Advanced Modelling Techniques in SAS Enterprise Miner

Dr Iain Brown, Senior Analytics Specialist Consultant, SAS UK & Ireland
Agenda

• SAS Presents – Thursday 11\textsuperscript{th} June 2015 – 15:45
• Advanced Modelling Techniques in SAS Enterprise Miner

• The session looks at:
  - Supervised and Unsupervised Modelling
  - Classification and Prediction Techniques
  - Tree Based Learners
The Analytics Lifecycle

**BUSINESS MANAGER**
Domain Expert
Makes Decisions
Evaluates Processes and ROI

**IT SYSTEMS / MANAGEMENT**
Model Validation
Model Deployment
Model Monitoring
Data Preparation

**BUSINESS ANALYST**
Data Exploration
Data Visualization
Report Creation

**DATA MINER / STATISTICIAN**
Exploratory Analysis
Descriptive Segmentation
Predictive Modeling

**IDENTIFY / FORMULATE PROBLEM**

**DATA PREPARATION**

**DATA EXPLORATION**

**TRANSFORM & SELECT**

**BUILD MODEL**

**VALIDATE MODEL**

**DEPLOY MODEL**

**EVALUATE / MONITOR RESULTS**
The Analytics Lifecycle

**BUSINESS MANAGER**
- Domain Expert
- Makes Decisions
- Evaluates Processes and ROI

**BUSINESS ANALYST**
- Data Exploration
- Data Visualization
- Report Creation

**IT SYSTEMS / MANAGEMENT**
- Model Validation
- Model Deployment
- Model Monitoring
- Data Preparation

**DATA MINER / STATISTICIAN**
- Exploratory Analysis
- Descriptive Segmentation
- Predictive Modeling
The Analytics Lifecycle

**IDENTIFY / FORMULATE PROBLEM**

**DATA PREPARATION**

**TRANSFORM & SELECT**

**BUILD MODEL**

**DEPLOY MODEL**

**VALIDATE MODEL**

**EVALUATE / MONITOR RESULTS**

**DETA EXPLORATION**

**BUSINESS MANAGER**
Domain Expert
Makes Decisions
Evaluates Processes and ROI

**BUSINESS ANALYST**
Data Exploration
Data Visualization
Report Creation

**IT SYSTEMS / MANAGEMENT**
Model Validation
Model Deployment
Model Monitoring
Data Preparation

**DATA MINER / STATISTICIAN**
Exploratory Analysis
Descriptive Segmentation
Predictive Modeling

Domain Expert Evaluates Processes and ROI
Supervised and Unsupervised Modelling
Taxonomy

Machine Learning

Supervised
- Classification
- Prediction

Unsupervised
- Clustering
- Affinity Analysis
Learning Methods

Supervised:

- Discover patterns in the data that relate attributes to labels.
- Patterns are used to predict the values of the label in future data instances.

Unsupervised:

- The data have no label attribute.
- Goal is to explore the data to find some intrinsic structures in them.
## Supervised Learning (Classification & Prediction)

<table>
<thead>
<tr>
<th>Method</th>
<th>Other Methods</th>
</tr>
</thead>
<tbody>
<tr>
<td>Logistic Regression</td>
<td>Neural Networks</td>
</tr>
<tr>
<td>Decision Trees, CART</td>
<td>Nonlinear SVMs</td>
</tr>
<tr>
<td>Decision Trees, CHAID</td>
<td>Bayesian Networks</td>
</tr>
<tr>
<td>Gradient Boosting</td>
<td></td>
</tr>
<tr>
<td>Random Forests</td>
<td></td>
</tr>
<tr>
<td>kth Nearest Neighbor</td>
<td></td>
</tr>
<tr>
<td>Regression, least square</td>
<td></td>
</tr>
<tr>
<td>Generalized Linear Models</td>
<td></td>
</tr>
<tr>
<td>LASSO, LAR</td>
<td></td>
</tr>
<tr>
<td>Splines, MARS</td>
<td></td>
</tr>
</tbody>
</table>
Unsupervised Learning

K-means
Fuzzy K-means
Hierarchical Clustering
Vector Quantization
Multidimensional Scaling
Principal Components
Nonnegative Matrix Factorization
Assocations, Apriori
Nonnegative Matrix Factorization
Regression

- Linear
- Logistic

- Computes a forward stepwise least-squares regression
- Optionally computes all 2-way interactions of classification variables
- Optionally uses AOV16 variables to identify non-linear relationships between interval variables and the target variable.
- Optionally uses group variables to reduce the number of levels of classification variables.
Generalised Linear Models

- Uses the high-performance HPGENSELECT procedure to fit a generalized linear model in a threaded or distributed computing environment.
- Several response probability distributions and link functions are available.
- Provides model selection methods.
Neural Networks

- Non-linear relationship between inputs and output
- Prediction more important than ease of explaining model
- Requires a lot of training data
Support Vector Machines

- Enables the creation of linear and non-linear support vector machine models.
- Constructs separating hyperplanes that maximize the margin between two classes.
- Enables the use a variety of kernels: linear, polynomial, radial basis function, and sigmoid function. The node also provides Interior point and active set optimization methods.
Ensemble

• Creates new models by combining the posterior probabilities (for class targets) or the predicted values (for interval targets) from multiple predecessor models.

• 3 Methods
  • Average
  • Maximum
  • Voting
Model Import

- Importing already scored records/cases
- Importing registered SAS Model Package
- Importing SAS Score Code

- Reads all model details from Metadata Repository
- Applies models to new data and generates all fit statistics
- Compatible with model selection tools
- Useful for sharing models with other users
- Useful testing old models with updated data
Tree Based Learners
SAS EM Tree Algorithms

- 3 key tree based learning algorithms:
  1. Decision Trees
  2. Gradient Boosting
  3. Random Forests
Decision Trees
Decision Trees

- Classify observations based on the values of nominal, binary, or ordinal targets
- Predict outcomes for interval targets
- Easy to interpret
- Interactive Trees available
- CART, CHAID, C4.5 approximate
Modelling Algorithms

- Sequential ensemble of many trees
- Extremely good predictions
- Very effective at variable selection
Gradient Boosting

• Approach that resamples the analysis data set several times to generate results that form a weighted average of the re-sampled data set.
• Tree boosting creates a series of decision trees which together form a single predictive model.
• A tree in the series is fit to the residual of the prediction from the earlier trees in the series.
• The residual is defined in terms of the derivative of a loss function.
• The successive samples are adjusted to accommodate previously computed inaccuracies.
Gradient Boosting

- A gradient boosting tree with an interval target (Median Home Value, MEDV):
  - Number of iterations, M=2; Maximum tree depth = 1
  - Resulting model is combination of two decision trees (T1 and T2) each with 2 leaves.
  - The value of 22.275 is the mean MEDV, while P_MEDV is the predicted value
  - An observation with LSTAT = 6 and RM = 5 would have a P_MEDV value of 22.275 + .95 - .17 = 23.055
Random Forest Node

What is a Random Forest?

At each node:
- choose some small subset of variables at random
- find a variable (and a value for that variable) which optimizes the split
HPForest

- HP node provides increased processing speed
- Random Forest ensemble methodology
  - Samples without replacement
  - Random selection of variables for each tree
  - Uses measures of association to select variable
  - Creates a prediction that is aggregated across the value in the leaf of each tree
Summary
Summary

• EM supports a variety of both supervised and unsupervised modelling algorithms

• Linear / Non-Linear modelling

• Benefits from Tree based learning algorithms include:
  • Interoperability
  • Model performance
  • Outliers/ Missing Values