Heuristic: find the "elbow" of the within-sum-of-squares (wss) plot as a function of K.

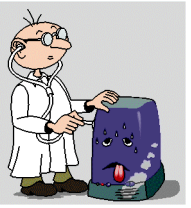
$$\text{wss} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} |x_{ij} - c_i|^2$$

K: # of clusters
n_i: # points in i^{th} cluster
c_i: centroid of i^{th} cluster
x_{ij}: jth point of i^{th} cluster

"Elbows" at k=2,4,6



Diagnostics – Evaluating the Model



- Do the clusters look separated in at least some of the plots when you do pair-wise plots of the clusters?
  - Pair-wise plots can be used when there are not many variables
- Do you have any clusters with few data points?
  - Try decreasing the value of K
- Are there splits on variables that you would expect, but don't see?
  - Try increasing the value K
- Do any of the centroids seem too close to each other?
  - Try decreasing the value of K

K-Means Clustering - Reasons to Choose (+) and Cautions (-)

Reasons to Choose (+)	Cautions (-)
Easy to implement	Doesn't handle categorical variables
Easy to assign new data to existing clusters Which is the nearest cluster center?	Sensitive to initialization (first guess)
Concise output Coordinates the K cluster centers	Variables should all be measured on similar or compatible scales Not scale-invariant!
	K (the number of clusters) must be known or decided a priori Wrong guess: possibly poor results
	Tends to produce "round" equi-sized clusters. Not always desirable

Check Your Knowledge

Your Thoughts?

1. Why do we consider K-means clustering as a unsupervised machine learning algorithm?
2. How do you use “pair-wise” plots to evaluate the effectiveness of the clustering?
3. Detail the four steps in the K-means clustering algorithm.
4. How do we use WSS to pick the value of K?
5. What is the most common measure of distance used with K-means clustering algorithms?
6. The attributes of a data set are “purchase decision (Yes/No), Gender (M/F), income group (<10K, 10-50K, >50K). Can you use K-means to cluster this data set?

Module 6: Advanced Analytics – Theory and Methods

Lesson 2: Association Rules

During this lesson the following topics are covered:

- Association Rules mining
- Apriori Algorithm
- Prominent use cases of Association Rules
- Support and Confidence parameters
- Lift and Leverage
- Diagnostics to evaluate the effectiveness of rules generated
- Reasons to Choose (+) and Cautions (-) of the Apriori algorithm

Association Rules

Which of my products tend to be purchased together?

What do other people like this person tend to like/buy/watch?

- Discover "interesting" relationships among variables in a large database
  - Rules of the form "When X observed, Y also observed"
  - The definition of "interesting" varies with the algorithm used for discovery
- Not a predictive method; finds similarities, relationships

Association Rules - Apriori

- Specifically designed for mining over transactions in databases
- Used over itemsets: sets of discrete variables that are linked:
  - Retail items that are purchased together
  - A set of tasks done in one day
  - A set of links clicked on by one user in a single session
- **Our Example: Apriori**

Apriori Algorithm - What is it?
Support

- Earliest of the association rule algorithms
- Frequent itemset: a set of items L that appears together "often enough“:
  - Formally: meets a minimum support criterion
  - **Support**: the % of transactions that contain L
- Apriori Property: Any subset of a frequent itemset is also frequent
  - It has at least the support of its superset

Apriori Algorithm (Continued)
Confidence

- Iteratively grow the frequent itemsets from size 1 to size K (or until we run out of support).
  - Apriori property tells us how to prune the search space
- Frequent itemsets are used to find rules X->Y with a minimum **confidence**:
  - Confidence: The % of transactions that contain X, which also contain Y
- Output: The set of all rules X -> Y with minimum support and confidence

Lift and Leverage

$$\text{Lift}(X \rightarrow Y) = \frac{\text{Support}(X \wedge Y)}{\text{Support}(X) * \text{Support}(Y)}$$

$$\text{Leverage}(X \rightarrow Y) = \text{Support}(X \wedge Y) - \text{Support}(X) * \text{Support}(Y)$$

Association Rules Implementations

- Market Basket Analysis
  - People who buy milk also buy cookies 60% of the time.
- Recommender Systems
  - "People who bought what you bought also purchased….".
- Discovering web usage patterns
  - People who land on page X click on link Y 76% of the time.

Use Case Example: Credit Records

Credit ID	Attributes
1	credit_good, female_married, job_skilled, home_owner, …
2	credit_bad, male_single, job_unskilled, renter, …

Minimum Support: 50%

Frequent Itemset	Support
credit_good	70%
male_single	55%
job_skilled	63%
home_owner	71%
home_owner, credit_good	53%

The itemset {home_owner, credit_good} has minimum support.

The possible rules are

credit_good -> home_owner

and

home_owner -> credit_good

Computing Confidence and Lift

Suppose we have 1000 credit records:

	free_housing	home_owner	renter	total
credit_bad	44	186	70	**300**
credit_good	64	527	109	**700**
	108	**713**	**179**	

713 home_owners, 527 have good credit.
home_owner -> credit_good has confidence 527/713 = 74%

700 with good credit, 527 of them are home_owners
credit_good -> home_owner has confidence 527/700 = 75%

The lift of these two rules is

$$0.527 / (0.700*0.713) = 1.055$$

A Sketch of the Algorithm

- If L_k is the set of frequent k-itemsets:
  - Generate the candidate set C_{k+1} by joining L_k to itself
  - Prune out the (k+1)-itemsets that don't have minimum support Now we have L_{k+1}
- We know this catches all the frequent (k+1)-itemsets by the apriori property
  - a (k+1)-itemset can't be frequent if any of its subsets aren't frequent
- Continue until we reach k_{max}, or run out of support
- From the union of all the L_k, find all the rules with minimum confidence

Step 1: 1-itemsets (L1)

- let min_support = 0.5
- 1000 credit records
- Scan the database
- Prune

Frequent Itemset	Count
credit_good	700
~~credit_bad~~	~~300~~
male_single	550
~~male_mar_or_wid~~	~~92~~
~~female~~	~~310~~
job_skilled	631
~~job_unskilled~~	~~200~~
home_owner	710
~~renter~~	~~179~~

Step 2: 2-itemsets (L2)

- Join L1 to itself
- Scan the database to get the counts
- Prune

Frequent Itemset	Count
~~credit_good, male_single~~	~~402~~
credit_good, job_skilled	544
credit_good, home_owner	527
~~male_single, job_skilled~~	~~340~~
~~male_single, home_owner~~	~~408~~
~~job_skilled, home_owner~~	~~452~~

Step 3: 3-itemsets

Frequent Itemset	Count
~~credit_good, job_skilled, home_owner~~	~~428~~

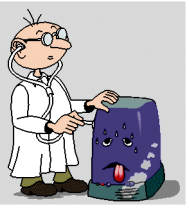
- We have run out of support.
- Candidate rules come from L2:
  - credit_good -> job_skilled
  - job_skilled -> credit_good
  - credit_good -> home_owner
  - home_owner -> credit_good

Finally: Find Confidence Rules

Rule	Set	Cnt	Set	Cnt	Confidence
IF credit_good THEN job_skilled	credit_good	700	credit_good AND job_skilled	544	544/700=77%
IF credit_good THEN home_owner	credit_good	700	credit_good AND home_owner	527	527/700=75%
IF job_skilled THEN credit_good	job_skilled	631	job_skilled AND credit_good	544	544/631=86%
IF home_owner THEN credit_good	home_owner	710	home_owner AND credit_good	527	527/710=74%

If we want confidence > 80%:
IF job_skilled THEN credit_good

Diagnostics



- Do the rules make sense?
  - What does the domain expert say?
- Make a "test set" from hold-out data:
  - Enter some market baskets with a few items missing (selected at random). Can the rules predict the missing items?
  - Remember, some of the test data may not cause a rule to fire.
- Evaluate the rules by lift or leverage.
  - Some associations may be coincidental (or obvious).

Apriori - Reasons to Choose (+) and Cautions (-)

Reasons to Choose (+)	Cautions (-)
Easy to implement	Requires many database scans
Uses a clever observation to prune the search space •Apriori property	Exponential time complexity
Easy to parallelize	Can mistakenly find spurious (or coincidental) relationships •Addressed with Lift and Leverage measures

Check Your Knowledge

Your Thoughts?

1. What is the Apriori property and how is it used in the Apriori algorithm?
2. List three popular use cases of the Association Rules mining algorithms.
3. What is the difference between Lift and Leverage. How is Lift used in evaluating the quality of rules discovered?
4. Define Support and Confidence
5. How do you use a "hold-out" dataset to evaluate the effectiveness of the rules generated?

Module 6: Advanced Analytics – Theory and Methods

Lesson 2: Association Rules - Summary

During this lesson the following topics were covered:

- Association Rules mining
- Apriori Algorithm
- Prominent use cases of Association Rules
- Support and Confidence parameters
- Lift and Leverage
- Diagnostics to evaluate the effectiveness of rules generated
- Reasons to Choose (+) and Cautions (-) of the Apriori algorithm

Lab Exercise 5 - Association Rules

- This Lab is designed to investigate and practice Association Rules.

  After completing the tasks in this lab you should be able to:

  - Use R functions for Association Rule based models

Lab Exercise 5 - Association Rules - Workflow

1. Set the Working Directory and install the “arules” package
2. Read in the Data for Modeling
3. Review Transaction data
4. Plot Transactions
5. Mine the Association Rules
6. Read in Groceries dataset
7. Mine the Rules for the Groceries Data
8. Extract the Rules in which the Confidence Value is >0.8 and high lift