Boosted Trees and Statistical Inference
A Roadmap

Giles Hooker
Yichen Zhou, Indrayudh Ghosal, Lucas Mentch
Zhengze Zhou, Sarah Tan

Stanford, May 2019
Inference and Machine Learning

Machine learning: highly successful *predictive models* but do not provide

- explanations/understanding of underlying model structures
- inferential procedures to quantify uncertainty about performance, or individual predictions.

e.g. Breiman’s Two Cultures, 2001.

*Breiman’s Bridge: statistical inference using ML*

Random Forests: a starting point

- bootstrap structure *but*
- used to average, not quantify uncertainty
- but using subsamples results in a *U*-statistic.
A Fundamental Result for Random Forests

Mentch and Hooker (2016):

**Theorem 1**

Let $F_{n,b_n,k_n}(x) = \frac{1}{b_n} \sum_{j=1}^{b_n} T(x; (X, Y)_{j1}, \ldots, (X, Y)_{j_{k_n}}, \omega_i)$ be an ensemble of $b_n$ learners based on $T$ using random subsamples of size $k_n$ and independent random inputs $\omega$ with

\[
\begin{align*}
\zeta_{k_n,k_n} &= \text{var} \left[ T(x; (X, Y)_{1}, \ldots, (X, Y)_{k_n}, \omega) \right] \\
\zeta_{1,k_n} &= \text{var}_{(X,Y)_1} \left[ E \left[ T(x; (X, Y)_{1}, \ldots, (X, Y)_{k_n}, \omega) \mid (X, Y)_1 \right] \right] 
\end{align*}
\]

such that $\lim (k_n/n) \zeta_{k_n,k_n}/k_n \zeta_{1,k_n} = 0$ then

\[
(F_{n,b_n,k_n}(x) - EF_{n,b_n,k_n}(x)) / \sqrt{k_n^2 \zeta_{1,k_n}/n + \zeta_{k_n,k_n}/b_n} \xrightarrow{d} N(0,1)
\]

Wager and Athey (2017): replace $EF(x)$ with $f_0(x) = E(Y \mid x)$ for specific $T$. 
Statistical Applications

Mentch and Hooker (2017): tests for structure, local or global:
- Tests of variable importance; Does changing $x_1$ change $F(x)$?
- Coleman et al. (2019), Does dropping $x_1$ change $F(x)$?
- Zhou and Hooker (2019b), based on importance measures of individual trees.
- Tests of additivity: $f(x_1, x_2, x_3) = g_1(x_1, x_3) + g_2(x_2, x_3)$ and other ANOVA structures.
- Adequacy of model distillation (Zhou et al., 2018)

Wager and Athey (2017) (and follow-up):
- inference for (localized) average treatment effect
- extensions to local regression structures

Bayesian Inference: Chipman et al. (2010) and colleagues for forests, Maddox et al. (2019) in neural networks.
Of Bagging and Boosting

Boosting: models built in sequence

\[ F_{t+1}(x) = F_t(x) + \eta T(x; (X_1, Y_1 - F_t(X_1)), \ldots, (X_n, Y_n - F_t(X_n))) \]

for squared error.

- Strategy for generic losses (Friedman et al. (2000); although see Athey et al. (2019)).
- Impose Structure by varying \( T \):
  - Forward stagewise regression (Friedman, 2001)
  - More generally in Bühlmann and Hothorn (2007) – see package mboost
  - Generalized additive models (Lou et al. (2012), Tan et al. (2017)).
  - Varying coefficient models (Zhou and Hooker, 2019a).
- Reduction in bias: model space of \( T_1(x) + T_2(x) \) is greater than one tree, see also BART, Additive Groves (Sorokina et al., 2007).
Barriers to Boosting Inference

Friedman (2001) algorithm uses all samples
- ⇒ overfits as boosting continues; requires early stopping
- ⇒ properties of $T$ become critical.

Adding subsampling reduces overfitting (Friedman, 2002)
- ⇒ $F_k(x)$ has non-degenerate stationary distribution (hopefully).
- ⇒ need to account for training variability.

Both processes give greatest influence to early trees.

Can we develop some process that allows for distributional results?
One-Step Boosted Forests


\[ F(x) = F_1(x) + F_2(x) \]

\( F_2 \) trained on out-of-bag residuals from \( F_1 \)

- \( F_2 \) captures signal missed by \( F_1 \).
- Correlated normal limit \( \Rightarrow \) can apply to sum.
- Main assumption: swapping \( Y - F_1(X) \) for \( Y - EF_1(X) \) in training \( F_2 \) is negligible.

Extensions to fixed numbers of forests are immediate.
## One-Step Boosted Random Forests

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Basics</th>
<th>Improvement</th>
<th>RF</th>
<th>BFv1</th>
<th>BFv2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>n</td>
<td>d</td>
<td>k</td>
<td>varY</td>
<td>GBM</td>
</tr>
<tr>
<td>yacht-hydrodynamics</td>
<td>308</td>
<td>6</td>
<td>60</td>
<td>229.55</td>
<td>92.92</td>
</tr>
<tr>
<td>BikeSharing-hour</td>
<td>17379</td>
<td>14</td>
<td>2000</td>
<td>2.21</td>
<td>20.39</td>
</tr>
<tr>
<td>Concrete</td>
<td>1030</td>
<td>8</td>
<td>200</td>
<td>279.08</td>
<td>31.89</td>
</tr>
<tr>
<td>airfoil</td>
<td>1503</td>
<td>5</td>
<td>300</td>
<td>46.95</td>
<td>-33.84</td>
</tr>
<tr>
<td>housing</td>
<td>506</td>
<td>13</td>
<td>150</td>
<td>0.17</td>
<td>-32.32</td>
</tr>
<tr>
<td>auto-mpg</td>
<td>392</td>
<td>7</td>
<td>50</td>
<td>61.03</td>
<td>7.98</td>
</tr>
<tr>
<td>winequality-white</td>
<td>4898</td>
<td>11</td>
<td>1000</td>
<td>0.79</td>
<td>-23.33</td>
</tr>
<tr>
<td>parkinsons</td>
<td>5875</td>
<td>16</td>
<td>1000</td>
<td>66.14</td>
<td>-24.87</td>
</tr>
<tr>
<td>winequality-red</td>
<td>1599</td>
<td>11</td>
<td>300</td>
<td>0.65</td>
<td>-9.81</td>
</tr>
<tr>
<td>SkillCraft</td>
<td>3338</td>
<td>18</td>
<td>600</td>
<td>2.10</td>
<td>2.24</td>
</tr>
<tr>
<td>communities</td>
<td>1994</td>
<td>96</td>
<td>400</td>
<td>0.05</td>
<td>-2.56</td>
</tr>
</tbody>
</table>
Boulevard: Boosting to $\infty$

Rather than summing, maintain an average:

$$F_{t+1}(x) = \frac{t}{t+1} F_t(x)$$

$$+ \frac{\lambda}{t+1} T(x; (X_{t_1}, Y_{t_1} - F_k(X_{t_1})), \ldots, (X_{t_k}, Y_{t_k} - F(X_{t_k})))$$

- progressively down-weights initial trees
- subsampling avoids over-fitting
- resulting learners exchangeable in the limit

Boulevard: sequence of trees, but all of the same size/influence.
Heuristic Justification

1. Finite sample convergence

\[ F_t(x) \xrightarrow{P} F_\infty(x) \]

2. Asymptotic exchangeability; write \( F_\infty(x) \) as

\[
\frac{1}{\binom{n}{k_n}} \sum E_\omega T(x; (X_{i_1}, Y_{i_1} - F_\infty(X_{i_1})), \ldots, (X_{i_{k_n}}, Y_{i_{k_n}} - F_\infty(X_{i_{k_n}})), \omega)
\]

3. U-statistic + estimating equation \( \Rightarrow \) normality result.

Note that process suggests

\[ F_\infty(x) \approx \frac{\lambda}{1 + \lambda} f_0(x) \]

but can be re-scaled.
Mathematical Complications: Convergence

Rogozhnikov and Likhomanenko (2017) “Infinite Boost”
- no randomization + contraction mapping theorem.

Requires:
2. Truncation of leaf values $\Gamma_M$.
3. Shrinkage $\lambda < 1$.

Honesty criterion is global, achieved by
- Randomly generated tree structures
- Split samples for tree-structures and leaves
Representation

For $x \in A_j$ using index set $I$ of observations with responses $Z$

$$T_t(x) = \sum \frac{I(i \in I, X_i \in A_j)}{\sum I(i \in I, X_i \in A_j)} Z_i = s_n(x)^T Z$$

and define the corresponding hat matrix

$$\hat{Z} = S_n Z$$

then

$$\lim_{t \to \infty} F_{t,n}(x) = E[s_n(x)] \left[ \frac{1}{\lambda} I + E[S_n] \right]^{-1} Y = k_n(x) Y$$

⇒ kernel ridge regression.

Expectations are w.r.t. subsamples and tree structures.
Results for $X \in [0, 1]^d$

Given

- leaf diameter $O(n^{-1/(d+1)})$
- leaf support $O(n^{1/(d+2)})$
- bounded, Lipschitz $f_0$,
- bounded covariate density
- sub-Gaussian errors, variance $\sigma_\epsilon^2$

we have the central limit theorem

$$
F_{\infty, n}(x) - \frac{\lambda}{1+\lambda} f_0(x) \frac{1}{\|k_n(x)\|} \overset{d}{\rightarrow} N(0, \sigma_\epsilon^2)
$$

Can re-multiply for inference about $f_0$. 
Performance

![Graph showing performance metrics for CCPP with ensemble size on the x-axis and MSE on the y-axis. The graph includes different models like GBT Train, SGBT Train, RF Train, BLV Train, and rBLV Train with their corresponding test metrics.]
Performance

Training and testing error curves of tree ensembles.
Reproduction Interval
Boosting in Structured Models

Cycling boosting iteration between model terms allows trees into structured models (inspired by Bühlmann and Hothorn (2007), Fahrmeir et al. (2013))

Including

- **Additive models:** $y = \sum_j g_j(x_j) + \epsilon$
- **Partially linear models:** $y = \beta x + g(z) + \epsilon$
- **Subsumed within varying coefficient models** $y = \beta(z)x + \epsilon$.

<table>
<thead>
<tr>
<th>Covariate</th>
<th>Slope</th>
<th>Covariate</th>
<th>Slope</th>
</tr>
</thead>
<tbody>
<tr>
<td>Elevation</td>
<td>9.65e-04</td>
<td>Barren</td>
<td>-1.59e+03</td>
</tr>
<tr>
<td>Shallow Ocean</td>
<td>-1.88e+03</td>
<td>Evergreen Broad</td>
<td>2.77e+02</td>
</tr>
<tr>
<td>CoastShore Lines</td>
<td>-6.51e+01</td>
<td>Deciduous Needle</td>
<td>2.57e+02</td>
</tr>
<tr>
<td>Shallow Inland</td>
<td>9.39e+01</td>
<td>Deciduous Broad</td>
<td>2.72e+02</td>
</tr>
<tr>
<td>Moderate Ocean</td>
<td>-1.18e+03</td>
<td>Mixed Forest</td>
<td>7.32e+01</td>
</tr>
<tr>
<td>Deep Ocean</td>
<td>-5.12e+03</td>
<td>Closed Shrubland</td>
<td>-1.19e+03</td>
</tr>
<tr>
<td>Evergreen Needle</td>
<td>-4.54e+02</td>
<td>Open Shrubland</td>
<td>8.60e+01</td>
</tr>
<tr>
<td>Grasslands</td>
<td>-4.49e+02</td>
<td>Woody Savannas</td>
<td>-5.75e+02</td>
</tr>
<tr>
<td>Croplands</td>
<td>-4.29e+02</td>
<td>Savannas</td>
<td>-7.46e+02</td>
</tr>
<tr>
<td>Urban Built</td>
<td>-6.62e+02</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Beijing Housing Data: Coefficients by Location

Zhou and Hooker (2019a): consistency of gradient boosting similar to Bühlmann et al. (2002). See also Buergin and Ritschard (2017); Berger et al. (2017).
Possible Extensions

- Less honest trees:
  - Boulevard without honest trees for example tree structures, use at random?
  - Swap structure and leaf subsets at large iteration gaps.
  - Looser conditions for contraction mapping?
- Dropout to reduce scaling effects (e.g. Rashmi and Gilad-Bachrach (2015), Wager et al. (2013))
  - Completely random tree structures with $100\rho\%$ dropout
    \[ \Rightarrow \text{limit is } \lambda / (1 + (1 - \rho)\lambda) f_0(x) \]
  - Drop every $K$th tree: BART-like $F_\infty^1(x) + \ldots + F_\infty^K(x)$.
- Non-$L2$ losses and Newton boosting
  - Averaging/rescaling no-longer exact.
  - Post-training calibration?


Bibliography II


