FINDING THE RIGHT-SIZED MODEL WITH ENTERPRISE MINER: MODEL TUNING AND GENERALIZATION

Lorne Rothman, PhD, PStat
Principal Statistician, SAS Institute (Canada) Inc.
Lorne.Rothman@sas.com
GENERALIZATION AND THE RIGHT SIZED MODEL: TOPICS

- Predictive modeling and empirical validation
- Model tuning with Validation Data
  - Regression, Decision Trees, Neural Networks
- Penalized measures of fit in Regression
- Ensemble models
Predictive modeling and empirical validation

Model tuning with Validation Data
- Regression, Decision Trees, Neural Networks

Penalized measures of fit in Regression

Ensemble Models
MODEL BUILDING

Target

Input Variables

\[ y, x_1, x_2, x_3, x_4, x_5, x_6, \ldots, x_k \]

Cases

1
2
3
4
5
\vdots
n
EXPLANATORY VS PREDICTIVE MODELING

Traditional Explanatory Modeling
- Smaller samples
- Small number of model inputs selected based on theory
- Strong emphasis on probability theory (p-values) to make inferences from samples to the population
- Purpose is explanation
- Models are fit and assessed on the same data.

Predictive Modeling
- Larger samples
- Large number of inputs selected based upon model performance
- Less emphasis on probability theory; focus is on Empirical Validation to generalize results to the population
- Purpose is prediction
- Models are assessed on hold-out data.
### GENERALIZATION

#### Unknown

<p>| | | | | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>$x_1$</td>
<td>$x_2$</td>
<td>$x_3$</td>
<td>$x_4$</td>
<td>$x_5$</td>
<td>$x_6$</td>
<td>...</td>
<td>$x_k$</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

#### New Cases

<p>| | | | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$&gt;n$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- $n$ new cases
- Input Variables
- Unknown
MODEL SELECTION

Underfitting

Overfitting

Just Right
THE OPTIMISM PRINCIPLE

Training
Accuracy = 70%

Test
Accuracy = 47%
DATA SPLITTING

- Comparison
- Selection
- Tuning
- Final Assessment
**Enterprise Miner supports Cross validation for decision trees and leave-one-out cross validation for regression.**
ASSESSMENT ON HOLDOUT DATA
 Predictive modeling and empirical validation

 Model tuning with Validation Data
   Regression, Decision Trees, Neural Networks

 Penalized measures of fit in Regression

 Ensemble Models
FIT VERSUS COMPLEXITY
Create a sequence of models with increasing complexity.
### General Model Tuning

Create a sequence of models with increasing complexity.

**Training Data**

<table>
<thead>
<tr>
<th></th>
<th></th>
<th>inputs</th>
<th></th>
<th>target</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Validation Data**

<table>
<thead>
<tr>
<th></th>
<th></th>
<th>inputs</th>
<th></th>
<th>target</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
## General Model Tuning

### Model Complexity

<table>
<thead>
<tr>
<th>Complexity</th>
<th>Validation Assessment</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>★★</td>
</tr>
<tr>
<td>2</td>
<td>★★★</td>
</tr>
<tr>
<td>3</td>
<td>★★★★</td>
</tr>
<tr>
<td>4</td>
<td>★★★★★</td>
</tr>
<tr>
<td>5</td>
<td>★★★★★</td>
</tr>
</tbody>
</table>

Rate model performance using validation data.

### Training Data

<table>
<thead>
<tr>
<th>Inputs</th>
<th>Target</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Validation Data

<table>
<thead>
<tr>
<th>Inputs</th>
<th>Target</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Rate model performance using validation data.
### General Model Tuning

Select the simplest model with the highest validation assessment.

**Training Data**

<table>
<thead>
<tr>
<th><strong>Inputs</strong></th>
<th><strong>Target</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Validation Data**

<table>
<thead>
<tr>
<th><strong>Inputs</strong></th>
<th><strong>Target</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Create a sequence of models with increasing complexity.
Create a sequence of models with increasing complexity.

A maximal tree is the most complex model in the sequence.
A maximal tree is the most complex model in the sequence.
MODEL TUNING: DECISION TREES

The subtree with the highest validation assessment is selected.
The subtree with the highest validation assessment is selected.
MODEL TUNING: DECISION TREES

Similarly this is done for subsequent models.
MODEL TUNING: DECISION TREES

Training Data

Validation Data

Prune two splits from the maximal tree,...
MODEL TUNING: DECISION TREES

Training Data

Validation Data

...rate each subtree using validation assessment, and...
MODEL TUNING: DECISION TREES

...select the subtree with the best assessment rating.
### Training Data

<table>
<thead>
<tr>
<th>Inputs</th>
<th>Target</th>
</tr>
</thead>
<tbody>
<tr>
<td>Inputs</td>
<td>Target</td>
</tr>
<tr>
<td>Inputs</td>
<td>Target</td>
</tr>
<tr>
<td>Inputs</td>
<td>Target</td>
</tr>
</tbody>
</table>

### Validation Data

<table>
<thead>
<tr>
<th>Inputs</th>
<th>Target</th>
</tr>
</thead>
<tbody>
<tr>
<td>Inputs</td>
<td>Target</td>
</tr>
<tr>
<td>Inputs</td>
<td>Target</td>
</tr>
<tr>
<td>Inputs</td>
<td>Target</td>
</tr>
</tbody>
</table>

Continue pruning until all subtrees are considered.
Continue pruning until all subtrees are considered.
MODEL TUNING: DECISION TREES

Compare validation assessment between tree complexities.
Choose the simplest model with highest validation assessment.
MODEL TUNING: REGRESSION STEPWISE SELECTION

Stop
MODEL TUNING: REGRESSION BACKWARD ELIMINATION

0
1
2
3
4
5
6
Stop

Stop
MODEL TUNING: REGRESSION

Evaluate each sequence step.
MODEL TUNING: REgression

Model fit statistic

Evaluate each sequence step.

Choose simplest optimal model.
MODEL TUNING: NEURAL NETWORKS

\[
\text{logit}(\hat{y}) = 0 + 0 \cdot H_1 + 0 \cdot H_2 + 0 \cdot H_3
\]

\[
H_1 = \tanh(-1.5 - 0.03x_1 + 0.07x_2)
\]

\[
H_2 = \tanh(0.79 - 0.17x_1 - 0.16x_2)
\]

\[
H_3 = \tanh(0.57 + 0.05x_1 + 0.35x_2)
\]

random initial input weights and biases
MODEL TUNING: NEURAL NETWORKS

![Graph showing ASE for training and validation iterations. The graph displays a steep decline in ASE for both training and validation data, with a plateau starting at iteration 20.](image_url)
MODEL TUNING: NEURAL NETWORKS

Iteration

ASE

training validation

0 5 10 15 20 23
This data focuses on North Carolina births for 2000 and 2001. The original data sets include more than 120,000 births in each year and contain data on the ethnicity, age, education level, and marital status of the parents; prenatal medical care received; and information about the mother's reproductive history including number of previous pregnancies and live births (State Center for Health Statistics 2001, 2002).
ENTERPRISE MINER DEMONSTRATION: VALIDATION AND CROSS VALIDATION TUNING

PROPENSITY TO GIVE BIRTH TO A LOW BIRTH WEIGHT BABY
GENERALIZATION AND THE RIGHT SIZED MODEL: OVERVIEW

- Predictive modeling and empirical validation
- Model tuning with Validation Data
  - Regression, Decision Trees, Neural Networks
- Penalized measures of fit in Regression
- Ensemble models
For least square training: $SBC = N \ln(SSE/N) + P \ln(N)$

$AIC = N \ln(SSE/N) + 2P$

For maximum likelihood training: $SBC = -2LL + P \ln(N)$

$AIC = -2LL + 2P$

(N: sample size  SSE: error sums of squares  P: number of model parameters  LL: log likelihood)

SMALLER IS BETTER
PENALIZED MEASURES OF FIT: REGRESSION

Fit Criterion for Price

Akaike Information Criterion

Number of Parameters

Best Model Evaluated at Number of Parameters

Sawa's Bayesian Information Criterion

Number of Parameters

Best Model Evaluated at Number of Parameters
ENTERPRISE MINER DEMONSTRATION:
PENALIZED FIT MEASURES
PROPENSITY TO DONATE TO CHARITY AMONGST LAPSED DONORS
GENERALIZATION AND THE RIGHT SIZED MODEL: OVERVIEW

- Predictive modeling and empirical validation
- Model tuning with Validation Data
  - Regression, Decision Trees, Neural Networks
- Penalized measures of fit in Regression
- Ensemble models
ENSEMBLE MODELS

Combine predictions from multiple models to create a single consensus prediction.
ENSEMBLE MODELS

- Combined Models
- Bagging
- Boosting
- Gradient Boosting
- Random Forests
BAGGING

\[
\begin{array}{cccc}
\text{case} & \text{freq} & \text{freq} & \text{freq} & \text{freq} \\
1 & 1 & 0 & 3 & 1 \\
2 & 0 & 1 & 1 & 1 \\
3 & 2 & 0 & 0 & 2 \\
4 & 0 & 2 & 2 & 0 \\
5 & 2 & 2 & 0 & 1 \\
6 & 1 & 1 & 0 & 1 \\
\end{array}
\]
### ARC-X4

<table>
<thead>
<tr>
<th>Case</th>
<th>$freq$</th>
<th>$m$</th>
<th>$freq$</th>
<th>$m$</th>
<th>$freq$</th>
<th>$m$</th>
<th>$freq$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1.5</td>
<td>1</td>
<td>.5</td>
<td>2</td>
<td>.97</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>0</td>
<td>.75</td>
<td>0</td>
<td>.25</td>
<td>0</td>
<td>.06</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>1</td>
<td>1.5</td>
<td>2</td>
<td>4.25</td>
<td>3</td>
<td>4.69</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>0</td>
<td>.75</td>
<td>1</td>
<td>.5</td>
<td>1</td>
<td>.11</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>0</td>
<td>.75</td>
<td>0</td>
<td>.25</td>
<td>0</td>
<td>.06</td>
</tr>
<tr>
<td>6</td>
<td>1</td>
<td>0</td>
<td>.75</td>
<td>0</td>
<td>.25</td>
<td>1</td>
<td>.11</td>
</tr>
</tbody>
</table>

![Trees](trees.png)
SINGLE, BAGGED, AND BOOSTED TREE
GRADIENT BOOSTING WITH DECISION TREES

Model

\[ F_M(x) = F_0 + \nu \beta_1 T_1(x) + \nu \beta_2 T_2(x) + \ldots + \nu \beta_M T_M(x) \]

where \( M \) is the number of iterations.

Update Formula

For \( m = 1 \) to \( M \), do...

\[ F_m(x) = F_{m-1}(x) + \nu \beta_m T_m(x) \]

Examples

\( m=1 \)

\[ F_1(x) = F_0 + \nu \beta_1 T_1(x) \]

\( m=2 \)

\[ F_2(x) = F_0 + \nu \beta_1 T_1(x) + \nu \beta_2 T_2(x) \]

and so on.
ENTERPRISE MINER DEMONSTRATION
PROPENSITY TO BUY INSURANCE PRODUCT AMONGST BANK CUSTOMERS
THANK YOU!

Lorne Rothman, PhD, PStat
Principal Statistician, SAS Institute (Canada) Inc.
Lorne.Rothman@sas.com