IST 3420: Introduction to Data Science and Management

Langtao Chen, Fall 2017

9. Unsupervised Machine Learning

Reading

- "Data Mining For Business Intelligence" : http://proquest.safaribooksonline.com/book/databases/business-intelligence/9780470526828
  - Chapter 14: Cluster Analysis
  - Chapter 13: Association Rule

- If off campus, you need S&T VPN for free access

Learning Objectives

- Cluster analysis
  - Understand cluster analysis (both hierarchical and nonhierarchical)
  - Be able to use R to conduct cluster analysis

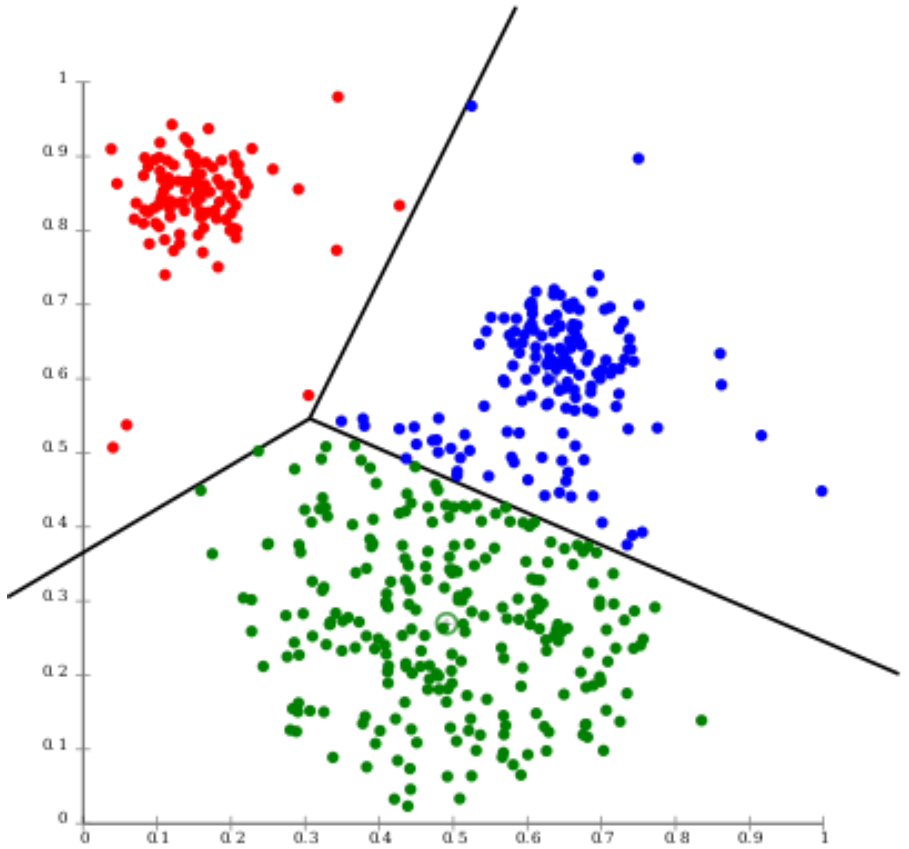
- Association rules
  - Understand association rules and measures including support, confidence, and lift; Be able to manually calculate them
  - Be able to use arules R package to conduct market basket analysis and properly interpret results
  - Be able to use aruleViz R package to visualize association rules and properly interpret results

Agenda

- Cluster Analysis
- Association Rules

Cluster Analysis

- Cluster analysis is to segment data into a set of clusters in such a way that objects in the same cluster are more similar.
- It is an unsupervised machine learning algorithm since the data are not labelled.



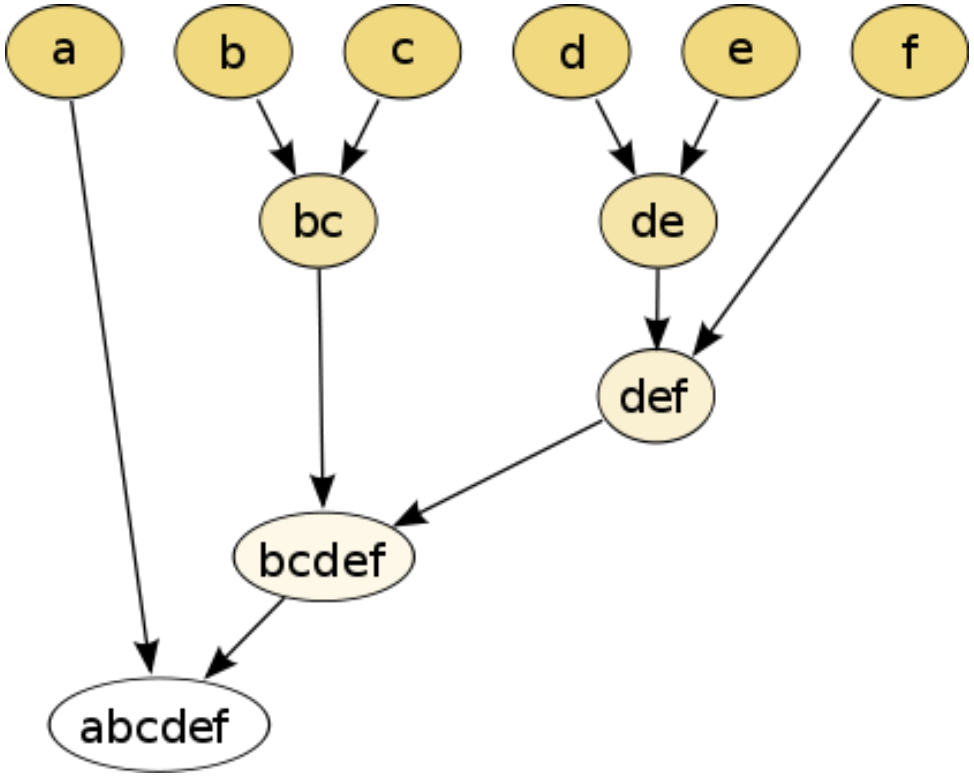
Clustering Algorithms

- Hierarchical clustering
- Non-hierarchical clustering: K-means algorithm

Hierarchical Clustering Analysis (HCA)

Hierarchical clustering analysis aims to build a hierarchy of clusters.

- Agglomerative:
  - Start with n clusters and sequentially merge similar clusters until all records are merged into one single cluster.
- Divisive
  - Start with one single cluster containing all records, then divide clusters into sub-clusters.



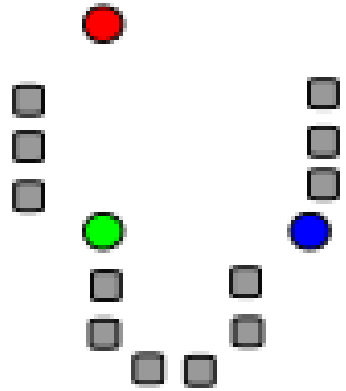
Non-Hierarchical Clustering Analysis

Non-hierarchical clustering analysis aims to obtain a set of clusters which maximizes or minimizes some evaluating criterion.

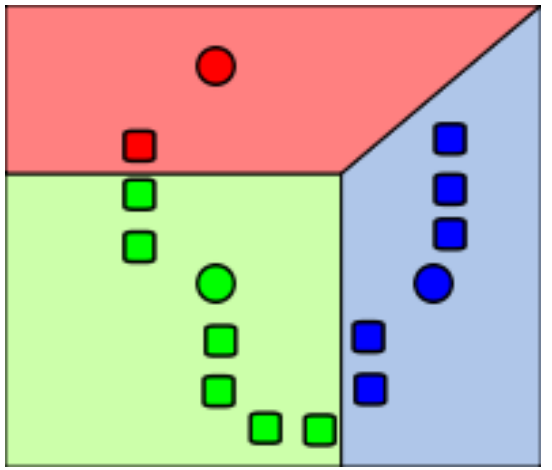
- K-means clustering:
  - Assign records to a *user-defined* number of clusters (*k*) in such a way that maximizes the separation of those clusters while minimizing intra-cluster distances.

K-Means Clustering

- Step 1. *k* initial "means" (in this case *k*=3) are randomly generated within the data domain (shown in color).



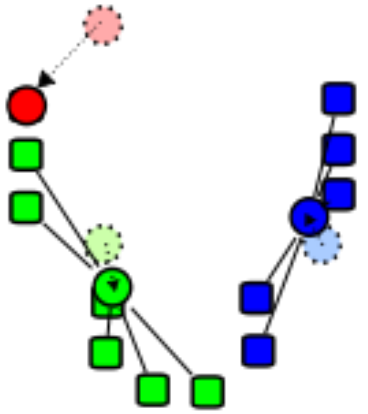
- Step 2. *k* clusters are created by associating every observation with the nearest mean. The partitions here represent the Voronoi diagram generated by the means.



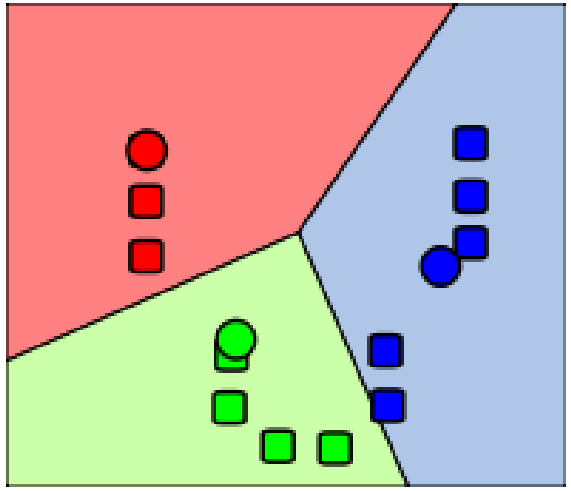
Source: https://en.wikipedia.org/wiki/K-means_clustering

K-Means Clustering (cont.)

- Step 3. The centroid of each of the *k* clusters becomes the new mean.



- Step 4. Steps 2 and 3 are repeated until convergence has been reached.



Source: https://en.wikipedia.org/wiki/K-means_clustering

Measuring Similarity between Records

- In the clustering process, similarity or distance needs to be assessed.

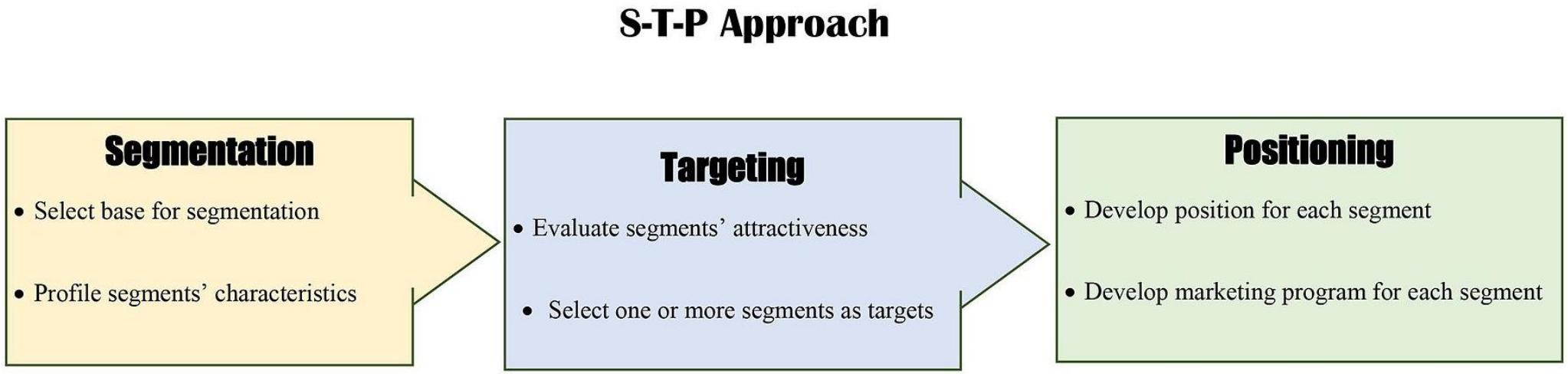
- For continuous variables, a commonly used distance metric is Euclidean distance.

$$d(p,q) = d(q,d) = \sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2 + \cdots + (q_n - p_n)^2} = \sqrt{\sum_{i=1}^{n} (q_i - p_i)^2}$$

- We can use other metrics such as standardized Euclidean distance, Mahalanobis distance, Minkowski distance, Chebychev distance, Cosine distance, Hamming distance, Manhattan distance, Jaccard distance, Spearman distance etc.

Market Segmentation: A Popular Use of Cluster Analysis

- Market segmentation is the process of dividing a broad consumer or business market, normally consisting of existing and potential customers, into sub-groups of consumers (known as segments) based on some type of shared characteristics.
- Consumers in a segment behave in similar way or have similar demands.



Source: https://en.wikipedia.org/wiki/Market_segmentation

Market Segmentation Sample Data

- Data source:
  - https://archive.ics.uci.edu/ml/datasets/Online+Retail
  - This is a transnational data set which contains all the transactions occurring between 01/12/2010 and 09/12/2011 for a UK-based and registered non-store online retail.
  - The company mainly sells unique all-occasion gifts.
  - Many customers of the company are wholesalers.

Dataset

- Attributes:
  - InvoiceNo: Invoice number. Nominal, a 6-digit integral number uniquely assigned to each transaction. If this code starts with letter 'c', it indicates a cancellation.
  - StockCode: Product (item) code. Nominal, a 5-digit integral number uniquely assigned to each distinct product.
  - Description: Product (item) name. Nominal.
  - Quantity: The quantities of each product (item) per transaction. Numeric.
  - InvoiceDate: Invice Date and time. Numeric, the day and time when each transaction was generated.
  - UnitPrice: Unit price. Numeric, Product price per unit in sterling.
  - CustomerID: Customer number. Nominal, a 5-digit integral number uniquely assigned to each customer.
  - Country: Country name. Nominal, the name of the country where each customer resides.

	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
1	536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	12/1/2010 8:26	2.55	17850	United Kingdom
2	536365	71053	WHITE METAL LANTERN	6	12/1/2010 8:26	3.39	17850	United Kingdom
3	536365	84406B	CREAM CUPID HEARTS COAT HANGER	8	12/1/2010 8:26	2.75	17850	United Kingdom
4	536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	6	12/1/2010 8:26	3.39	17850	United Kingdom
5	536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	6	12/1/2010 8:26	3.39	17850	United Kingdom
6	536365	22752	SET 7 BABUSHKA NESTING BOXES	2	12/1/2010 8:26	7.65	17850	United Kingdom
7	536365	21730	GLASS STAR FROSTED T-LIGHT HOLDER	6	12/1/2010 8:26	4.25	17850	United Kingdom
8	536366	22633	HAND WARMER UNION JACK	6	12/1/2010 8:28	1.85	17850	United Kingdom

RFM Method for Market Segmentation

- RFM (Recency, Frequency, Monetary value) analysis is a popular technique to market segmentation.

- RFM analysis tries to cluster customers with similar buying patterns based on the following characteristics:
  - **R**ecency – *How recently did the customer purchase?*
  - **F**requency – *How often do they purchase?*
  - **M**onetary Value – *How much do they spend?*

Data Analysis

Refer to "Market_Segmentation.pdf"

Agenda

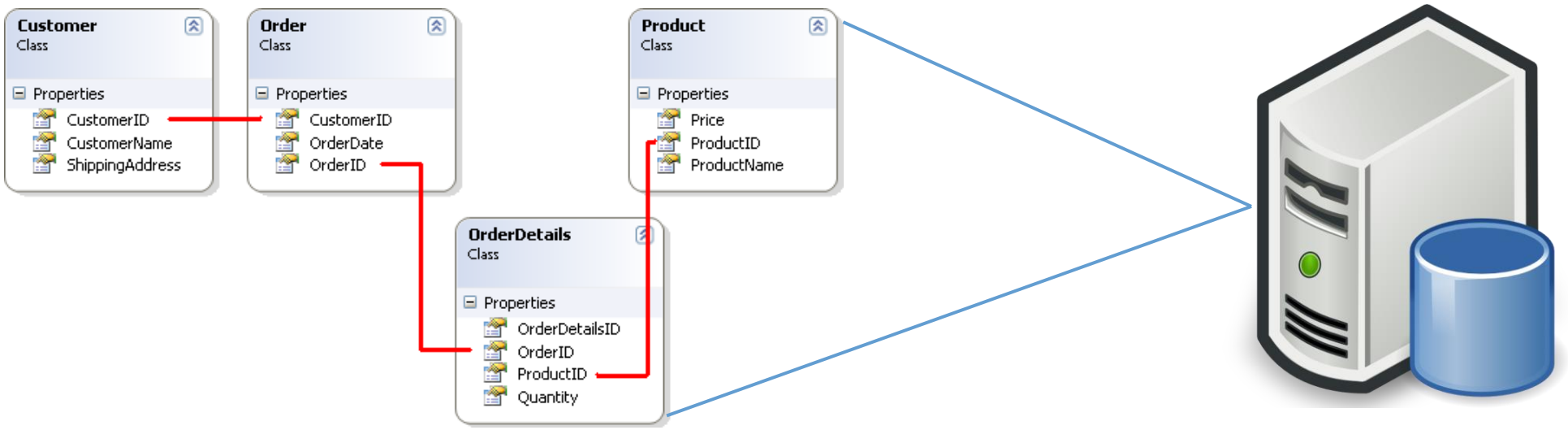
- Cluster Analysis
- Association Rules

Market Basket Analysis

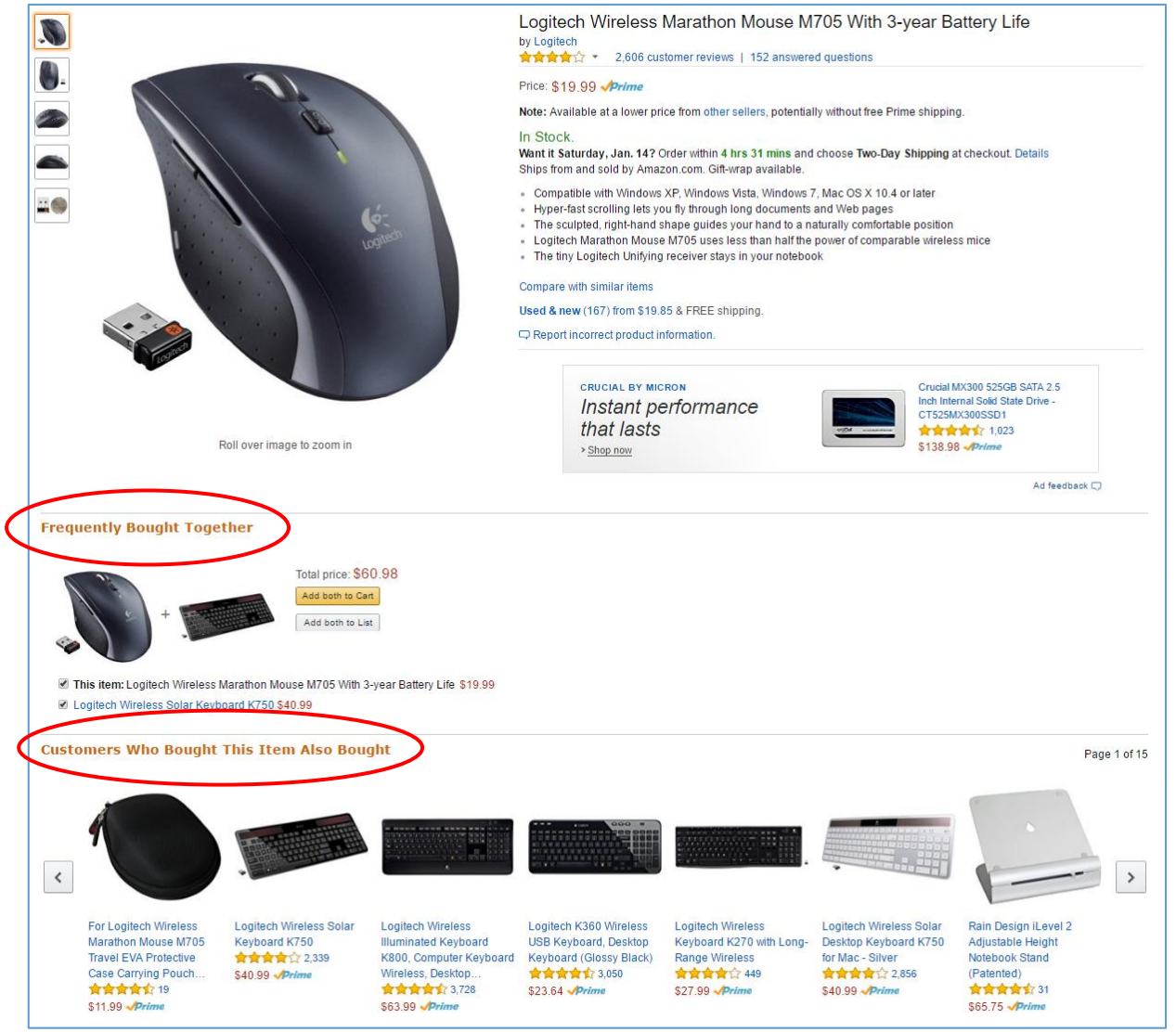
- Association rules are used to conduct market basket analysis
- The objective is to identity item clusters in transaction-type database (a.k.a. "affinity analysis").
- It is an unsupervised learning method since we don't need to guide the machine learning process.
- Having such customer behavior knowledge would help so many decision making cases in business especially in marketing
  - What items should be placed together in retail stores?
  - What items should be offered in post-transaction coupons?
  - What products should be recommended in online shopping sites?
  - ……

Why Market Basket Analysis?

- Transactional data are readily available in databases of many companies.
- Transactional data are objective (not like subjective data collected from customer survey).



Application Case: Amazon



The Target

- To generate clear and simple rules of the form:

IF X is purchased, THEN Y is also likely to be purchased.

For example:

LHS		RHS
{beef, dairy produce}	=>	{vegetables}

- General Steps:
  - Generate a set of candidate rules based on frequent itemsets (Apriori is the most popular algorithm);
  - Select the rules that indicate the strongest association between items.

Example: 10 Association Rules

	lhs	rhs	support	confidence	lift
[1]	{beef,dairy produce}	=> {vegetables}	0.030	0.61	2.2
[2]	{poultry}	=> {vegetables}	0.029	0.57	2.1
[3]	{dairy produce,fruit,sausage}	=> {vegetables}	0.027	0.57	2.1
[4]	{beef}	=> {vegetables}	0.046	0.56	2.0
[5]	{dairy produce,vinegar/oils}	=> {vegetables}	0.031	0.54	2.0
[6]	{fruit,sausage}	=> {vegetables}	0.034	0.53	1.9
[7]	{bread and backed goods,dairy produce,fruit}	=> {vegetables}	0.041	0.53	1.9
[8]	{pork}	=> {vegetables}	0.030	0.52	1.9
[9]	{cheese,fruit}	=> {vegetables}	0.027	0.52	1.9
[10]	{dairy produce,fruit,non-alc. drinks}	=> {vegetables}	0.033	0.52	1.9

Association Rule Measures

lhs	rhs	support	confidence	lift
[1] {beef,dairy produce}	=> {vegetables}	0.030	0.61	2.2

- **Support**: The fraction of which the item set occurs in the transaction dataset.

$$Support(X => Y) = \frac{\#\ of\ transactions\ containing\ X\ and\ Y}{\#\ of\ alll\ transactions}$$

- **Confidence**: The probability that a rule is correct for a new transaction with items on the left hand side.

$$Confidence(X => Y) = \frac{\#\ of\ transactions\ containing\ X\ and\ Y}{\#\ of\ transactions\ containing\ X}$$

- **Lift**: The ratio of the confidence of a rule and the expected confidence of the rule.

$$Lift(X => Y) = \frac{Confidence(X => Y)}{Support(Y)} = \frac{(\#\ of\ transactions\ containing\ X\ and\ Y) * (\#\ of\ alll\ transactions)}{(\#\ of\ transactions\ containing\ X) * (\#\ of\ transactions\ containing\ Y)}$$

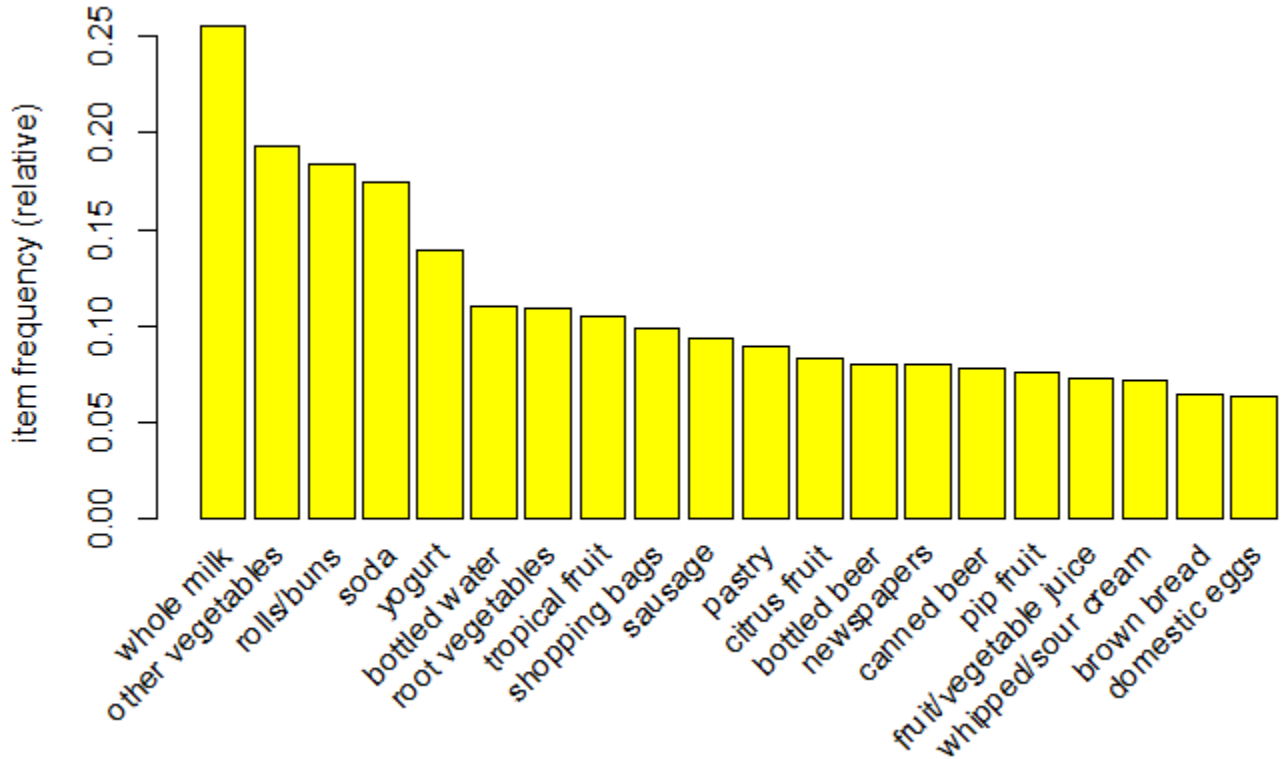
Exercise

- Given a transaction database:

Transaction	Items Purchased		
1	fruit	coffee	
2	yogurt	milk	bread
3	coffee	cheese	
4	fruit	drink	
5	coffee		

- Calculate measures of the rule {fruit}=>{coffee}
  - Refer to "Association Rule Exercise.xlsx"

Frequency


item frequency (relative)
0.00
0.05
0.10
0.15
0.20
0.25
whole milk
other vegetables
rolls/buns
soda
yogurt
bottled water
root vegetables
tropical fruit
shopping bags
sausage
pastry
citrus fruit
bottled beer
newspapers
canned beer
pip fruit
fruit/vegetable juice
whipped/sour cream
brown bread
domestic eggs

The arules and arulesViz R Packages

- arules Package
  - Call read.transactions() to read in transaction data file and create a transactions object for association rule mining
  - Call itemFrequencyPlot() to draw frequency plot
  - Call apriori() to mine associations rules using Apriori algorithm

- arulesViz Package
  - Call plot() to visualize association rules and itemsets

- To learn more, refer to:

Introduction to arules.pdf
Visualizing Association Rules.pdf

Case: Grocery Basket

- 9835 transactions, 169 items, 10 level 1 item categories, 55 level 2 item categories

	TransactionID	Item	ItemCategory1	ItemCategory2
1	1	citrus fruit	fruit and vegetables	fruit
2	1	semi-finished bread	fresh products	bread and backed goods
3	1	margarine	processed food	vinegar/oils
4	1	ready soups	processed food	soups/sauces
5	2	tropical fruit	fruit and vegetables	fruit
6	2	yogurt	fresh products	dairy produce
7	2	coffee	drinks	coffee
8	3	whole milk	fresh products	dairy produce
9	4	pip fruit	fruit and vegetables	fruit
10	4	yogurt	fresh products	dairy produce
11	4	cream cheese	fresh products	cheese
12	4	meat spreads	canned food	meat spreads

Data Analysis

Refer to "Market_Basket_Analysis.pdf"

Q & A