Lecture 4 - Review

• Error of the hypothesis vs error of the learning algorithm?
• Know the training and test set error, good estimate of the learner’s performance?
• Learners Error = noise + bias² + variance
• How we calculate bias and variance for a learner*
  – $T_{1\ldots n}$: Training sets drawn randomly from population
• Bias is the expected (mean) error over all training sets
• Variance is the variability of the error.
• Why would a decision tree be biased? Have a high variance?

Ensemble Techniques Reduce Error

• Decision trees are known to have a high variance, particularly when overfitted.
• BMA
  – Expected cost of Bayesian prediction is the noise.
  – Why?
• Bagging
  – Reduces variance but not bias
• Boosting
  – Reduces what?

Boosting – The Idea

• Take weak learners (marginally better than random guessing) make them stronger.
• Freund and Schapire, 95 – AdaBoost
• AdaBoost premise
  – Each training instances has equal weight
  – Build first Model from training instances
  – Training instances that are classified incorrectly given more weight
  – Build another model with re-weighted instances and so on and so on.
Boosting Pseudo Code

\[ D_0(i) = \frac{1}{I} \] Initially training instances have same weight

For \( j = 1 \) to \( J \) // \( J \) is the number of rounds (trees)

Build \( H_j \) from \( D_j \)

\[ \alpha_j = 0.5 \log \left( \frac{1 - \text{Error}(H_j)}{\text{Error}(H_j)} \right) \]

For \( i = 1 \) to \( I \)

if instance \( i \) is misclassified then

\[ D_{j+1}(i) = D_j(i) e^{\alpha_j} \]

else

\[ D_{j+1}(i) = D_j(i) e^{-\alpha_j} \]

endif

endfor

endfor

Elaborate calculation of \( \alpha_j \) is so that \( \sum D_j(i) = 1 \)

Prediction(\( x \)) = \( \sum \alpha_j H_j(x) \), \( H_j(x) \) produces a 0 or 1

Some Implementation Comments

• Difficult to parallelize
• Factoring instance weights into decision tree induction.
• Tree vote is weighted inversely to error.
• Adaptive Boosting (AdaBoosting) according to the tree error
• Free scaled down version of C5.0 incorporates boosting available at http://www.rulequest.com/download.html

Toy Example (Freund COLT 99)

Round 1

\[ D_0 \]
Round 2 + 3

Final Hypothesis

Some Insights into Boosting

- Final aggregate model will have no training error (given some conditions).
- Seems to over-fit but reduces test set error
- Larger margins on training set correspond to better generalization error
  
  \[
  \text{Margin}(x) = y \sum \alpha_i h_i(x) / \sum \alpha_i
  \]
Ensemble Technique Scorecard

<table>
<thead>
<tr>
<th></th>
<th>BMA</th>
<th>Bagging</th>
<th>Boosting</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reduces Variance of base</td>
<td>Yes</td>
<td>Yes</td>
<td>No*</td>
</tr>
<tr>
<td>Testing Scheme</td>
<td>Degree of Belief in Model</td>
<td>Equal</td>
<td>Depends on Model Error</td>
</tr>
<tr>
<td>Requirement of Learners</td>
<td>Regression</td>
<td>Estimable</td>
<td>Weak (commonly better than random guessing)</td>
</tr>
</tbody>
</table>

Retrospective on Decision Trees

- Representation and search
- Does Bagging and Boosting change model representation space?
- Do they change search preference?
- Order of data presented does not count.

Reading for Next Classes

- Additional reading
  - Various papers at http://www.boosting.org
- Lecture 6 – 02/11/02 and onwards
  - Mitchell chapter 4 (Neural Networks)