International Journal of Combined Research & Development (IJCRD)
eISSN:2321-225X; pISSN:2321-2241 Volume: 2; Issue: 5; May -2014

STATISTICS AND DATA MINING

A Theoretical Overview

Dr. J Katyayani , Professor, Department of MBA, SPMVV, Tirupati, India
M Jahnavi , Research Scholar, Department of MBA, SPMVV, Tirupati, India

Abstract: Data mining is viewed as an computer automated exploratory data analysis for large complex data, it is viewed from statistical perspectives. Most of the methodologies used in data mining are originated in fields other than statistics. When the data sets are relatively large and homogeneous it is advised to use statistical techniques. Data mining algorithms work well with data sets of modest size may fail or take an unreasonably along time to run in really large data sets. In this paper we are explaining what are the common areas where both data mining and statistics are applicable and also along with the brief explanation of those techniques in data mining and statistics.

Keywords: Statistical Techniques, Data Mining, Exploratory data analysis.

DATA MINING AND STATISTICS:

Both data mining and statistics aims to discover the structure in data. Statistics is at the core of data mining - helping to distinguish between random noise and significant findings, and providing a theory for estimating probabilities of predictions, etc.
However Data Mining is more than Statistics. DM covers the entire process of data analysis, including data cleaning and preparation and visualization of the results, and how to produce predictions in real-time, etc.
Some people regard that data mining as a subset of statistics. Especially in the process of developing and assessing the models statistical procedures plays a major role. Most of the learning algorithms use statistical test at the time of constructing the rules. Some of the commonly used statistical analysis techniques are:

1.1 Descriptive and Visualization Techniques include simple descriptive statistics such as:
• averages and measures of variation,
• counts and percentages, and
• cross-tabs and simple correlations

They are useful for understanding the structure of the data. Visualization is primarily a discovery technique and is useful for interpreting large amounts of data; visualization tools include histograms, box plots, scatter diagrams, and multi-dimensional surface plots [Tegarden 1999].

Cluster Analysis seeks to organize information about variables so that relatively homogeneous groups, or "clusters," can be formed. The clusters formed with this family of methods should be highly internally homogenous (members are similar to one another) and highly externally heterogeneous (members are *not* like members of other clusters).

Correlation Analysis measures the relationship between two variables. The resulting correlation coefficient shows if changes in one variable will result in changes in the other. When comparing the correlation between two variables, the goal is to see if a change in the independent variable will result in a change in the dependent variable. This information helps in understanding an independent variable's predictive abilities. Correlation findings, just as regression findings, can be useful in analyzing causal relationships, but they do not by themselves establish causal patterns.

Discriminant Analysis is used to predict membership in two or more mutually exclusive groups from a set of predictors, when there is no natural ordering on the groups. Discriminant analysis can be seen as the inverse of a one-way multivariate analysis of variance (MANOVA) in that the levels of the independent variable (or factor) for MANOVA become the categories of the dependent variable for discriminant analysis, and the dependent variables of the MANOVA become the predictors for discriminant analysis.

Factor Analysis is useful for understanding the underlying reasons for the correlations among a group of variables. The main applications of factor analytic techniques are to reduce the number of variables and to detect structure in the relationships among variables; that is to classify variables.

Therefore, factor analysis can be applied as a data reduction or structure detection method. In an exploratory factor analysis, the goal is to explore or search for a factor structure.
Confirmatory factor analysis, on the other hand, assumes the factor structure is known a priori and the objective is to empirically verify or confirm that the assumed factor structure is correct. Communications of the Association for Information Systems (Volume 8, 2002) 267-296 275 Data Mining: A Conceptual Overview by J. Jackson

2. Difference between data mining and statistical analysis:

2.1Data Mining and Statistics:

Data mining is the extension of standard statistics, because many of the data mining techniques invented by statisticians. And now these data mining techniques have been integrated into statistical software. Both the data miners and statisticians use similar techniques to solve similar type of problems, but the data mining approach differs from the standard statistical approach in several areas such as:

SI.No	Statistics	Data mining
1	It deals with structured data in order to solve structured problems	It deals with structured data in order to solve unstructured business problems
2	Inference reflects statistical hypothesis testing	Inference reflects computational properties of data mining algorithm at hand

It assumes that there is more than enough data and processing power, as well as dependency on time.
It is very hard to design experiments in the business world without data mining.

Another major difference between business data and scientific data is that the latter is non-censored data and the former is censored data. Given a methodology or an algorithm to analyze data, it is often very hard to say whether it is "Statistics" or "Data Mining". While dealing with real life business problems, no one will ask whether you are a data miner or statistician. Because their main interest is to solve the problem. And for them it is immaterial what label they are using. Businesses increase revenues, maximize operating efficiency, minimizing the costs and improves customer satisfaction by employing advanced data mining techniques. Whereas with the help of statistics, one can transform the raw data into a knowledgeable for the business processes. Because without Statistics, there is no effective analysis. Without effective analysis, there is no business intelligence. Without business intelligence, how can you hope to assimilate gigabytes of data and consistently make decisions that will keep you ahead of your competition? Nowadays, because of statistical software we are improving our competitiveness from the shop floor to the sales floor and also to the executive floor.

It is very important for each and every business in today's environment, is mining of the data for all it's worth in order to increase the market share, improving the operational effectiveness, and have a keep record of market trends and prediction of future outcomes otherwise if you are not mining data you are guilty of underuse of your company's greatest assets.

The central aim of data mining is discovery it not concerned with the areas of statistics about how best to collect the data in order to answer specific questions such as experimental design and survey design at the first place. Data mining is concerned in discovering the secrets from the data when the data is already collected.

When data mining technology is appropriate?
When the problems are unstructured
If we are giving more importance to accurate prediction rather than explanation
When the data is a mixture of interval, nominal, ordinal, count
If the relationship between the variables are non-linear
When the data is highly heterogeneous with large percentage of missing values, outliers and leverage points.
The sample size is relatively large

2.2 Does Data Mining Mean Statistics or More Than Statistics?

The synonymous of data mining are data dredging or data fishing which is used to describe the process of identifying the patterns through the trawling of data. The term 'Data Mining' conveys to the statisticians that the sense of naïve hope vainly struggling against the cold realties of chance, Statisticians have typically not concerned themselves with data sets containing many millions or even billions of records. Moreover, special storage and

manipulation techniques required to handle them have been developed by entirely different intellectual communities from statisticians. Most of the statisticians concerned with primary data analysis, On the other hand data mining is entirely concerned with secondary data analysis. In fact we might define 'Data Mining' as the process of secondary analysis of large databases aimed at finding unsuspected relationships, which are of interest or value to the database owners. From this we can see that 'data mining' is opposed to hypothetic-deductive approach.

One more major difference is statistics deals with numeric data but data mining can extract interesting patterns and structures from the databases when databases contains other kinds of data such as image data, audio data, text data and geographical data. It is not possible to ask simply the computer to search interesting patterns or find any structure in the data. Before that one needs to define what is mean by patterns or structure. And before that one needs to decide what one means by "interesting".

In principle, a statistical expert system would embody a large base of intelligent understanding of the data analysis process, which it could apply automatically to a relatively small set of data. Whereas a data mining system, embodies a small base of intelligent understanding, and applies to a large data set. In both cases the applications are automatic and also both the cases require interaction with the researcher and we consider it as fundamental. In statistical expert system, the program drives the analysis following a statistical strategy, whereas in data mining application, the program drives the analysis because the user has insufficiently resources to manually examine billions of records and hundreds of thousands of potential patterns.

2.3 Should Data Mining be Included in the 'Statistics' Curriculum?

Statistical data may be characterized by the data which are small, clean, static and randomly sampled, and often collected to answer a specific question. None of this is applicable to data mining context. A data set with few thousand observations may be considered as large for classical statistician, whereas the same thing is considered as too small for a data miner.

Majority of the data miners appear to have little formal statistical expertise, because there is a chance of committing errors which trained statisticians would avoid obviously. So to avoid this it's better that the data miners must have broad statistical insights regarding the potentials for spurious associations and issues of substance versus statistical significance. This leads to the requirement of training to the data miners in statistics or statistics graduates in data mining. Because of this some of them are redesigning the statistics curriculum which is mainly focusing on changing trends in data collection and analysis,. The last decade or so have seen hundred of computer software manufacturers jumping onto the data mining bandwagon. Major statistical software packages such as SAS, S-PLUS, SPSS, and STATISTICA, etc. are being marketed as data mining tools rather than statistical tools.

Now a day if we observe that computer scientists have beaten the statisticians in offering data mining courses from since 1990's because of the advent of substantial improvements in computing power. There is a mutual ignorance between statisticians and data miners because of conservativeness in statistics versus a risk taking attitude in computing. If we see that there is a possibility of merging insights of computing specialists with those of statisticians due to progress in data mining. It is indictment to the statistical profession that some of the statisticians involved in a deep way with data mining. Most of the data miners ignorant to statistics and client's domain; whereas statisticians are ignorant to data mining and client's domain not but not the least client's domain are ignorant to statisticians and data mining. Because unfortunately statisticians focus on identifying and handling uncertainties, computer scientists focus upon database manipulations and clients focus upon integrating knowledge into the knowledge domain. In the near future both the data mining and statistics will grow towards each other because without statistical thinking data mining will not become knowledge discovery, without data mining approaches statistics failed to succeed on massive and complex datasets.

A successful knowledge discovery requires a substantial collaboration from all the three communities that is from computer science, statistics and client knowledge. Knowledge discovery rests on all these three domains, it cannot stand without the absence of any one of these domains. Now the major maturity challenge is the recognition of their dependence on each other for the data miners, statisticians and clients. All of them have to widen their focus until true collaboration becomes reality. There is no doubt that data mining is statistically intellectual. Data mining is essentially a statistical data analysis. Previous experience tells us statistics

has a poor record for the timely recognition of important ideas, because of this there is a chance of improving the reputation in this discipline. We can consider it as a great loss for the reputation of statistics as well as for individual statisticians if these were not grasped.

"Difference between statistics and data mining" concluded that it is immaterial what you can call it either data mining or statistics. Since 'computation' plays a major role in the process of data mining so computer scientists have significant claim over the ownership of data mining. Nevertheless, data mining techniques, in general, have a statistical base; and statisticians are beginning to show a significant interest in the area, including offering tertiary courses in statistical data mining.

3. Introduction to forecasting:

The following chapter explains some advanced models for forecasting such as multivariate time series models, neural network models etc with necessary derivations within each model. This is the area where we can apply either data mining techniques or statistical techniques of both for computing forecasting:

The following points needs to keep in mind in the case of advanced forecasting models.

Some of the advanced forecasting methods are still in developing stage and lot of research is going on in this direction.

Many of the advanced forecasting methods are applicable to some specific scenarios only. They cannot be generalized to all the forecasting problems.

The empirical investigation involves neural network model and hence the more emphasizes focus is given on neural network forecasting in the current research study.

All the advanced forecasting models involve some complex computational algorithms which can be carried out by analytical software like SAS, SPSS etc.

3.1 Why to forecast?

Frequently there is a time lag between awareness of an impending event or need and occurrence of that event. This lead time is the main reason for planning and forecasting. If the lead time is zero or very small, there is no need for planning. If the lead time is long, and the outcome of the final event is conditional on identifiable factors, planning can perform an important role. In such situations, forecasting is needed to determine when an event will occur or a need arise, so that appropriate actions can be taken.

In management and administrative situations the need for planning is great because the lead time for decision making ranges from several years to a few days or hours to a few seconds. Forecasting is an important aid in effective and efficient planning.

Regardless of these, two important comments must be kept in view. The first is that successful forecasting is not always directly useful to managers and others. More than 100 years ago, Jules Verne correctly predicted such developments as submarines, nuclear energy and travel to the moon. Similarly, in the mid-1800s, Charles Babbage not only predicted the need for computers, but also proposed the design and did the actual construction for one. In spite of the accuracy of these forecasts, they were of little value in helping organizations to profit from such forecasts, they were of little value in helping organizations to profit from such forecasts or achieve greater success.

A wide variety of forecasting methods are available to management. These range from the most naïve methods, such as use of the most recent observation as a forecast, to highly complex approaches such as neural nets and econometric systems of simultaneous equations. In addition, the widespread introduction of computers has led to readily available software for applying forecasting techniques. Complementing such software and hardware has been the availability of data describing the state of economic events and natural phenomena. These data in conjunction with organizational statistics and technological know-how provide the base of past information needed for the various forecasting methods.

Forecasting is an integral part of the decision making activities of management. An organization establishes goals and objectives, seeks to predict environmental factors, then selects actions that it hopes will result in attainment of these goals and objectives. The need for forecasting is increasing as management attempts to decrease its dependence on chance and becomes more scientific in dealing with its environment. Some of the areas in which forecasting currently plays an important role are:

Scheduling: Efficient use of resources requires the scheduling of production, transportation, cash, and personnel and so on. Forecasts of the level of demand for product, material, labor, financing, or service are an essential input to such scheduling.

Acquiring resources: The lead time for acquiring raw materials, hiring personnel, or buying machinery and equipment can vary from a few days to several years. Forecasting is required to determine future resource requirements.

Determining resource requirements: All organizations must determine what resources they want to have in long term. Such decisions depend on market opportunities, environmental factors and the internal development of financial, human, product and technological resources. These determinations all require good forecasts and managers who can interpret the predictions and make appropriate decisions.

Although there are many different areas requiring forecasts, the preceding three categories are typical of the short, medium, and long term forecasting requirements of today's organizations. This range of needs requires that a company develop multiple approaches to predicting uncertain events and build up a system for forecasting.

A forecasting system must establish linkages among forecasts made by different management areas. There is a high degree of interdependence among the forecasts of various divisions or departments which cannot be ignored if forecasting is to be successful. For example, errors in sales projections can trigger a series of reactions affecting budget forecasts, operating expenses, cash flows, inventory levels, pricing, and so on. Similarly, budgeting errors in projecting the development, modernization of equipment, hiring of personnel and advertising expenditures. This in turn, will influence, if not determine, the level of sales, operating costs and cash flows. Clearly there is a strong interdependence among the different forecasting areas in and organization.

3.3 An overview of forecasting techniques:

Forecasting situations vary widely in their time horizons, factors determining actual outcomes, types of data patterns, and many other aspects.

To deal with diverse applications, several techniques have been developed. These falls into two major categories: quantitative and qualitative methods. The below table summarizes this categorization scheme and provides examples of situations that might be addressed by forecasting methods in these categories.

Quantitative: *Sufficient quantitative information is available.* *Time series:* predicting the continuation of historical patterns such as the growth in sales or gross national product. *Explanatory:* understanding how explanatory variables such as prices and advertising affect sales.
Qualitative: *Little or no quantitative information is available, but sufficient qualitative knowledge exists.* Predicting the speed of telecommunications around the year 2020 Forecasting how a large increase in oil prices will affect the consumption of oil
Unpredictable: *Little or no information is available.* Predicting the effects of interplanetary travel. Predicting the discovery of a new, very cheap form of energy that produces no pollution.

Categories of forecasting methods and examples of their applications

Quantitative forecasting can be applied when three conditions exist:

1. Information about the past is available
2. This information can be quantified in the form of numerical data
3. It can be assumed that some aspects of the past pattern will continue into the future

Quantitative forecasting technique varies considerably, having been developed by diverse disciplines for different purposes. Each has its own properties, accuracies and costs that must be considered in choosing a specific method. Quantitative forecasting procedures fall on a continuum between two extremes: intuitive or ad hoc methods and formal quantitative methods based on statistical principles. The first type is based on empirical experience that varies widely from business to business, product to product and forecaster to forecaster. Intuitive methods are simple and easy to use but not always as accurate as formal quantitative methods. Also, they usually give little or no information about the accuracy of the forecast. Because of these limitations, their use has declined as formal methods have gained in popularity. May businesses still use these methods, either because they do not know about simple formal methods or because they prefer a judgmental approach to forecasting instead of more objective approaches?

Quantitative forecasting technique varies considerably, having been developed by diverse disciplines for different purposes. Each has its own properties, accuracies and costs that must be considered in choosing a specific method. Quantitative forecasting procedures fall on a continuum between two extremes: intuitive or ad hoc methods and formal quantitative methods based on statistical principles. The first type is based on empirical experience that varies widely from business to business, product to product and forecaster to forecaster. Intuitive methods are simple and easy to use but not always as accurate as formal quantitative methods. Also, they usually give little or no information about the accuracy of the forecast. Because of these limitations, their use has declined as formal methods have gained in popularity. May businesses still use these methods, either because they do not know about simple formal methods or because they prefer a judgmental approach to forecasting instead of more objective approaches?

There are several formal methods, often requiring limited historical data, that are inexpensive and easy to use and that can be applied in a mechanical manner. These methods are useful when forecasts are needed for a large number of items and when forecasting errors on a single item will not be extremely costly.Persons unfamiliar with quantitative forecasting methods often think that the past cannot describe the future accurately because everything is constantly changing. After some familiarity with data forecasting techniques, however, it becomes clear that although nothing remains exactly the same, some aspects of history do repeat themselves in a sense. Application of the right methods can often identify the relationship between the variable to be forecasted and time itself, making improved forecasting possible.

3.4 Qualitative forecasting:

Qualitative forecasting methods, on the other hand, do not require data in the same manner as quantitative forecasting methods. The inputs required depend on the specific method and are mainly the product of judgment and accumulated knowledge. Qualitative approaches often require inputs form a number of specially trained people.

It is more difficult to measure the usefulness of qualitative forecasts. They are used mainly to provide hints, to aid the planner, and to supplement quantitative forecasts, rather than to provide a specific numerical forecast. Because of their nature and cost, they are used almost exclusively for medium and long range situations such as formulating strategy, developing new products and technologies and developing long range plans.

Qualitative methods can be used successfully in conjunction with quantitative methods in such areas as product development, capital expenditures, goal and strategy formulation, and mergers by even medium and small organizations.

3.5 The basic steps in forecasting tasks:

3.5.1 *Problem definition:* The definition of the problem is sometimes the most difficult aspect of the forecaster's task. It involves developing a deep understanding of how the forecasts will be used, who requires the forecasts, and how the forecasting function firs within the organization. It is worth spending time talking to everyone who will be involved in collecting data, maintaining databases, and using the forecasts for future planning.

3.5.2 *Gathering information*: There are always at least two kinds of information available: (a). statistical and (b). The accumulated judgment and expertise of key personnel. Both kinds of information must be tapped.

It is necessary to collect historical data of the items of interest. We can use the historical data to construct a model which can be used for forecasting.

3.5.3 Preliminary (exploratory) analysis: What do the data tell us? We start by graphing the data for visual inspection. Then we compute some simple descriptive statistics e.g., mean, standard deviation, minimum, maximum, percentile associated with each set of data. Where more than one series of historical data is available and relevant, we can produce scatter plots of each pair of series and related descriptive statistics (correlation). Another useful tool is decomposition analysis to check the relative strengths of trend, seasonality, cycles, and to identify unusual data points.

The purpose in all cases at this stage is to get a feel for the data. Are there consistent patterns? Is there a significant trend? Is seasonality important? Is there evidence of the presence of business cycles? Are there any outliers (extreme points) in the data that need to be explained by those with expert knowledge? How strong are the relationships among the variables available for analysis?

Such preliminary analyses will help suggest a class of quantitative models that might be useful in the forecasting assignment.

3.5.4 Choosing and fitting models: This step involves choosing and fitting several quantitative forecasting models. Each model is based on a set of assumptions and usually involves one or more parameters which must be "fitted" using the known historical data. When forecasting the long term, a less formal approach is often better. This can involve identifying and extrapolating mega trends going back in time, using analogies, and constructing scenarios to consider future possibilities.

3.5.5 Using and evaluating a forecasting model: Once a model has been selected judiciously and its parameters estimated appropriately, the model is to be used to make forecasts, and the users of the forecasts will be evaluating the pros and cons of the model as time progresses. A forecasting assignment is not complete when the model has been fitted to the known data. The performance of the model can only be properly evaluated after the data for the forecast period have become available.

In addition the accuracy of future forecasts is not the only criterion for assessing the success of a forecasting assignment. A successful forecasting assignment will usually also be a stimulus to action within the organization. If the forecasts suggest a gloomy picture ahead, then management will do its best to try to change the scenario so that the gloomy forecast will not come true. If the forecasts suggest a positive future, then the management will work hard to make that come true. In general, forecasts act as new information and management must incorporate such information into its basic objective to enhance the likelihood of a favorable outcome. Implementing forecasting is often at least as important as the forecasts themselves.

3.6. DATA MINING TECHNIQUES USED FOR PREDICTION AND CLASSIFICATION:

Task	Microsoft algorithms to use
Predicting a discrete attribute. For example, to predict whether the recipient of a targeted mailing campaign will buy a product.	Decision Trees Algorithm Naive Bayes Algorithm Clustering Algorithm Neural Network Algorithm (SSAS)
Predicting a continuous attribute. For example, to forecast next year's sales.	Decision Trees Algorithm Time Series Algorithm
Predicting a sequence. For example, to perform a click stream analysis of a company's Web site.	Sequence Clustering Algorithm

predict one or more discrete variables, based on the other attributes in the dataset. An example of a classification algorithm is the Microsoft Decision Trees Algorithm.

3.6.1 Microsoft Decision Trees Algorithm:

The Microsoft Decision Trees algorithm is a classification and regression algorithm used in predictive modeling of both discrete and continuous attributes.

For discrete attributes, the algorithm makes predictions based on the relationships between input columns in a dataset. It uses the values, known as states, of those columns to predict the states of a column that you designate as predictable. Specifically, the algorithm identifies the input columns that are correlated with the predictable column. For example, in a scenario to predict which customers are likely to purchase a bicycle, if nine out of ten younger customers buy a bicycle, but only two out of ten older customers do so, the algorithm infers that age is a good predictor of bicycle purchase. The decision tree makes predictions based on this tendency toward a particular outcome.

For continuous attributes, the algorithm uses linear regression to determine where a decision tree splits. If more than one column is set to predictable, or if the input data contains a nested table that is set to predictable, the algorithm builds a separate decision tree for each predictable column

3.6.2 Microsoft Naive Baye's Algorithm:

The Microsoft Naive Bayes algorithm is a classification algorithm used in predictive modeling. The name Naive Bayes derives from the fact that the algorithm uses Bayes theorem but does not take into account dependencies that may exist, and therefore its assumptions are said to be naive.

This algorithm is less computationally intense than other Microsoft algorithms, and therefore is useful for quickly generating mining models to discover relationships between input columns and predictable columns. You can use this algorithm to do initial explorations of data, and then later you can apply the results to create additional mining models with other algorithms that are more computationally intense and more accurate.

3.6.3Microsoft Neural Network Algorithm:

Neural Network algorithm combines each possible state of the input attribute with each possible state of the predictable attribute, and uses the training data to calculate probabilities. You can later use these probabilities for classification or regression, and to predict an outcome of the predicted attribute, based on the input attributes.

A mining model that is constructed with the Microsoft Neural Network algorithm can contain multiple networks, depending on the number of columns that are used for both input and prediction, or that are used only for prediction. The number of networks that a single mining model contains depends on the number of states that are contained by the input columns and predictable columns that the mining model uses.

3.6.4 Microsoft Clustering Algorithm

The Microsoft Clustering algorithm is a segmentation algorithm. The algorithm uses iterative techniques to group cases in a dataset into clusters that contain similar characteristics. These groupings are useful for exploring data, identifying anomalies in the data, and creating predictions.

Clustering models identify relationships in a dataset that you might not logically derive through casual observation. The clustering algorithm differs from other data mining algorithms, such as the Microsoft Decision Trees algorithm, in that you do not have to designate a predictable column to be able to build a clustering model. The clustering algorithm trains the model strictly from the relationships that exist in the data and from the clusters that the algorithm identifies.

3.6.5 Microsoft Sequence Clustering Algorithm

The Microsoft Sequence Clustering algorithm is a sequence analysis algorithm. You can use this algorithm to explore data that contains events that can be linked by following paths, or *sequences*. The algorithm finds the most common sequences by grouping, or clustering, sequences that are identical. This algorithm is similar in many ways to the Microsoft Clustering algorithm. However, instead of finding clusters of cases that contain similar attributes, the Microsoft Sequence Clustering algorithm finds clusters of cases that contain similar paths in a sequence.

3.6.6 Microsoft Time Series Algorithm

The Microsoft Time Series algorithm provides regression algorithms that are optimized for the forecasting of continuous values, such as product sales, over time. Whereas other Microsoft algorithms, such as decision trees, require additional columns of new information as input to predict a trend, a time series model does not. A time series model can predict trends based only on the original dataset that is used to create the model. You can also add new data to the model when you make a prediction and automatically incorporate the new data in the trend analysis.

An important feature is that it can perform cross prediction. If you train the algorithm with two separate, but related, series, you can use the resulting model to predict the outcome of one series based on the behavior of the other series. For example, the observed sales of one product can influence the forecasted sales of another product. Cross prediction is also useful for creating a general model that can be applied to multiple series.

4. *Statistical Techniques Used For Prediction and Classification:*

4.1. Moving Average Method:

A moving average forecast uses a number of most recent historical actual data values to generate a forecast. The moving average for ‘n’ number of periods in the moving average is calculated as :

$$\text{Moving average} = \frac{\sum \text{demand in previous n periods}}{N}$$

N may be 3,4,5 or 6 periods for 3,4,5 or 6 period moving average.

The “simple moving average method” is used to estimate the average of a demand time series and remove the effects of random fluctuation. It is most useful when demand has no pronounced trend or seasonal fluctuation.

In the weighted moving average method each historical demand in the moving average can have its own weight and the sum of the weight equals to one. For *example,* in a 3 period weighted moving average model, the most recent period might be assigned a weight of 0.50, the second most recent period might be assigned a weight of 0.30 and the third most recent period with a weight of 0.20.

Then forecast,

$$F_{t+1} = \frac{0.5\,D_t + 0.3\,D_{t-1} + 0.2\,D_{t-2}}{\text{Sumof weights } (0.5+0.3+0.2)}$$

4.2 Exponential smoothing method:
It is a sophisticated weighted moving average method that is still relatively easy to understand and use. It requires only three items of data: this periods forecast, the actual demand for this period and α which is referred to as smoothing constant and having a value between 0 and 1. The formula used is
Next periods forecast = This period forecast + (α*(This periods actual demand-This periods forecast))
Selecting a smoothing constant is basically a matter of judgment or trial and error. Commonly used values of α ranges from 0.05 to 0.5.

4.3 Trend projection method:

The trend component of a time series reflects the effect of any long term factors on the series. Analysis of trend involves developing an equation that will suitably describe trend. The trend component may be linear or may not.

Trend equation: A linear trend can be expressed as

$Y = a+bx$

x=specified number of time periods from x=0

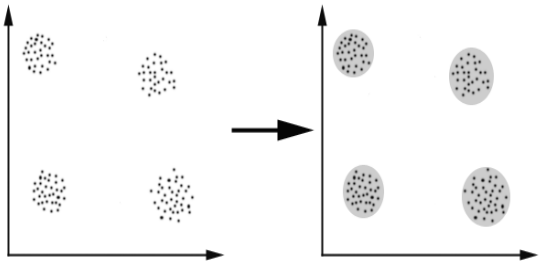
y= forecast for period x
a=value of yt at x=0
b=slope of the straight line

The coefficient of the equation for the trend line, a and b can be computed from historical data using these two equations.

$$b = \frac{n\sum xy - \sum x\sum y}{n\sum x^2 - (\sum x)^2}$$

$$a = \frac{\sum y - b\sum x}{n}$$

In linear regression, one dependent variable is related to one or more independent variables by a linear equation. The dependent variable is demand for a product and the independent variable such as advertising expenditure or new housing are assumed to affect the dependent variable and thereby cause the results observed in the past. In linear regression, one dependent variable is related to one or more independent variables by a linear equation. Whereas in simple linear regression model the dependent variable is a function of one independent variable. The main objective of linear regression is to determine the values of the coefficients in order to minimize the sum of the squared deviations of the actual data points from the graphed straight line.
4.4 Clustering:
It is one of the most unsupervised learning problems. It deals with finding a structure from the collection of unlabeled data. It is the process of organizing objects into groups whose members are similar in some way. Therefore it is a collection of objects which are similar between them and are dissimilar to the objects belonging to other clusters. A simple graphical example of clustering is as follows:



Here we identified four clusters from which the data can be divided; the similar criterion is distance: two or more objects belong to the same cluster if they are close according to given distance this is called as *distance-based clustering.* Another kind of clustering is *conceptual clustering:* two or more objects belong to the same cluster if this one defines a concept common to all objects. Here objects are grouped according to their fit to descriptive concepts not according to simple similarity measures.

Reference:

1. Friedman, J.H. 1997. Data Mining and Statistics. Proceedings of the 29th Symposium on the Interface: Computing Science and Statistics, May 1997, Houston, Texas
2. http://www.knowledgetechnologies.org/proceedings/presentations/treloar/nathantreloar.ppt
3. http://www.amstat.org/publications/tas/hand.pdf
4. Hannu Toivonen Sampling Large Database for Association Rules, Proceedings of the 22nd VLDB Conference Mumbai, India 1996.
5. Friedman, N. and Goldszmidt, M. 1996. Building classifiers using Bayesian networks. In Proceedings AAAI-96 Thirteenth National Conference on Artificial Intelligence, Portland, OR, Menlo Park, CA: AAAI Press, pp. 1277-284.
6. Bernardo, J. and Smith, A. 1994. Bayesian Theory. New York: John Wiley and Sons.
7. Heckerman, D., Geiger, D., and Chickering, D. 1995a. Learning Bayesian networks: The combination of knowledge and statistical data. Machine Learning, 20:197-43.