Soft Decision Trees

Ozan İrsoy\textsuperscript{1} \quad Olcay Taner Yıldız\textsuperscript{2} \quad Ethem Alpaydın\textsuperscript{1}

\textsuperscript{1}Department of Computer Engineering, Boğaziçi University, TR-34342, Istanbul, Turkey

\textsuperscript{2}Department of Computer Engineering, İşık University, TR-34980, Istanbul, Turkey

ICPR 2012
Hard Decision Tree

- Each decision node \( m \) applies a test \( g_m(x) \) and chooses one of the children accordingly.

\[
F_m(x) = \begin{cases} 
F_m^L(x) & \text{if } g_m(x) > 0 /* true */ \\
F_m^R(x) & \text{otherwise /* false */}
\end{cases}
\]

- Classification: Leaves carry the label of one of \( K \) classes
- Regression: Leaves carry a constant which is the numeric regression value.
Hard Decision Tree Types

- **Univariate tree:** $g_m(x) = x_j + w_{m0} > 0$ (Quinlan, 1993).
- **Multivariate linear tree:** $g_m(x) = w^T_m x + w_{m0} > 0$ (Murthy and Salzberg, 1994), (Yildiz and Alpaydin, 2005).
- **Multivariate nonlinear tree:** $g_m(x) = \sum_{j=1}^k w_j \phi_j(x) > 0$ (Guo and Gelfand, 1992).
- **Omnivariate tree:** $g_m(x)$ can be any of the above, chosen by a statistical model selection procedure (Yildiz and Alpaydin, 2001).
Soft Decision Tree

- Soft decision node redirects instances to all its children with probabilities calculated by a *gating function* $g_m(x)$.

\[
F_m(x) = F^L_m(x)g_m(x) + F^R_m(x)(1 - g_m(x))
\]

\[
g_m(x) = \frac{1}{1 + \exp[-(w^T_m x + w_{m0})]}
\]

- Gating model implements a discriminative (logistic linear) model estimating the posterior probability of the left child.
Hard vs. Soft Tree (Toy Dataset)
Response and Error

```
1 function F_m(x)
2   if m is leaf node
3       y = z_m /* leaf value at m */
4   else
5       g_m(x) = 1/(1 + exp(-(w^T_m x + w_m0)))
6   y = F^L_m(x)g_m(x) + F^R_m(x)(1 - g_m(x))
7   return y
```

- Classification: $E = r \log y + (1 - r) \log(1 - y)$
- Regression: $E = (r - y)^2$
1 function LearnSoftTree(m, X, V)
2  
3 initialize \(w_{mj}, z^L_m, \) and \(z^R_m\)
4 repeat
5     for all \((x, r) \in X\)
6         \(\delta(x) = (F_{\text{root}}(x) - r)(g_p(x))^{\text{left}}(1 - g_p(x))^{\text{right}}\)
7         for \(j = 0, \ldots, d\)
8             \(w_{mj} = w_{mj} - \eta \delta(x)(F^L_m(x) - F^R_m(x))v_m(x)(1 - v_m(x))x_j\)
9             \(z^L_m = z^L_m - \eta \delta(x)v_m(x)\)
10            \(z^R_m = z^R_m - \eta \delta(x)(1 - v_m(x))\)
11 until convergence
12 \(E_{\text{after}} = \text{ErrorOfTree}(V)\)
13 if \(E_{\text{after}} < E_{\text{before}}\)
14     LearnSoftTree(m.left, X, V)
15     LearnSoftTree(m.right, X, V)
## Results: Regression

<table>
<thead>
<tr>
<th></th>
<th>Mean Square Error</th>
<th>Tree Size</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Soft</td>
<td>Hard</td>
</tr>
<tr>
<td>ABA</td>
<td>0.439</td>
<td>0.557</td>
</tr>
<tr>
<td>ADD</td>
<td>0.094</td>
<td>0.267</td>
</tr>
<tr>
<td>BOS</td>
<td>0.271</td>
<td>0.344</td>
</tr>
<tr>
<td>CAL</td>
<td>0.312</td>
<td>0.326</td>
</tr>
<tr>
<td>COM</td>
<td>0.037</td>
<td>0.046</td>
</tr>
<tr>
<td>CON</td>
<td>0.264</td>
<td>0.286</td>
</tr>
<tr>
<td>8FH</td>
<td>0.383</td>
<td>0.418</td>
</tr>
<tr>
<td>8FM</td>
<td>0.057</td>
<td>0.074</td>
</tr>
<tr>
<td>8NH</td>
<td>0.388</td>
<td>0.416</td>
</tr>
<tr>
<td>8NM</td>
<td>0.054</td>
<td>0.084</td>
</tr>
</tbody>
</table>
## Results: Classification

<table>
<thead>
<tr>
<th></th>
<th>Accuracy</th>
<th></th>
<th></th>
<th>Tree Size</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Soft</td>
<td>Hard</td>
<td>LDT</td>
<td>Soft</td>
</tr>
<tr>
<td>BRE</td>
<td>95.34</td>
<td>93.80</td>
<td>95.09</td>
<td>17</td>
</tr>
<tr>
<td>GER</td>
<td>75.74</td>
<td>69.07</td>
<td>74.16</td>
<td>16</td>
</tr>
<tr>
<td>MAG</td>
<td>81.27</td>
<td>84.09</td>
<td>83.07</td>
<td>17</td>
</tr>
<tr>
<td>MUS</td>
<td>92.25</td>
<td>94.62</td>
<td>93.59</td>
<td>22</td>
</tr>
<tr>
<td>PIM</td>
<td>70.85</td>
<td>69.41</td>
<td>76.89</td>
<td>26</td>
</tr>
<tr>
<td>POL</td>
<td>77.41</td>
<td>69.81</td>
<td>77.45</td>
<td>21</td>
</tr>
<tr>
<td>RIN</td>
<td>88.94</td>
<td>87.54</td>
<td>77.25</td>
<td>368</td>
</tr>
<tr>
<td>SAT</td>
<td>83.90</td>
<td>84.01</td>
<td>83.30</td>
<td>11</td>
</tr>
<tr>
<td>SPA</td>
<td>78.38</td>
<td>90.14</td>
<td>89.86</td>
<td>22</td>
</tr>
<tr>
<td>TWO</td>
<td>97.92</td>
<td>87.59</td>
<td>98.00</td>
<td>41</td>
</tr>
</tbody>
</table>
Proposed decision tree model with soft decisions, which makes use of a soft gating function to merge the decisions of the subtrees.

The model is shown to have better or comparable performance to hard trees, while having fewer nodes.

One drawback of soft trees is gradient-descent which is prone to get stuck at local minima.
Conclusions: Soft Trees vs. Hard Trees

- Soft trees have smoother fits and hence lower bias around the split boundaries.
- Linear gating function enables soft trees to make oblique splits in contrast to the axis-orthogonal splits made by hard trees.