Predictive Modeling with SAS® Enterprise Miner™

Practical Solutions for Business Applications

Predictive Analytics-Emphasis

- Study the three main predictive modeling tools: Decision Tree, Neural Network, and Regression
- Examine the SAS code generated by each node and show the correspondence between the theory and the results produced by Enterprise Miner.
- Give intuitive explanations of the way that various nodes such as Decision Tree, Neural Network, Regression, and Variable Selection operate and how different options such as Model Selection Criteria and Model Assessment are implemented.

Chapter 2: Getting Started with Predictive Modeling

This chapter introduces you to

- SAS Enterprise Miner 13.2
- Some of the preprocessing and data cleaning tools (nodes) needed for data mining
- Predictive modeling projects

Objectives

- Open Enterprise Miner
- Define a new project
- Create data sources for projects
- Perform data cleaning, exploration, data mining
- Use several nodes
- Carry out predictions using predictive models

Specifically, use data sets to demonstrate how to use the

- Input Data
- Data Partition,
- Filter
- File Import
- Time Series
- Merge
- Append
- StatExplore
- MultiPlot

Specifically, use data sets to demonstrate how to use the

- Graph Explore
- Variable Clustering
- Cluster
- Variable Selection
- Drop
- Replacement
- Impute
- Interactive Binning
- Principal Components
- Transform
- Variables and SAS Code nodes.

SAS Enterprise Miner 13.2

- Chapter 2 discusses a step-by-step approach for SAS Enterprise Miner 13.2
- We will practice how to use SAS Enterprise Miner 13.2 to analyze real life data
- You will be required to suggest possible real life data
- We will also explore the use of existing data

Creating a SAS Data Source

- You must create a data source before you start working on your SAS Enterprise Miner Project.
- After the data source is created, it contains all the information associated with your data—the directory path to the file that contains the data, the name of the data file, the names and measurement scales of the variables in the data set, and the cost and decision matrices and target profiles you specify.
- Profit matrix is also known as decision weights in SAS Enterprise Miner 13.2 which is used in decisions such as assigning a target class to an observation and assessing the models.
- This section shows how a data source is created, covering the essential steps.
- For additional capabilities and features of data source creation, refer to the Help menu in SAS Enterprise Miner.
- SAS Enterprise Miner saves all of this information, or *metadata*, as different data sets in a folder called **Data Sources** in the project directory.

Creating a Process Flow Diagram

- To create a process flow diagram, right-click **Diagrams** in the project panel and click **Create Diagram**.
- You are prompted to enter the name of the diagram in a text box labeled **Diagram Name**.
- After entering a name for your diagram, click **OK**. A blank workspace opens, where you create your process flow diagram.

Sample Nodes

- If you open the **Sample** tab, the tool bar is populated with the icons for the following nodes:
- Append, Data Partition, File Import, Filter, Input Data, Merge, Sample, and Time Series.
- In this section, an overview of some of these nodes are provided, starting with the **Input Data** node.

Tools for Initial Data Exploration

- In this section the following nodes which are useful in predictive modeling projects are discussed:
- **StatExplore**
- **MultiPlot**
- **GraphExplore**
- **Variable Clustering**
- **Cluster Variable Selection**

Tools for Data Modification

This section discuss the following:

Drop Node

Replacement Node

Impute Node

Interactive Binning Node

Principal Components Node

Transform Variables Node

Note: A step-by-step process is given in the text.

Additional useful concepts

Utility Nodes

SAS Code Node

The Type, the Measurement Scale, and the Number of Levels of a Variable

Eigenvalues, Eigenvectors, and Principal Components

Cramer's V

Calculation of Chi-Square Statistic and Cramer's V for a Continuous Input

Case Study-Choose a project to apply the following:

- Creamer's V
- Chi-Squared

Systems of Two Linear Algebraic Equations

- **Review linear algebraic systems**

Consider the system

$$a_{11}x_1 + a_{12}x_2 = b_1,$$
$$a_{21}x_1 + a_{22}x_2 = b_2,$$

In matrix notation, system is **Ax** = **b**, where

$$\boldsymbol{A} = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}, \boldsymbol{x} = \begin{pmatrix} x_{11} \\ x_{21} \end{pmatrix}, \boldsymbol{b} = \begin{pmatrix} b_1 \\ b_2 \end{pmatrix}.$$

Here **A** is a given 2 × 2 matrix, **b** is a given 2 × 1 column vector, and **x** is a 2 × 1 column vector to be determined.

Eigenvalues and Eigenvectors

- **Eigenvalues** (*λ*) of the matrix **A** are the solutions to **Ax** = *λ***x.** The **eigenvector x** corresponding to the eigenvalue *λ* is obtained by solving **Ax** = *λ***x** for **x** for the given *λ*.
- For a 2X2 matrix **Ax** = *λ***x** reduces to

$$\begin{pmatrix} a_{11}-\lambda & a_{12} \\ a_{21} & a_{22}-\lambda \end{pmatrix} x = 0$$

Since det(**A**-λ**I**)=0, get

$$\det(\mathbf{A}-\lambda\mathbf{I}) = \begin{vmatrix} a_{11}-\lambda & a_{12} \\ a_{21} & a_{22}-\lambda \end{vmatrix} = 0.$$

Characteristic equation

The **characteristic equation** of the matrix **A** is

$\lambda^2 - (a_{11} + a_{22})\lambda + a_{11}a_{22} - a_{12}a_{21} = 0.$

Solutions determine the eigenvalues.

The two solutions, the eigenvalues λ_1 and λ_2, may be real and different, real and equal, or complex conjugates.

Summary for a Two-Dimensional Linear Algebraic System

- Matrix notation for a linear algebraic system of two equations in two unknowns is $\mathbf{Ax} = \mathbf{b}$.

1. If $\det \mathbf{A} \neq 0$, the unique solution of $\mathbf{Ax} = \mathbf{b}$ is $\mathbf{x} = \mathbf{A}^{-1}\mathbf{b}$.
2. If $\det \mathbf{A} = 0$, $\mathbf{Ax} = \mathbf{b}$ may have (i) no solution, or (ii) a straight line of solutions in the plane; in particular, if $\mathbf{b} = \mathbf{0}$ and $\mathbf{A} \neq \mathbf{0}$, the solution set is a straight line passing through the origin.

- **The eigenvalue problem:** $(\mathbf{A} - \lambda\mathbf{I})\mathbf{x} = \mathbf{0}$. The **eigenvalues** of $\mathbf{A}$ are solutions of the **characteristic equation** $\det(\mathbf{A} - \lambda\mathbf{I}) = 0$. An **eigenvector** for the eigenvalue λ is a nonzero solution of $(\mathbf{A} - \lambda\mathbf{I})\mathbf{x} = \mathbf{0}$. Eigenvalues may be real and different, real and equal, or complex conjugates.

Examples

Find the eigenvalues and eigenvectors of the matrix **A.**

1. $\mathbf{A} = \begin{pmatrix} 1 & 1 \\ 4 & 1 \end{pmatrix}$

2. $\mathbf{A} = \begin{pmatrix} -\frac{1}{2} & 1 \\ -1 & -\frac{1}{2} \end{pmatrix}$

3. $\mathbf{A} = \begin{pmatrix} 1 & -1 \\ 1 & 3 \end{pmatrix}$