Data Mining

Mohammed Almulla

College of I.T

Principles of Data Mining

As I explain the meaning of data mining above now, I will talk about the data mining process. It has three stages. The first one is preparing input data that consists of data collection, data selection to select relevant features and records, data preprocessing to clean and transform data and deal with unknown data and reduce dimensionality (Reduce the number of attributes), and data formatting into acceptable form. The second stage is mining patterns that competent in mining from input data to patterns that a sensible data mining task must be designed to comply with the objectives. It determines data mining tasks, assigning roles for data for certain tasks, selecting data mining solution(s) to each task, setting necessary parameters for the data mining technique, and collecting results and patterns. The final stage is post-processing patterns and its specialist in evaluate, select, and interpret the pattern. Figure 7 is shown the data mining process with each stage.

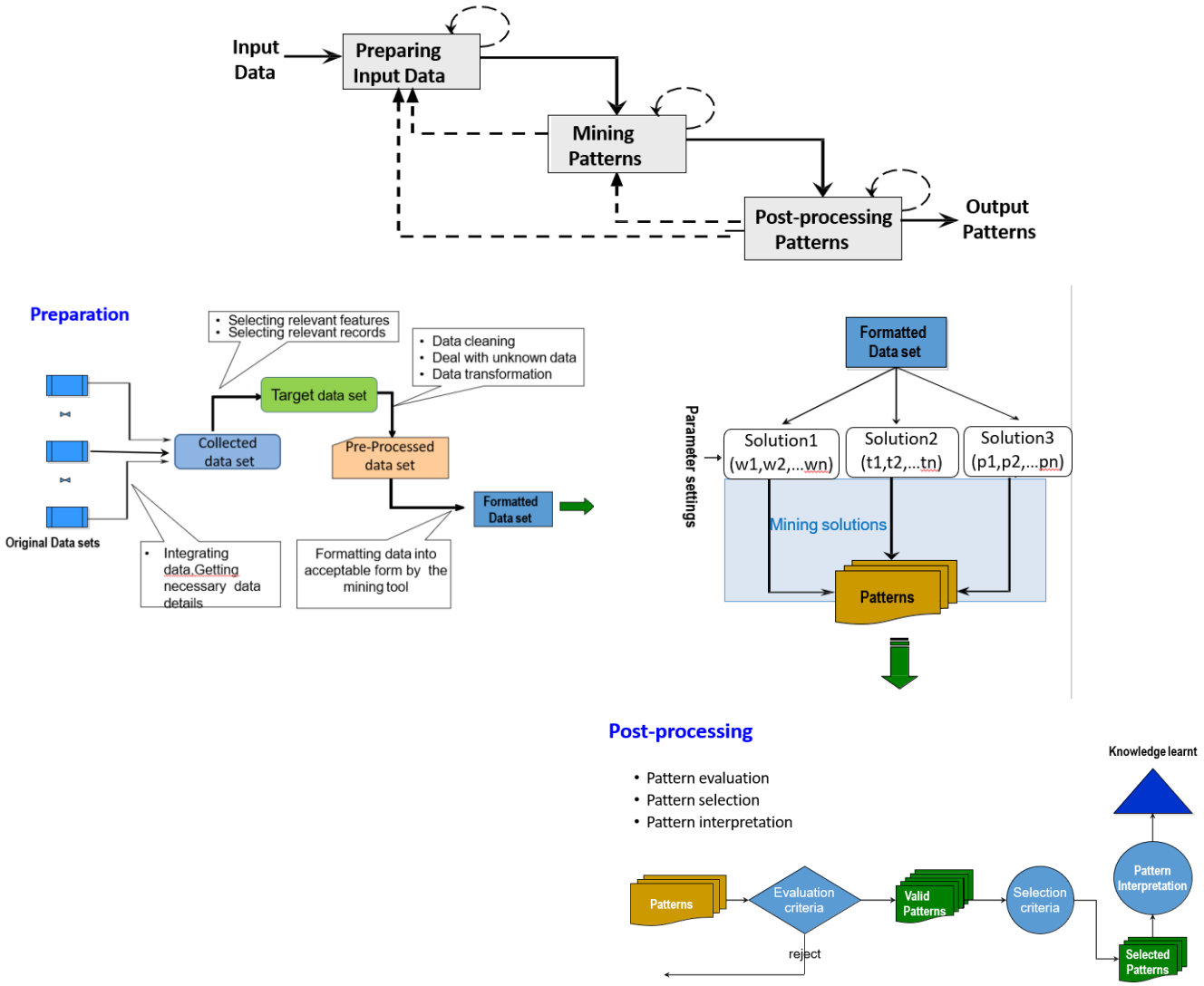


Figure 7: Data mining process

As I talk in the introduction about the two categories that data mining objectives classified to, I will explain each one in details. Firstly, supervised learning and its goal is to predict a single target or outcome variable by using training data with target value is known and then score to unseen data or new data where target value is unknown. Some of the methods of supervised learning are classification, prediction, and regression. Classification is output a target variable in binary form such as yes/no, or 0/1. And for more than two targets such as low/medium/high we called it multi-classification problem. Some of the representations of classification models are input space, classification rules, decision trees, and many more. As shown in figure 8 an illustration of classification process represented in input space. The input space or sometimes called instance space is multidimensional space defined by descriptive

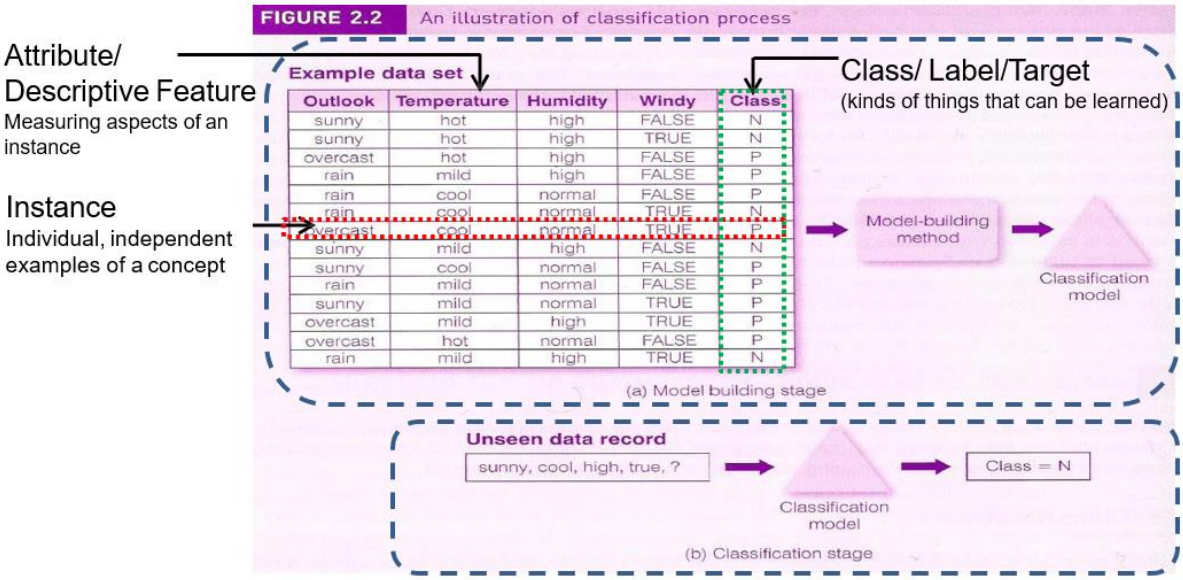


Figure 8: An illustration of classification process (input space)

feature variables. They held in computer to predict new data. In figure 9 below shown an example of classification in instance space with several real-life objects using k-Nearest-Neighbor (kNN) that classify objects based on distance or similarity between it and other objects.

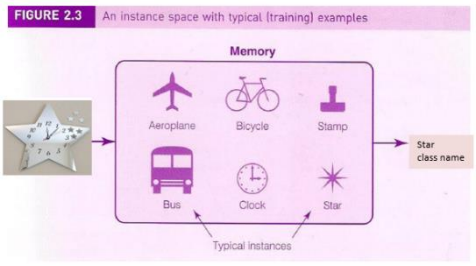


Figure 9: An instance space with training examples

Another representation is classification rules that are an if-then assertion that states a fact about the data set. For example, IF (20,000 < Income < 40,000) AND (OwnHouse = 'yes') THEN Class = 'Safe customer' (Shown in figure 10 an example on classification rules).

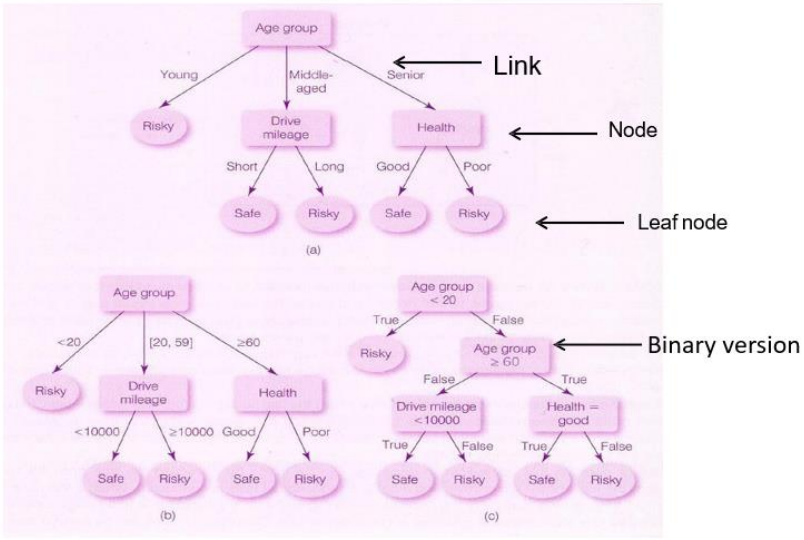
Age	marital status	gender	own House	income	class
39	Never-married	Male	yes	30000	safe
50	Married-civ-spouse	Male	yes	45000	safe
38	Divorced	Male	yes	12000	not safe
53	Married-civ-spouse	Male	no	43000	not safe
37	Married-civ-spouse	Female	yes	23000	safe
49	Married-spouse-absent	Female	no	34333	not safe
31	Never-married	Female	yes	65000	safe
37	Married-civ-spouse	Male	no	51000	not safe

Is this a safe customer?

Age	marital status	gender	own House	income	class
33	-	Female	No	45000	??

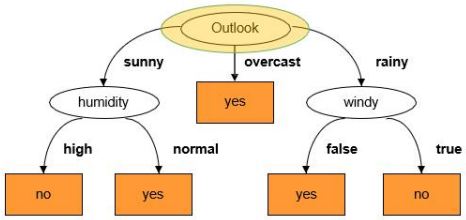
Figure 10: classification rules table

Moreover, a decision tree is a representation for classification models that is a tree structure consisting of connected nodes. Figure 11 shows examples of decision trees.



Outlook	Temperature	Humidity	Windy	Class
sunny	hot	high	FALSE	N
sunny	hot	high	TRUE	N
overcast	hot	high	FALSE	P
rain	mild	high	FALSE	P
rain	cool	normal	FALSE	P
rain	cool	normal	TRUE	N
overcast	cool	normal	TRUE	P
sunny	mild	high	FALSE	N
sunny	cool	normal	FALSE	P
rain	mild	normal	FALSE	P
sunny	mild	normal	TRUE	P
overcast	mild	high	TRUE	P
overcast	hot	normal	FALSE	P
rain	mild	high	TRUE	N

Attribute	Rules	Errors	Total errors
Outlook	Sunny →No	2/5	4/14
	Overcast →Yes	0/4	
	Rainy →Yes	2/5	
Temp	Hot →No*	2/4	5/14
	Mild →Yes	2/6	
	Cool →Yes	1/4	
Humidity	High →No	3/7	4/14
	Normal →Yes	1/7	
Windy	False →Yes	2/8	5/14
	True →No*	3/6	



Would some one play tennis when:

Outlook	Temperature	Humidity	Windy	Class
Sunny	Mild	Normal	True	?
Rainy	-	-	True	?

Figure 11: Examples of decision trees

The second method for supervised learning is regression or estimation that uses a function that represents a mapping pattern from an argument to a value. It finds the best fit line or curves among the data points. For example, $F(n) = n^2 + 2n - 10$ is a function where n represents yesterday's price and F(n) is today's price. Figure 12 shows examples of a line or curve.

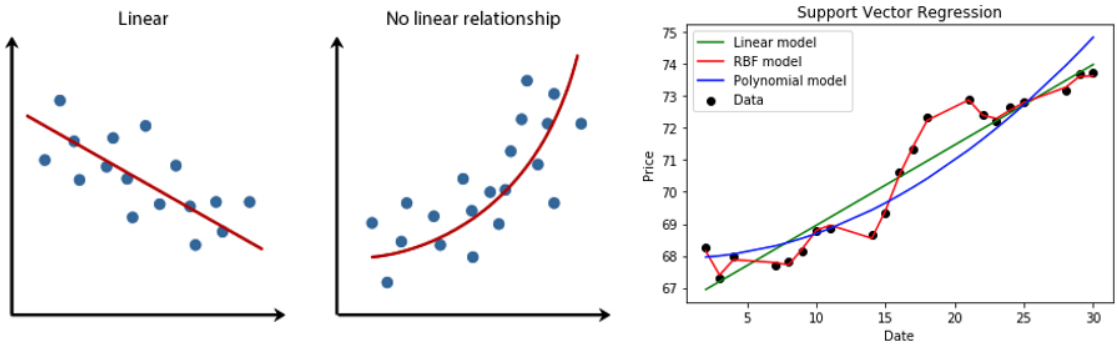


Figure 12: Examples of regression line or curve

Another model or pattern in supervised learning that can be used in classification or estimation is an artificial neural network (ANN). It emulates the human brain and consists of nodes and links or edges located on different layers. Each node has a value, and each edge has a weight. As shown in figure 13 a structure of an ANN. As well for supervised learning, we have various methods or patterns such as SVM or many more.

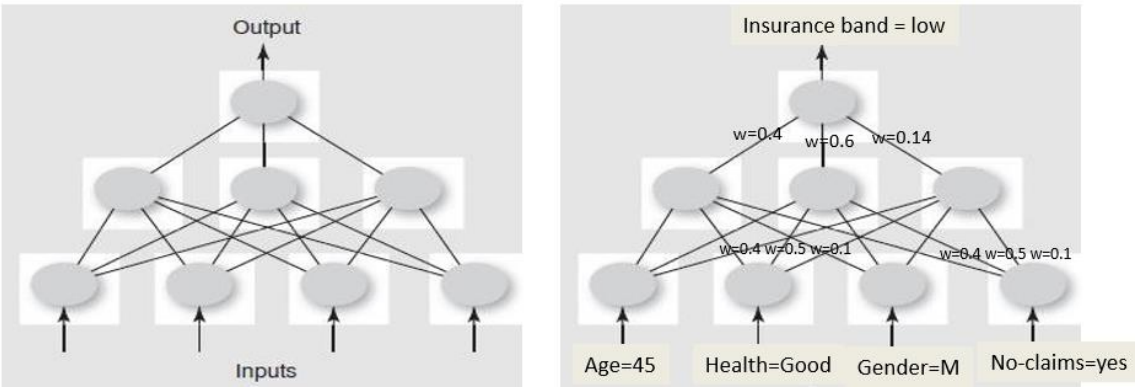


Figure 13: Structure of an ANN

The second category for data mining objectives is unsupervised learning that takes no outcomes or labels and leaves the machine to learn by itself. It is goal is to segment data into meaningful segments and detect patterns. And as I said there is no target or outcome variable to predict or classify. It groups or associates unseen data. Some of the methods for unsupervised learning are clustering, association rules, data reduction & exploration, visualization, and many more. Clustering or cluster detection measures similarity among data objects and group them into clusters accordingly (As shown in figure 14).

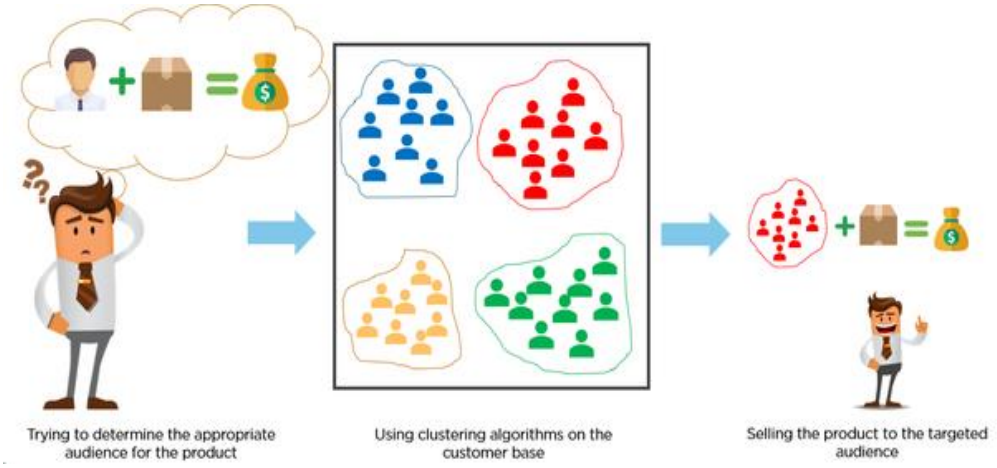


Figure 14: Clustering

Also, association rules or it called affinity analysis are another unsupervised learning model whose goal is to produce rules that define what goes with what. For example, if X was purchased, Y was also purchased. It is used in recommender systems. Association rule mining is discovering significant relationships between data objects. Figure 15 shows the association rules example.

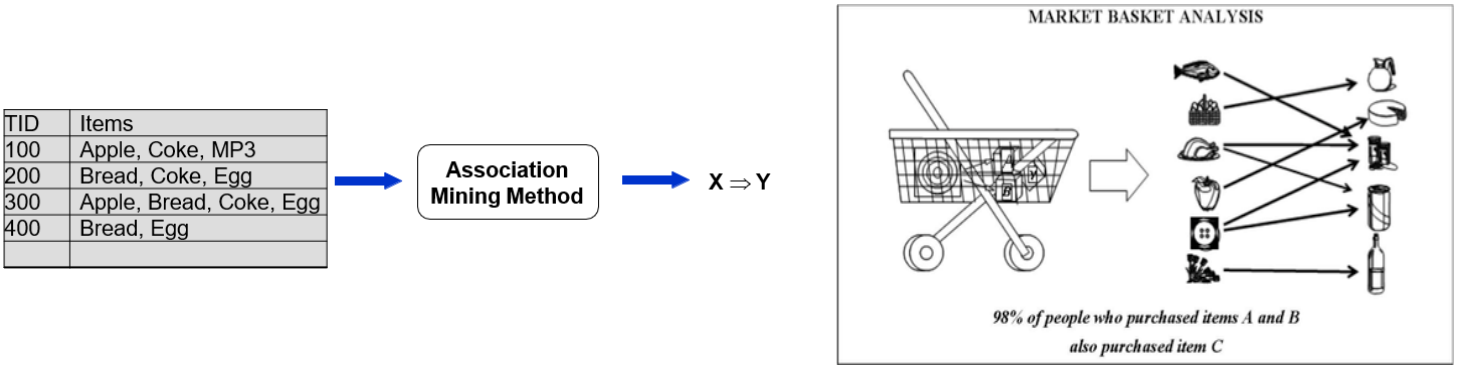


Figure 15: Association rules example

Furthermore, an unsupervised learning method is data reduction that means reduce the number of records, or instances, or rows in our data set. While reducing the number of attributes or columns is called dimensionality reduction. Data exploration is also an unsupervised learning method that understands the global landscape of the data and detecting unusual values to clean the data such as from data aggregations and summaries by looking at each variable and each relationship among them separately. Then, data visualization or also called visual analytics that is data exploration by creating charts and dashboards.

After talking about models or patterns we need to know how to evaluate them. So, I will start with the importance of evaluating result patterns that classification model must be accurate enough to be creditable, clusters must genuinely exist, association rules must have enough strengths to be believed, and data descriptions must be general enough to cover a large part of the data set. We have two possible

measures of interestingness or importance of the knowledge we received. Firstly, objective measures based on data and pattern that consists of conciseness that is number and complexity of patterns and this called in philosophy Occam's razor that means more things should not be used than are necessary or the simplest explanation is usually the best one (Figure 16 shows the complexity of pattern), coverage also is from the objective measures that are the percentage of data that supports the pattern, reliability that is a high percentage of occurrence among applicable cases, a peculiarity that measures of difference from the norm, and diversity that is how different from other patterns or the tendency of clusters. Secondly and finally, subjective measures based on domain knowledge and consists of novelty that is not known before & not trivial to be inferred, usefulness that helps achieve objectives or goals, surprisingness, and applicability that can be incorporated in business process and lead to benefit of some kind. The subjective measurements are very related to business and managers while objective measurements are for data miners.

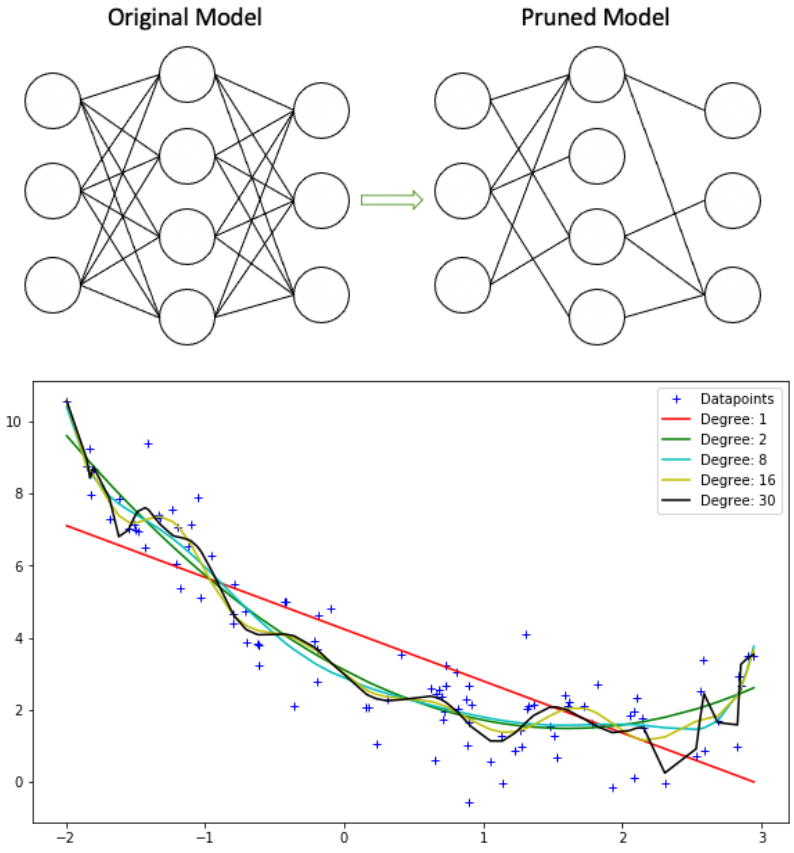


Figure 16: Complexity of pattern

Let's see now how to calculate the coverage or accuracy rate for classification that is by first split our data into training records that are what we will give to the model, validation or unseen data set that are records not used in training but used for the optimization, and a test set that is as a new data our model must predict the output of it. For training records, we must calculate rule coverage that is how much the left-hand side of the rule (X ➔ Y) is covering. And calculate accuracy that is how much both the left- & right-hand side of the rule is covering. Also, calculate the error rate for our training records. So, one of the commonly used measures is the accuracy rate or error rate for classification models as I explain above, and one of the techniques used in this measure is k-fold cross-validation that is a procedure used to estimate the skill of the model on new data. It partitions the data set into k-folds, test on 1-fold and train on the rest, and iterate for each k (As shown in figure 17).

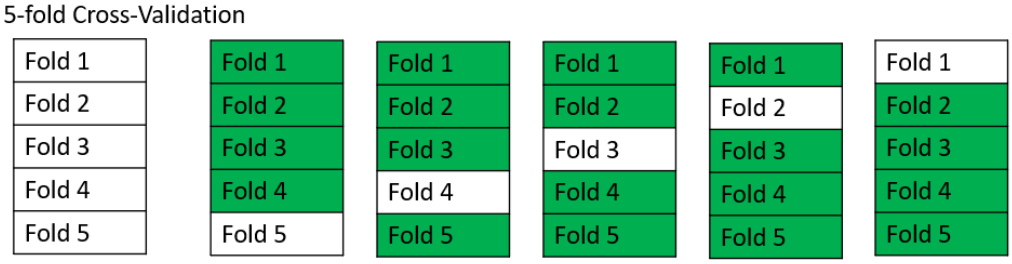


Figure 17: 5-fold cross-validation

Second commonly used measure specifically used in clustering called Quality of a cluster, and the other called Overall quality of all clusters. I will talk about quality of a cluster only that is consist of two measures. First one is Inter-cluster distance that maximize the distance between clusters, and second or finally is Intra-cluster distance that minimize the distance between point inside the cluster (As shown in figure 18 clustering example).

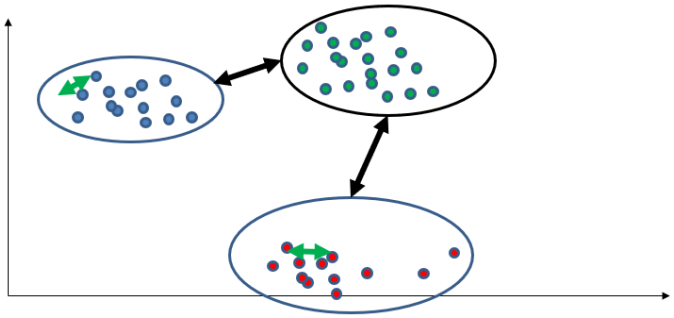


Figure 18: Clustering example

Finally, commonly used measures in association rules are support that is an indication of how frequently the itemset appears in the dataset, confidence that is an indication of how often the rule is true, and lift that is a measure of the performance of a targeting model. Figure 19 shows support, confidence, lift laws.

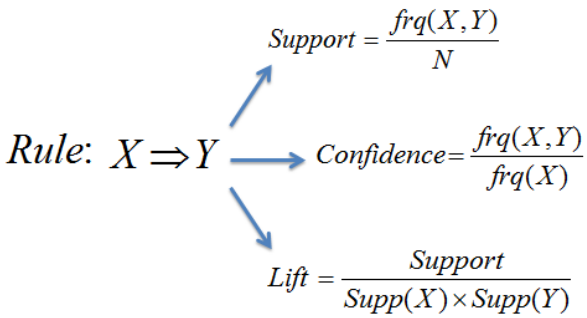


Figure 19: support, confidence, lift laws

References & Useful Resources:

- CSBP 320: Data Mining, UAE University
- http://www.saedsayad.com/association_rules.htm