Errors - Review

• Error of the hypothesis vs error of the algorithm?
• Know the training and test set error, good estimate of the classifier’s performance?
• Classifier Error = noise + bias² + variance
• How we calculate bias and variance for a classifier*
  – $T_{1...n}$: 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?
Errors

The true error of hypothesis $h$ with respect to target function $f$ and distribution $\mathcal{D}$ is the probability that $h$ will misclassify an instance drawn at random according to $\mathcal{D}$.

$$error_\mathcal{D}(h) \equiv \Pr_{x \in \mathcal{D}}[f(x) \neq h(x)]$$

The sample error of $h$ with respect to target function $f$ and data sample $S$ is the proportion of examples $h$ misclassifies.

$$error_S(h) \equiv \frac{1}{n} \sum_{x \in S} \delta(f(x) \neq h(x))$$

Where $\delta(f(x) \neq h(x))$ is 1 if $f(x) \neq h(x)$, and 0 otherwise.

How well does $error_S(h)$ estimate $error_\mathcal{D}(h)$?
Bias and Variance

1. **Bias**: If $S$ is training set, $\text{error}_S(h)$ is optimistically biased

   \[ \text{bias} \equiv E[\text{error}_S(h)] - \text{error}_D(h) \]

   For unbiased estimate, $h$ and $S$ must be chosen independently

2. **Variance**: Even with unbiased $S$, $\text{error}_S(h)$ may still vary from $\text{error}_D(h)$

   ??? What else ???
Model Uncertainty

• What’s wrong with making predictions from one model?
  – May have two or more equally accurate models that give different predictions.
  – May have two models that are quite fundamentally different
Ensemble of Models Techniques

• Bayesian Modeling Averaging
  – $\Pr(c, x \mid D, H) = \sum_{h \in H} \Pr(c, x \mid h) \cdot \Pr(h \mid D)$
  – Weight each model’s prediction by how good the model is.
  – Can this approach be applied to C4.5 Dtrees?

• Bagging (Bootstrap Aggregation), 1996.
  – Improves accuracy
    • Seminal paper says on 19 of 26 data sets improves accuracy by 4%.
The Bagging Algorithm

• Building the Models
  For $i = 1$ to $k$ // $k$ is the number of bags
  $T_i = \text{BootStrap}(D)$ // $D$ is the training set
  Build Model $M_i$ from $T_i$ (ie. Induce the tree)
  End

• Applying the Models To Make a Prediction
  For a test set example, $x$
  For $i = 1$ to $k$ // $k$ is the number of bags
  $C_i = M_i(x)$
  End
  Prediction is the class with the most vote.
Take A Bootstrap Sample

Sample with replacement
Bootstrapping and model building can be easily parallelized

<table>
<thead>
<tr>
<th>Original</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td>Training set 1</td>
<td>2</td>
<td>7</td>
<td>8</td>
<td>3</td>
<td>7</td>
<td>6</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>Training set 2</td>
<td>7</td>
<td>8</td>
<td>5</td>
<td>6</td>
<td>4</td>
<td>2</td>
<td>7</td>
<td>1</td>
</tr>
<tr>
<td>Training set 3</td>
<td>3</td>
<td>6</td>
<td>2</td>
<td>7</td>
<td>5</td>
<td>6</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>Training set 4</td>
<td>4</td>
<td>5</td>
<td>1</td>
<td>4</td>
<td>6</td>
<td>4</td>
<td>3</td>
<td>8</td>
</tr>
</tbody>
</table>
Example of Bagging

Problem

Single DT Solution

100 DT’s

Bagging Solution
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

- Initialize distribution over the training set $D_1(i) = 1/m$
- For $t = 1, \ldots, T$:
  1. Train Weak Learner using distribution $D_t$.
  2. Choose a weight (or confidence value) $\alpha_t \in \mathbb{R}$.
  3. Update the distribution over the training set:

$$D_{t+1}(i) = \frac{D_t(i)e^{-\alpha_t y_i h_t(x_i)}}{Z_t}$$ (2)

Where $Z_t$ is a normalization factor chosen so that $D_{t+1}$ will be a distribution.

- Final vote $H(x)$ is a weighted sum:

$$H(x) = \text{sign}(f(x)) = \text{sign} \left( \sum_{t}^{T} \alpha_t h_t(x) \right)$$ (3)
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

$D_1$

$D_2$

$\varepsilon_1 = 0.30$
$\alpha_1 = 0.42$
Round 2 + 3

$\varepsilon_2 = 0.21$
$\alpha_2 = 0.65$

$\varepsilon_3 = 0.14$
$\alpha_3 = 0.92$

DM Spring 2004 - Lecture 20-21
Final Hypothesis

\[ H_{\text{final}} = \text{sign}(0.42 + 0.65 + 0.92) \]

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
  – Margin($x$) = $y \sum \alpha_j h_j(x) / \sum \alpha_j$
The Performance of Models and Learners

• 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*
  – \( T_{1,...,n} \) : Training sets drawn randomly from population
• Bias is the difference in error over all training sets – true error.
• 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?
## Ensemble Technique Scorecard

<table>
<thead>
<tr>
<th></th>
<th>BMA</th>
<th>Bagging</th>
<th>Boosting</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reduces Variance Or bias</td>
<td>Both</td>
<td>Variance</td>
<td>Bias*</td>
</tr>
<tr>
<td>Voting 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>Bayesian</td>
<td>Unstable</td>
<td>Weak (consistently better than random guessing)</td>
</tr>
<tr>
<td></td>
<td>?</td>
<td>?</td>
<td>?</td>
</tr>
</tbody>
</table>

DM Spring 2004 - Lecture 20-21
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.