Tree Ensembles: The Power of Post-Processing

December 2012
Dan Steinberg
Mikhail Golovnya
Salford Systems
Course Outline

- **Salford Systems** – quick overview
- **Treenet** – an ensemble of boosted trees
- **GPS** – modern highly versatile regularized regression
- **ISLE** – an effective way to compress ensemble models
- **RuleLearner** – extracting most interesting sets of rules
- **Examples** of compression

© Salford Systems 2012
Brief Overview of Salford Systems

• **Salford Systems** principally well known for three reasons
  o **Products** (CART®, MARS®, TreeNet®, RandomForests®)
  o **Brains trust** (J. Friedman, L. Breiman, R. Olshen, C. Stone)
  o **Awards** (multiple data mining competitions since 2000 until today)

• **Marquee** list of major customers in banking, insurance, retail trade bricks-and-mortar and online, pharmaceuticals, etc.)
  o NAB Bank
  o American Express
  o VISA USA (credit cards)
  o Capital One (credit cards)
  o Traveler’s Insurance (formerly Citigroup)
  o Johnson & Johnson

• **Partners** in China and Japan (Asian language versions)

© Salford Systems 2012
Salford Competitive Awards

• 2010 Direct Marketing Association. 1st place*
• 2009 KDDCup IDAnalytics*, and FEG Japan* 1st Runner Up
• 2008 DMA Direct Marketing Association 1st Runner Up
• 2007 Pacific Asia PAKDD: Credit Card Cross Sell. 1st place
• 2006 DMA Direct Marketing Association: Predictive Modeling*
• 2006 PAKDD Pacific Asia KDD: Telco Customer Type Profiling
• 2005 BI-Cup Latin America: Predictive Modeling E-commerce* 1st place
• 2004 KDDCup: Predictive Modeling ‘Most Accurate’ *
• 2002 NCR/Teradata Duke University: Predictive Modeling-Churn
  o all four separate predictive modeling challenges 1st place
• 2000 KDDCup: Predictive Modeling- Online behavior 1st place
• 2000 KDDCup: CRM Analysis 1st place

*Won either by Salford or by client using Salford tools
Salford Predictive Modeler SPM

• Download a current version from our website http://www.salford-systems.com

• Version will run without a license key for 10-days

• Request a license key from unlock@salford-systems.com

• Request configuration to meet your needs
  o Data handling capacity
  o Data mining engines made available
Stochastic Gradient Boosting

- New approach to machine learning/function approximation developed by Jerome H. Friedman at Stanford University
  - Co-author of CART® with Breiman, Olshen and Stone
  - Author of MARS®, PRIM, Projection Pursuit
- Also known as Treenet®
- Good for classification and regression problems
- Built on small trees and thus
  - Fast and efficient
  - Data driven
  - Immune to outliers
  - Invariant to monotone transformations of variables
- Resistant to over training – generalizes very well
- Can be remarkably accurate with little effort
- BUT resulting model may be very complex
The Algorithm

• Begin with a very small tree as initial model
  o Could be as small as ONE split generating 2 terminal nodes
  o Typical model will have 3-5 splits in a tree, generating 4-6 terminal nodes
  o Model is intentionally “weak”

• Compute “residuals” (prediction errors) for this simple model for every record in data
• Grow a second small tree to predict the residuals from the first tree
• Compute residuals from this new 2-tree model and grow a 3rd tree to predict revised residuals
• Repeat this process to grow a sequence of trees
Illustration: Saddle Function

- 500 \( \{X_1, X_2\} \) points randomly drawn from a \([-3, +3]\) box to produce the XOR response surface \( Y = X_1 \times X_2 \)
- Will use 3-node trees to show the evolution of Treenet response surface
Notes on Treenet Solution

• The solution evolves slowly and usually includes hundreds or even thousands of small trees.
• The process is myopic – only the next best tree given the current set of conditions is added.
• There is a high degree of similarity and overlap among the resulting trees.
• Very large tree sequences make the model scoring time and resource intensive.

• Thus, ever present need to simplify (reduce) model complexity.
Real Life Dataset

Dataset supplied by 2006 DMC Data Mining competition

8,000 iPod auctions held at eBay from May 2005 to May 2006 in Germany

Auction items were grouped into 15 mutually exclusive categories based on size, type (regular, mini, nano), and color

Goal: predict whether the closing price will be above or below the category average
## The Data: Available Fields

<table>
<thead>
<tr>
<th>Variable</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>CATEGORY_NAME</td>
<td>Product category</td>
</tr>
<tr>
<td>LISTING_DURTN_DAYS</td>
<td>duration of auction</td>
</tr>
<tr>
<td>LISTING_TYPE_CODE</td>
<td>type of auction (normal auction, multi auction, etc)</td>
</tr>
<tr>
<td>QTY_AVAILABLE</td>
<td>amount of offered items for multi auction</td>
</tr>
<tr>
<td>FEEDBACK_SCORE</td>
<td>feedback-rating of the seller of this auction listing</td>
</tr>
<tr>
<td>START_PRICE</td>
<td>start price in EUR</td>
</tr>
<tr>
<td>BUY_IT_NOW_PRICE</td>
<td>buy it now price in EUR</td>
</tr>
<tr>
<td>BUY_IT_NOW_FLAG</td>
<td>option for buy it now on this auction listing</td>
</tr>
<tr>
<td>BOLD_FEE_FLAG</td>
<td>option for bold font on this auction listing</td>
</tr>
<tr>
<td>CATEGORY_FEATURED_FEE_FLAG</td>
<td>show this auction listing on top of category</td>
</tr>
<tr>
<td>GALLERY_FEE_FLAG</td>
<td>auction listing with picture gallery</td>
</tr>
<tr>
<td>RESERVE_FEE_FLAG</td>
<td>auction listing with reserve-price</td>
</tr>
<tr>
<td>...</td>
<td>Variety of additional fees</td>
</tr>
</tbody>
</table>
A Tree is a Variable Transformation

Any tree in a Treenet model can be represented by a derived continuous variable as a function of inputs.

\[ \text{TREE}_1 = F(X_1, X_2) \]
The original Treenet model combines all trees with equal coefficients. **ISLE** accomplishes model compression by removing redundant trees and changing the relative contribution of the remaining trees by adjusting the coefficients. **Regularized Regression** methodology provides the required machinery to accomplish this task!
• **GPS (Generalized Path Seeker)** introduced by Jerome Friedman in 2008
• Dramatically expands the pool of potential linear models by including different sets of variables in addition to varying the magnitude of coefficients
• The optimal model of any desirable size can then be selected based on its performance on the TEST sample
Path Building Process

- **Elasticity Parameter** – controls the variable selection strategy along the path (using the LEARN sample only), it can be between 0 and 2, inclusive
  - Elasticity = 2 – fast approximation of Ridge Regression, introduces variables as quickly as possible and then jointly varies the magnitude of coefficients – lowest degree of compression
  - Elasticity = 1 – fast approximation of Lasso Regression, introduces variables sparingly letting the current active variables develop their coefficients – good degree of compression versus accuracy
  - Elasticity = 0 – fast approximation of Stepwise Regression, introduces new variables only after the current active variables were fully developed – excellent degree of compression but may lose accuracy

© Salford Systems 2012
ISLE managed to half the model complexity while slightly improving the TEST sample performance.
Lasso variable selection strategy turned out to be the winning compression approach here.
Note that even higher compression ratios can be achieved if one is willing to sacrifice a little bit of accuracy.
All ISLE models show uniformly better performance than the corresponding TN models.
A Node is a Variable Transformation

- Any node in a Treenet model can be represented by a derived dummy variable as a function of inputs.
RuleLearner Compression

- Create an exhaustive set of dummy variables for every node (internal and terminal) and every tree in a TN model.
- Run GPS regularized regression to extract an informative subset of node dummies along with the modeling coefficients.
- Thus, a model compression can be achieved by eliminating redundant nodes.
- Each selected node dummy represents a specific rule-set which can be interpreted directly for further insights.
ISLE Compression in Adserving

- Web portal leveraging hundreds of Treenet models to optimize the user experience
- Optimal models may each have hundreds or thousands of trees
- Have sufficient time to score no more than 30 trees (each with 6 nodes) to execute page quickly enough
- Originally simply took the first 30 trees from the model
  - May have reworked the model manipulating the learn rate
- With ISLE much better accuracy is possible because we can now work with an optimal subset of 30 trees instead first 30
**RuleLearner Compression on eBay Data**

- RuleLearner has reduced model complexity to 1/10\textsuperscript{th} while improved its performance
- A large number of HotSpots has been identified
- All RuleLearner models show uniformly better performance than the corresponding Treenet models

---

© Salford Systems 2012
RuleLearner Compression in Banking

- Loan default dataset from a major bank, 16,000 accounts, 43 predictors, 10% overall default rate
- 64% compression using RuleLearner, no accuracy loss
- The top rules have lift of 6 and provide significant insights in terms of hot spots on loan defaults
Big Data – The Future

• In model compression we can think of the tree building as simply a search engine for interesting transforms of the raw data

• Ideal transforms:
  o Handle missing values
  o Impervious to predictor outliers
  o Invariant to how the raw data is expressed (eg. Scale)

• Transforms can be discovered on very small samples of the data (ideally stratified samples)

• Second stage regression is much more easily run as a MapReduce job on the transforms
  o Challenge is that there can be many transforms
  o Results of such experiments very promising