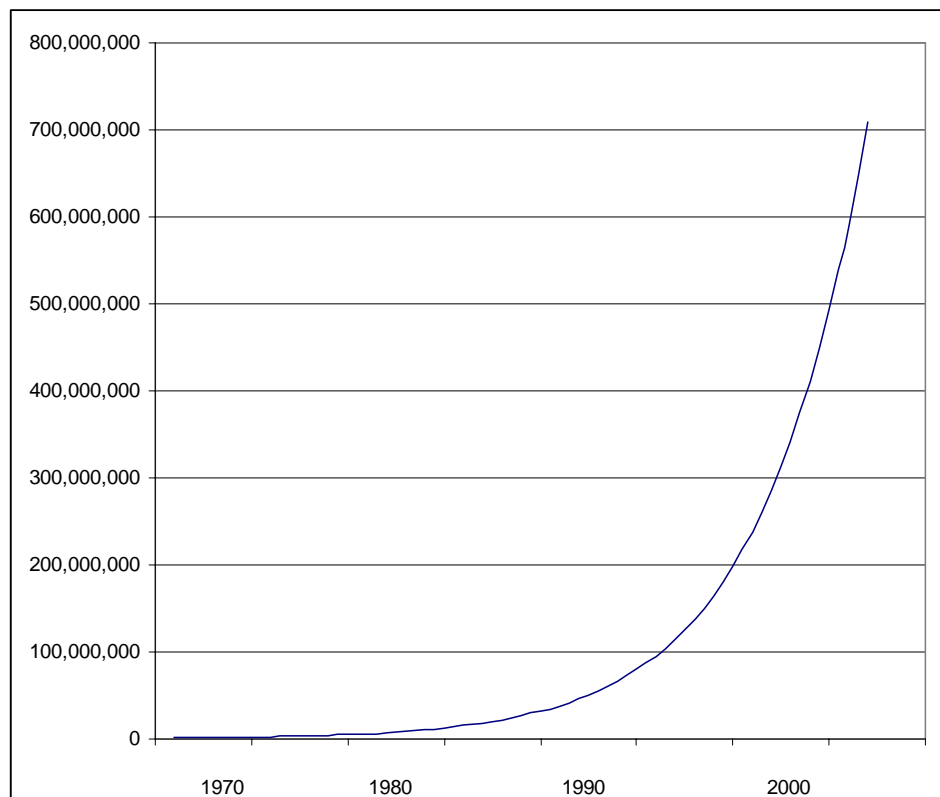


Data Mining Introduction

Every day more and more data becomes available to us at less and less cost. It has been estimated that the amount of information doubles every 20 months! A hypothetical 1 megabyte database growth from 1970 is shown in figure 1. This is a highly non-linear explosion of information. Today, the Walmart retail store chain captures 20 million transactions per day in a 10 terabyte (10×10^{12}) database. How to make the best use of it? How do we turn these large data storage fortresses into useful information? To answer these and other related database related questions we turn to the field of data mining. This ever expanding field can help your company's lean-sigma project deployment or your lean six-sigma project obtain useful information from the vast storage of data you already have at your disposal.

Figure 1
Simulated growth of a 1 megabyte knowledge base from 1970



What is Data Mining?

Data mining is often described as “obtaining desired, useful information from large data storage areas”. It often comprises statistical and logical techniques into hybrid approaches uniquely suited to the task of prying information from thousands and thousands of data records and information fields. Often times the definitions of data mining are now so broad that most any type of data analysis will fit into the topic. In this article we will focus on the most widely used approaches and how they fit into Lean-Sigma.

Often data mining is appropriately classified by the objectives for your specific effort. These objectives can be grouped according to the following broad areas:

- **Association:**
 - Which traits tie to reject mortgage applications?
- **Reduction:**
 - Which variables to process further for information?
- **Exploration:**
 - What messages can be learned here?
- **Classification:**
 - Which transactions are “good” or defect free?
- **Prediction:**
 - Which customers will buy or not in key areas?

The first three areas, **Association**, **Reduction** and **Exploration** are often grouped into a broader title of **Unsupervised Learning**. The goal of Unsupervised Learning is to create knowledge by observation and discovery. The last two groups, **Classification** and **Prediction**, fall into a broader grouping called **Supervised Learning**. In Supervised Learning, the goal is to create through training and practice an algorithm or series of algorithms that classify the data and make predictions about the future or new incoming data.

Let’s look at the individual areas, one at a time. When pursuing **Association**, we are attempting to identify the variables and their specific values or ranges that are correlated with and hopefully causal to an outcome or series of events. This is a lot like what we do in Lean Sigma during the analysis phase of a project, when we analyze the results of passive observation or a Multi-Vari study. When our data-mining goal is **Reduction**, we are attempting to screen out or group together ranges of variables that are appropriate to our objectives or issues. One might group transactions into certain types, product sales into groups, customers by size, and suppliers by geographic or other similarities before proceeding to explore further. In data **Exploration**, we are sifting through the data, learning about the variables, ranges, and pre-qualifying the ranges to screen out bad data points. We may also be pre-processing the data so that algorithm development can proceed more smoothly. In data mining **Classification**, we are developing a rule set to classify transactions or incoming data into two or more groups so that appropriate decisions or further appropriate analysis may take place. Here a past data set is used to create the classification rules, and at least one more data set is used to test the classification approach. Often more than one set is used so that the rule set may be developed or optimized under differing conditions.

Data Mining Techniques

When we consider the various techniques currently available in the software, there are several primary types to choose from. In the **Unsupervised Learning** areas, all the tools we use in Lean Sigma apply. The approach of Practical-

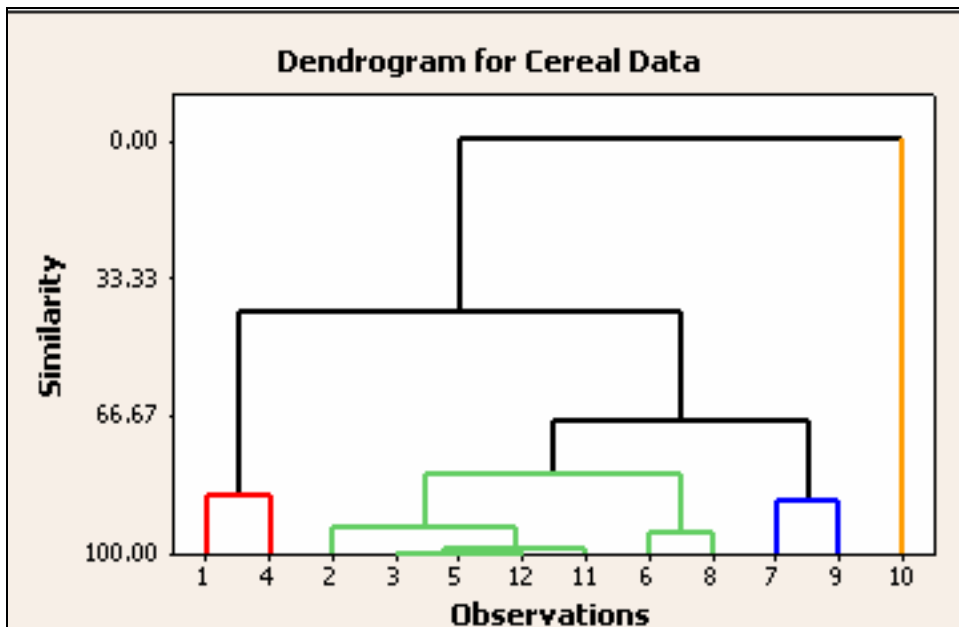
Graphical and then Analytical or Statistical still holds true for Unsupervised Learning goals. The Lean Sigma approaches developed for the DMAIC roadmap in the Define Measure and Analyze phases fit the bill nicely for Unsupervised Learning. The advanced graphics offered in most of the standard software packages will make this data mining work fairly accessible to most any Lean Sigma practitioner. In **Supervised Learning**, we are concerned with three additional approaches beyond the tried and true approach of Multiple Linear Regression, which should not be overlooked for continuous variables!

The first approach in Supervised learning is broadly known as clustering. This is a form used in Multivariate Statistics, and seeks to group data based on specific likeness or differencing criteria. This approach seeks to build a classification algorithm with no assumptions about the $Y=f(x)$ function. In a small example of 12 cereal brands, grouped according likeness in 5 different attributes, an abbreviated clustering output from Minitab might look like this:

Figure 2 Sample Minitab Clustering Output

Final Partition: Number of clusters: 4

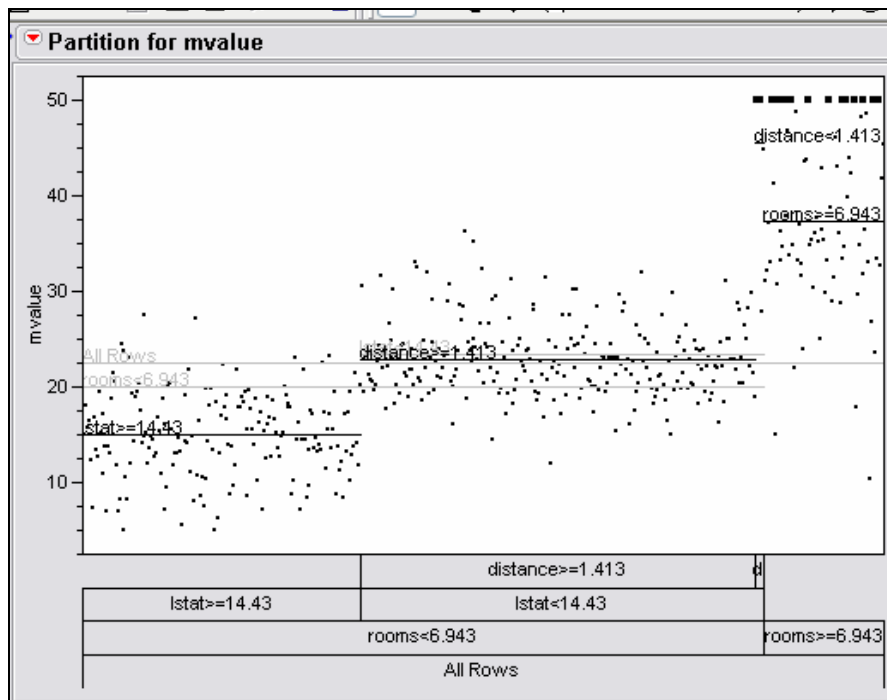
	# Obs.	Within Cluster SS	Avg. Dist ...	Max Dist ...
Cluster1	2	2.48505	1.11469	1.11469
Cluster2	7	8.99868	1.04259	1.76922
Cluster3	2	2.27987	1.06768	1.06768
Cluster4	1	0.00000	0.00000	0.00000



Another significant approach used in Supervised Learning is that of Classification and Regression Trees (CART) or sometimes called partitioning. If you had to choose only one approach to live with, this one generally gets the nod for elegance of use and ease of interpretation by the recipients. In this approach, we seek to maximize the homogeneity within a rectangle by choosing the best independent variable and the best level of that variable to split by. Each successive split creates another branch in the tree. We then prune according to validation data.

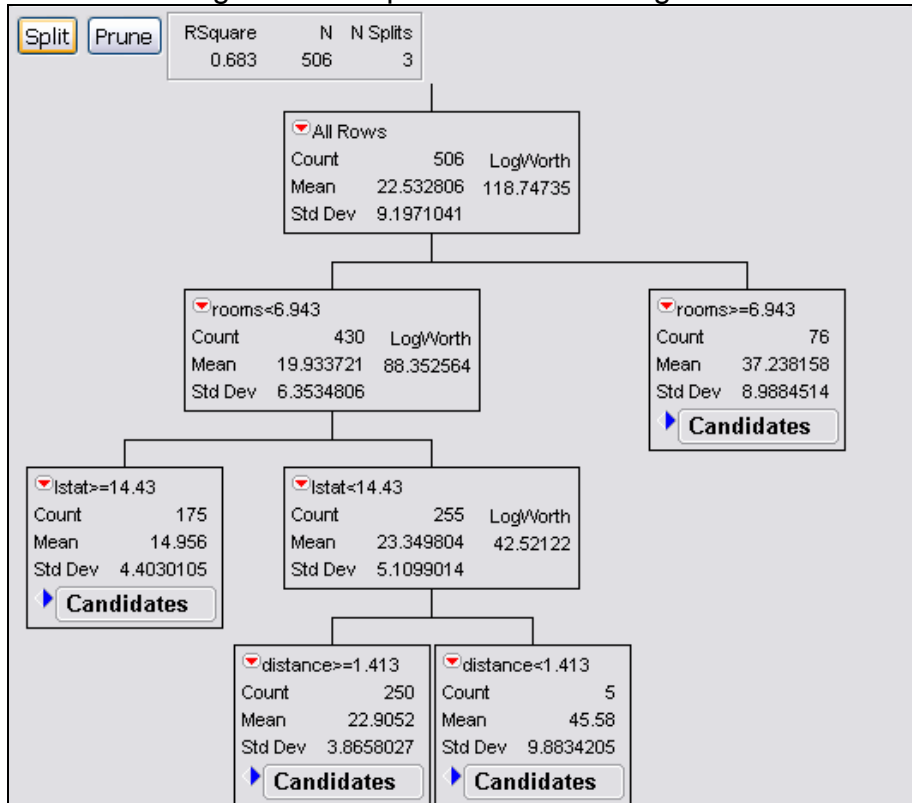
A sample output from JMP's Partitioning features is shown in Figures 3 and 4 for an old data set of the median value on Boston Housing residences versus several potential prediction and classification variables for tax assessment:

Figure 3 – Sample JMP Partitioning Output



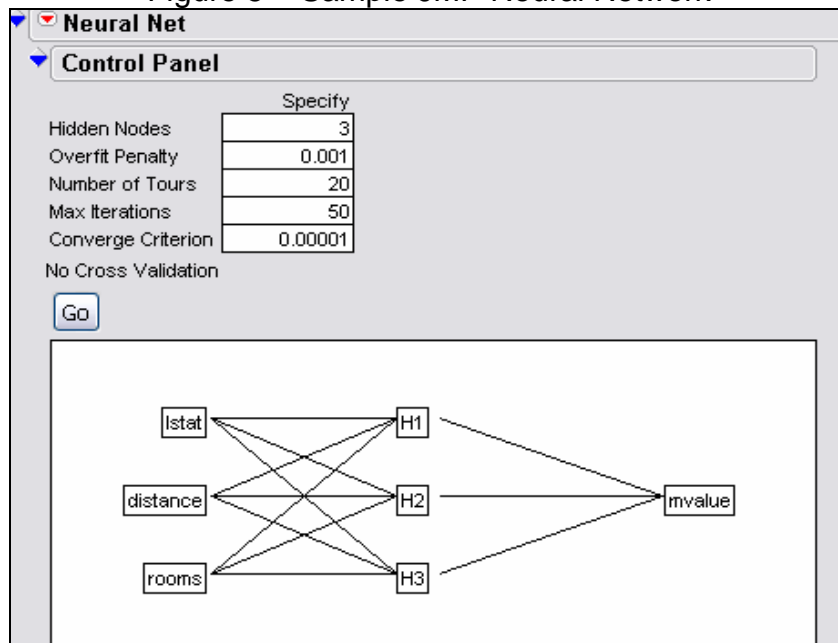
Here the data has been partitioned according to first the average number of rooms being less or greater than 6.943 and then according to a status percentage of 14.43% and finally according to the weighted distance from 5 employment centers. The tree that matches and defines these splits is shown in Figure 4.

Figure 4. Sample JMP Partitioning Tree



The final broad approach to Supervised Learning in data mining is the use of Neural Networks. In a neural network we are training using weighted sums across an intermediate layer to achieve a desired output. Each of the lines in a neural network has a weighting for summation. Such a network, like the CART approach needs training, optimization, validation and test data before applying the final model.

Figure 5 – Sample JMP Neural Network



Neural Networks have been used in all sorts of artificial intelligence applications, anywhere where a decision on known multiple criteria will be dynamic and made regularly. The advances in neural network applications include: unmanned aerial vehicles, lip-reading, automatic security identification and screening, cyber creature programming and autonomous walking and navigating robots.

Data Mining Techniques and Software

Many of the very familiar aspects of JMP, Minitab and SigmaFlow apply quite well in **Unsupervised Learning**. In most PC-based statistical packages, working with 10 million rows of multi-column data is now possible. Issues with large data sets are in most cases a thing of the past. In **Supervised Learning** however, there may be some software limitations that can be overcome with specialty software. Of the three aforementioned statistical packages used in Lean Sigma, JMP has added all the primary techniques necessary to do supervised learning on a PC. If you are using Minitab or SigmaFlow, you will need another software package or an Excel based add-in called XLMiner, that adds this functionality into MS Excel. However, be cautioned that MS Excel has a limitation to 65,536 rows, which may be incompatible with your data mining goals. You may want to explore the higher end data mining offerings such as Rapid Insight or even mainframe applications like SAS Enterprise Miner.

Data Mining in Lean Sigma Deployments

When expanding or re-invigorating your Lean Sigma deployments, one of the great ways to find new projects is applying data mining techniques to the abundant financial data most companies have available. Depending upon your focus (Growth, cost efficiency, inventory, safety, capital spending, etc.), you can apply the data mining principles to identify, explore and discover opportunities for lean sigma projects. One SBTI client recently trained a top financial executive as a BB. They are still getting 4% bottom line savings on their projects five years into the deployment of Six Sigma !

Data Mining in Lean Sigma Projects

So how do the principles and approaches of data mining relate to DMAIC? If we list the 5 DMAIC stages and their goals, we can relate them to Data Mining:

- | | | |
|---|---|-----------------------|
| <ol style="list-style-type: none"> 1. Define – scope and possibilities 2. Measure – size and breadth of project issues 3. Analyze – potential relationships 4. Improve – develop and test new models 5. Control – test and apply proven model controls | } | Unsupervised Learning |
| <ol style="list-style-type: none"> 4. Improve – develop and test new models 5. Control – test and apply proven model controls | } | Supervised Learning |

The Exploratory, Reduction and Association aspects of data mining align quite nicely with the Define, Measure and Analyze phases of the Lean Sigma Roadmap. Unsupervised Learning Data Mining principles can be integrated with Lean Sigma in these early project stages. The Classification and Prediction oriented aspects of Data Mining align with the Improve and Control phases of

Lean Sigma. The Supervised Learning methods and techniques can be integrated with the DMAIC roadmap in Lean Sigma at the later stages of project work.

Summary

The development of Data Mining tools and techniques in statistical packages now makes these methods available to Lean Sigma Black Belts, and can be integrally applied during the project work. This is particularly valuable in transactional and service design projects where an abundance of data is usually available. Watch for updates to the SBTI Healthcare programs, MBB electives and the Transactional Lean Sigma BB curriculums for inclusion of data mining into the training !

This quarter's best data mining tip – always include a subject matter expert! All the data and analysis in the world can go completely off track without someone who knows the process that generated the data, how it was collected and most importantly, what isn't in the data!

Printed References:

- **Applied Data Mining : Statistical Methods for Business and Industry (Statistics in Practice) (Paperback)**
by [Paolo Giudici](#)
- **Data Mining Techniques : For Marketing, Sales, and Customer Relationship Management (Paperback)**
by [Michael J. A. Berry](#), [Gordon S. Linoff](#)
- **Data Mining Lecture Notes, MIT, Nitin R. Patel, Peter C. Bruce, May 2003**

Web Based References:

1. databases.about.com/od/datamining/a/datamining.htm
2. www.pcc.qub.ac.uk/tec/courses/datamining/stu_notes/dm_book_1.html
3. www.jmp.com
4. www.minitab.com
5. www.sas.com
6. www.resample.com/xlminer/index.shtml
7. www.sigflow.com
8. www.rapidinsightinc.com/