

Introduction

An Overview of Enterprise Miner

Enterprise Miner is a great new product SAS has recently introduced in version 8. It consists of a variety of analytical tools, like neural networks, to support data mining to enhance traditional statistical modeling. *Data mining* is an analytical tool that is used in solving critical business decisions by analyzing enormous amounts of data in order to discover relationships and unknown patterns in the data. Enterprise Miner is a powerful product now available within the SAS software. The Enterprise Miner data mining SEMMA methodology is specifically designed in handling enormous data sets in preparation to subsequent data analysis. In SAS Enterprise Miner, the abbreviation SEMMA stands for Sampling, Exploring, Modifying, Modeling and Assessing large amounts of data. Neural network modeling with regard to the data mining tasks falls under nonlinear regression modeling or classification modeling. In regression modeling, the aim is building a model that will produce values of one variable to be predicted based on a set of known values of other variables. In classification modeling, the difference is that the variable to predict is categorical based on a set of known continuous or categorical variables in the model.

Purpose of Writing this Book

One reason for writing this book is because there is not a tremendous amount of written literature on neural network modeling using SAS Enterprise Miner. The book is a step-by-step approach to neural network modeling using SAS Enterprise Miner 4.3. This book will also get you familiar with the SAS neural network procedure called PROC NEURAL. This book consists of step-by-step instructions in getting started in designing a process flow diagram in SAS Enterprise Miner. The book will also explain the various statements and options to the PROC NEURAL procedure. There are numerous examples in explaining the various complex neural network designs and optimization techniques used in neural network modeling with numerous examples taken from various SAS literature comparing the prediction results between both neural network and traditional regression statistical modeling techniques with an explanation to the SAS modeling results. The book's introductory is a brief overview to traditional regression modeling and the various statistical assumptions that must be satisfied.

Highlights to this Book

Chapter 2 discusses basic model building and the various modeling assumptions that need to be satisfied. These modeling assumptions in order of importance are independence, equal variance and normality in the modeling terms that must be satisfied in both traditional regression modeling and neural network designs. However, it should be noted that some neural network modelers might ignore these extremely important statistical modeling assumptions. This section will explain the various diagnostic statistic used in identifying outliers and influential data points that have a profound effect to the modeling results. And finally explaining the various goodness-of-fit statistics used in determining the best linear combination of input variables among a pool of all possible combination of input variables in the predictive model.

Chapter 3 explains the neural network design and the various configuration settings. The section will first explain a simple perceptron design for a binary-valued target variable. The following section will graphically display the similarities between a multiple linear regression model and a neural network model. Next, we will discuss the neural network designs and the various configurations to the design like the various layers, weights, combination functions, transfer functions, objective or error functions and optimization techniques that are used. The section will explain the various

optimization techniques such as the various line search and grid search techniques. It will be followed by various numerical examples in order to simplify the complexity of the various optimization techniques that are applied in calculating the neural network weight estimates and determining the smallest error to the neural network model. A numerical example of the backpropagation algorithm will be presented that is typically used in a neural network MLP design. Pruning techniques used in pre-processing the model will be discussed leading to the general strategies in interpreting important input variables to the neural network model and the basic ideas in constructing a well-designed neural network model. The section will conclude with a brief summary to the advantages and disadvantages to neural network modeling.

Chapter 4 will explain both the SAS neural network procedure called the PROC NEURAL procedure and the SAS data mining regression procedure called the PROC DMREG procedure and the various option statements to these predictive data-mining procedures. The chapter will then display diagrams of the neural network architecture based on a couple of the modeling comparison examples presented later in the book. Thereby, enabling the reader to graphically understand the neural network configuration between the various layers and the weight estimates associated with these same neural network layers. Followed by SAS output listings from the Enterprise Miner results in order for the reader to understand the process in calculating the predicted values from the neural network model based on the corresponding weight estimates and biases assigned to the various neural network connections. The section will then go through the basic steps in constructing an SAS Enterprise Miner process flow diagram to the various modeling designs.

Chapter 5 discusses the various SAS Enterprise Miner menu options and setting-up the various start-up configuration settings to the Enterprise Miner project. The section will explain the basic steps in constructing Enterprise Miner projects and diagrams with a brief overview to the SAS Enterprise Miner interface.

Chapter 6 will have a brief overview to the data mining SEMMA process. The section will briefly explain the relationship between each of the EM nodes with regard to the SEMMA design. The section will explain the purpose of each one of the nodes in Enterprise Miner. The next section is a general overview and explanation of the configuration settings and results to each one of the EM nodes currently available in SAS Enterprise Miner for data mining analysis.

Chapter 7 will explain the configuration settings and the corresponding results of the Enterprise Miner nodes used in designing the neural network modeling comparisons. That is, explaining each one of the option settings of the Input Data Source node, Data Partition node, Regression node, Neural Network node, SAS Code node, Assessment node and the Reporter node. The section will first explain the purpose of the numerous tabs and the various options or configuration settings within each tab of the corresponding the Enterprise Miner node followed by an explanation of the results that are generated from each one of the nodes.

Chapter 8 will compare the neural network estimates with the various statistical modeling designs such as multiple linear regression estimates, nonlinear regression estimates, logistic regression estimates, time series estimates and discriminant analysis. Each section will consist of a brief summary to the various statistical modeling methods and the associated SAS procedure and option statements needed in producing the modeling results. The various statistical modeling comparison examples are taken from various SAS manuals and literature in order for the reader to cross-reference the statistical results. The process in assessing the fit of both models was done by graphically comparing the modeling estimates and analyzing the various modeling assessment statistics from both models in determining the best fit.

Advantages and Disadvantages to Neural Network Modeling

The general purpose of neural network modeling is to estimate, classify and make predictions. For example, predicting a person's weight based on their age and height to classifying a type of wild flower species based on its petal and sepal length and width. Neural network modeling is typically designed in fitting data with an enormous number of records with numerous predictor variables in the model. That is, data consisting of millions of records with hundreds or even thousands of predictor variables in the nonlinear model. Nowadays, this phenomenon might actually occur with the advancements in technology capable of handling enormous amount of data. Some of the biggest advantages of neural network modeling is its flexibility in modeling a wide variety of statistical models and interpolating extremely complex nonlinear functions. Unlike traditional regression modeling, neural network modeling does not require any distributional assumptions between the input variables and the target variables to the model. That is, neural network model building does not depend on you specifying the exact mathematical functional form between the target variable and the input variables to the model. Neural network modeling can also interpolate many functional forms with extreme accuracy assuming that there is a sufficient number of hidden layer units, an adequate amount of data and a reasonable amount of computational time. In classification problems, neural network modeling can approximate any decision boundary in assigning incoming observations into distinct groups with great precision. Conversely, the disadvantages to neural network modeling is that the model needs a sufficient amount of data to assure convergence to the correct parameter estimates. In neural network modeling, it's very complicated in measuring the importance of the predictor variables or the input variables in the model. Also unlike traditional regression parameter estimates, the weight estimates do not tell you the effect, magnitude or the rate of change in the relationship between the outcome variable and the predictor variables. The reason is because the input variable in the model are associated with both the hidden layer and the output layer weight estimates. Therefore, neural network modeling is designed in producing predictions without the ability in making any interpretations in the relationship between the variables in the model. Similar to nonlinear regression there is no guarantee in the neural network model that the iteration algorithm that is applied will converge to the global minimum in determining the best linear combination of parameter estimates. And currently, there does not seem to be available certain diagnostic statistics such as testing for lack of fit to the neural network model, identifying influential and outlier data points and significant test in the neural network modeling effects.

One reason in considering neural network modeling is that the neural network model might perform better than traditional regression modeling at certain data points of interest, a range of data points of interest or beyond the range of the actual data points. And for classification problems, the neural network design might result in a better modeling performance in classifying the categorical responses. Also, for categorical responses or responses stratified into separate levels, various predictive modeling designs may be performed based on various decision scenarios in order to maximize expected profit or minimize expected loss based on predetermined profit, loss, revenue and cost amounts for each target-specific decision consequence at each target level or accurately determining the appropriate response levels by specifying corresponding prior probabilities that accurately represent the true proportion of occurrences to each level of the categorical response.

Enterprise Miner Ease Of Use

This book is also an overview to neural network model with step-by-step instructions in designing and constructing projects and diagrams for predictive model using SAS Enterprise Miner 4.3. The book will explain the SAS Enterprise Miner prediction results and the NEURAL procedure results. SAS Enterprise Miner is a powerful new module introduced in version 8. But, more than anything else SAS Enterprise Miner is a very easy application to learn and very easy to use. SAS Enterprise

Miner is visual programming with a GUI interface. The power of the SAS Enterprise Miner product is that you do not even need to know SAS programming and have very little statistical expertise in designing an EM project in order to develop a completely comprehensive statistical analysis reporting system. Yet, an expert statistician can adjust the default settings and run the EM SEMMA process flow diagram to their own personal specifications. SAS Enterprise Miner is visual programming with the SAS icons selected from the EM tool palette or the menu bar that are dragged onto a graphical EM diagram workspace. That is, it is as simple as dragging and dropping icons to the EM diagram graphical workspace. The nodes are then connected to one another within a graphical EM diagram workspace. The Enterprise Miner diagram workspace environment looks similar to the desktop in Windows 95, 98 and XP. A dialog box will appear with numerous tabs and subtabs when you open each one of the Enterprise Miner nodes to either specify the various configuration settings or viewing the various results from the nodes. The various option setting that are specified within the corresponding Enterprise Miner node instructs SAS to automatically write code behind the scenes based on the associated procedure. That is, the results that are generated from the various nodes are based on the associated data-mining procedures that are running behind the scenes once you execute the node. Enterprise Miner is very easy to use and can save a tremendous amount of time having to program in SAS. SAS Enterprise Miner has a powerful EM SAS Code node that brings in the capability of SAS programming into the EM SEMMA process through the use of a SAS data step in accessing a wide range of the powerful SAS procedures into the SAS Enterprise Miner process flow diagram. Enterprise Miner is designed to perform exploratory analysis by producing various charts and graphs, descriptive modeling such as cluster analysis and predictive modeling from regression modeling, neural network modeling to classification modeling. In addition to neural network modeling, Enterprise Miner performs a wide variety of data mining analysis from association analysis, link analysis, decision tree analysis, principal component analysis, cluster analysis, kernel smoothing models called memory-based reasoning, average estimation by combining models called ensemble models, two-stage modeling in predicting a categorical-valued variable and a continuous-valued variable in succession, multiple regression modeling, logistic regression modeling and various modeling assessment routines. These nodes are designed to automatically generate various statistical results and graphs that can be redirected to the SAS output window or the SAS graphics window. Enterprise Miner v4.1 can now perform time series analysis and autoregressive time series modeling that will be introduced in later versions of SAS Enterprise Miner.

In SAS v9.1, SAS has two parallel versions of Enterprise Miner 4.3 and Enterprise Miner 5.1. These two releases are very different from a GUI standpoint. The same type of interface in Enterprise Miner 4.3 is similar to the GUI in Enterprise Miner 4.1 from SAS 8.2. However, Enterprise Miner 5.1 is Java based. The previously mentioned algorithms are the same. Although, Enterprise Miner 5.1 has some added tools such as StatExplore, Rule Induction, Path Analysis and Automated networks. The StatExplore tool is designed to compute univariate and bivariate distribution statistics for interval or class variables. Rule Induction is designed in building models by recursively identifying target events. That is particularly useful for modeling rare events. The Path Analysis node uses a new PATH procedure that includes a referrer variable for Web log analysis. Automated networks uses an algorithm for automated MLP network building. The node is designed to select the type of network architecture and the number of activation functions from the four different architectures.

Purpose of the Book

This book is a step-by-step tutorial to SAS Enterprise Miner. The book is also a brief overview to neural network modeling. I hope after reading this book Enterprise Miner will become very easy SAS analytical tool for you to use while at the same time making you feel very comfortable in using the SAS neural network procedure to incorporate in your SAS statistical modeling code.