Black-box Inference: Efficient, Scalable, Model-Free Tests for Variable Importance

Joint Statistical Meetings 2019

Overview

- Random Forests produce highly accurate predictions with little tuning but are poor at producing interpretable conclusions - "Black-boxes"
- Distributional results for random forest predictions have been developed in works such as (Mentch and Hooker, 2016; Wager and Athey, 2018) and extended in (Peng et al., 2019)
- Valid inference for variable importance (like the F-test for linear regression) has been developed too (Coleman et al., 2019)

Random Forest Definitions

- Random forests are ensembles of randomized decision trees trained on data $\mathcal{D}$, $T(\cdot;\xi,\mathcal{D})$ which make predictions according to

$$RF(x;\mathcal{D}) = \mathbb{E}_{\xi} T(x;\xi,\mathcal{D}) \approx \frac{1}{B}\sum_{k=1}^{B} T(x;\xi_k,\mathcal{D}) = RF_B(x;\mathcal{D})$$

- Intuition: $\text{Bias}(RF(x;\mathcal{D})) = \text{Bias}(T(x;\xi,\mathcal{D}))$ which tends to be small, but

$$\text{Var}(RF(x;\mathcal{D})) = \text{Cor}(T(x;\xi,\mathcal{D}), T(x;\xi',\mathcal{D}))\text{Var}(T(x;\xi,\mathcal{D})) \le \text{Var}(T(x;\xi,\mathcal{D})),$$

so that randomness ξ decreases correlation, which stabilizes predictions
- Averaging of deep decision trees, thus hard to interpret predictions

Random Forest Inference Challenges

- Correlation in random forest average makes analysis challenging
- Typical measures of variable importance are neither statistically valid nor reliable heuristics (Strobl et al., 2007), but are widely used
- Often overstate influence of correlated variables and understate influence of categorical variables

References

Coleman, T., Peng, W., and Mentch, L. (2019). Scalable and efficient hypothesis testing with random forests. *arXiv preprint arXiv:1904.07830.*

Mentch, L. and G. Hooker (2016). Quantifying uncertainty in random forests via confidence intervals and hypothesis tests. *The Journal of Machine Learning Research 17*(1), 841–881.

Peng, W., Coleman, T., and Mentch, L. (2019). Asymptotic distributions and rates of convergence for random forests and other resampled ensemble learners. *arXiv preprint arXiv:1905.10651.*

Strobl, C., A.-L. Boulesteix, A. Zeileis, and T. Hothorn (2007). Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinformatics 8*(1), 25.

Wager, S. and S. Athey (2018). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*

Distributional Random Forest Results

- Recent work takes advantage of averaging nature of random forest predictions to establish central limit theorems for random forests build on subsamples of size $k < n$. Let $\mathcal{D}_{k-c} \cup \mathcal{D}_c$ be a data frame with the last c rows from $\mathcal{D}_c$, then define

$$\zeta_c = \text{Cov}(T(x;\xi,\mathcal{D}_{k-c} \cup \mathcal{D}_c), T(x;\xi,\mathcal{D}_{k-c} \cup \mathcal{D}_c'))$$

- Mentch and Hooker (2016) showed that **if M1:** $k = o(\sqrt{n})$ **and M2:** $\zeta_1 \not\to 0$, then

$$\frac{\sqrt{B}\,[RF_B(x;\mathcal{D}) - \mathbb{E}RF(x;\mathcal{D})]}{\sqrt{\frac{k^2}{\alpha}\zeta_1 + \zeta_k}} \xrightarrow{d} \mathcal{N}(0,1) \quad \text{and } \alpha = \lim_{n\to\infty} \frac{n}{B}$$

by illuminating the connection between U-statistics and subsampled learners.
- Wager and Athey (2018) showed that if additional constraints **(W1: honesty, W2: regularity)** are placed on tree construction, then if $k = o(n^{\beta})$ for $\beta \in (0.5, 1)$:

$$\frac{[RF(x;\mathcal{D}) - \mathbb{E}(Y|X=x)]}{\sigma_n(x)} \xrightarrow{d} \mathcal{N}(0,1)$$

and they further provide consistent estimators for $\sigma_n(x)$.
- **M1**, **W1** and **W2** place many restrictions on tree building, and **M2** is impossible to verify in practice, and additionally don't inform rates of convergence

Relaxing These Assumptions and Berry-Esseen Bounds

- In Peng et al. (2019), **M1** is relaxed and **M2** is eliminated (without enforcing **W1, W2**), so that so long as **P1:** $\frac{k}{n}\frac{\zeta_1}{k\zeta_k} \to 0$ and $k = o(n)$, then

$$\frac{\sqrt{B}\,[RF_B(x;\mathcal{D}) - \mathbb{E}RF(x;\mathcal{D})]}{\sqrt{\frac{k^2}{n}\zeta_1 + \frac{1}{B}\zeta_k}} \xrightarrow{d} \mathcal{N}(0,1)$$

- For a bagged p nearest-neighbor estimator, can be shown that $\frac{\zeta_1}{k\zeta_k} \le c(p) \le 2$ for

$$c(p) = \lim_{k\to\infty} 2p \Big/ \sum_{i=0}^{p-1}\sum_{j=0}^{p-1}\left[\frac{\binom{k-1}{i}\binom{k-1}{j}}{\binom{2k-2}{i+j}}\right]$$

- Peng et al. (2019) also provide a rate of convergence to a normal distribution, commonly referred to as *Berry-Esseen* bounds. Subject to moment conditions,

$$\sup_{z\in\mathbb{R}} |F_{n,B,x}(z) - \Phi(z)| \le C\left(\frac{\mathbb{E}|T(x;\xi,\mathcal{D}_1 \cup \mathcal{D}_{k-1})|^3}{n^{1/2}(\mathbb{E}|T(x;\xi,\mathcal{D}_1 \cup \mathcal{D}_{k-1})|^2)^{3/2}} + \frac{\mathbb{E}|T(x;\xi,\mathcal{D}) - \theta|^3}{B^{1/2}(\mathbb{E}|T(x;\xi,\mathcal{D}) - \mathbb{E}T(x;\xi,\mathcal{D})|^2)^{3/2}} + \left[\frac{k}{n}\left(\frac{\zeta_k}{k\zeta_1} - 1\right)\right]^{1/2} + \left(\frac{k}{n}\right)^{1/3}\right)$$

where $F_{n,B,x}(z)$ is the actual cdf of a random forest prediction at x.

Feature Importance Using Distributional Results

- Often interested in comparing a full model versus a nested model. Then, the difference in RF predictions at given test points $\mathcal{T}$ is a U-statistic. Mentch and Hooker (2016) showed that for $\hat{D}_B(x) = RF_B(x;\mathcal{D}) - RF_B(x;\mathcal{D}^{\pi})$,

$$H_0 : \mathbb{E}\hat{D}_B(x) = 0 \;\forall\; x \in \mathcal{T} \implies \hat{D}_B^T \hat{\Sigma}_D^{-1} \hat{D}_B \xrightarrow{d} \chi^2_{|\mathcal{T}|}$$

where $\hat{\Sigma}_D$ is a $|\mathcal{T}| \times |\mathcal{T}|$ covariance matrix.
- The Monte Carlo estimation errors associated with $\hat{\Sigma}_D$ are large enough to affect power/Type I error of procedure, leads to requirements like $B_n = O(n)$

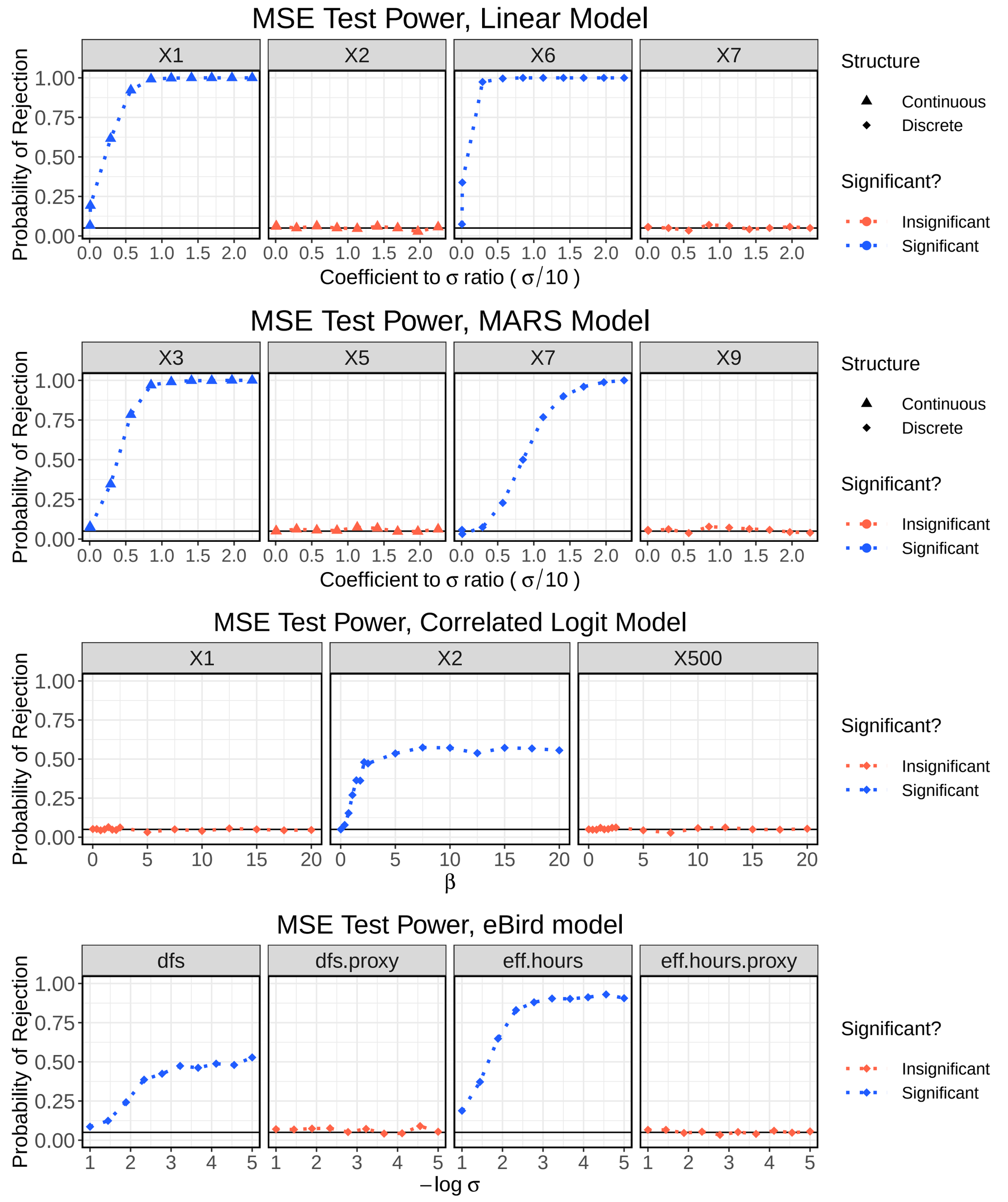

Figure: Power for MSE Test procedure from Coleman et al. (2019). **Top:** Linear model **Second:** Nonlinear regression model **Third:** High dimensional correlated signal **Bottom:** Data model is exactly a random forest trained on real data

An Efficient Modification

- Permutation tests allow for testing feature importance w/o variance estimation. Let $g_y(r) = (r-y)^2$. Then, the MSE (conditional on both x, y) is given by $g_y(RF(x))$.
- If $\exists$ sequence a_n such that

$$a_n\left[RF_{B_n}(x) - \mathbb{E}RF(x)\right] \xrightarrow{d} \mathcal{N}(0,1)$$

$$RF_{B_n}(x) \xrightarrow{p} \mathbb{E}RF(x)$$

then delta method applies and

$$a_n\left[g_y(RF_{B_n}(x)) - \mathbb{E}g_y(RF_{B_n}(x))\right] \xrightarrow{d} \mathcal{N}(0, (g_y'(\mathbb{E}RF(x))^2)$$

- Can show that permutation distribution of MSE statistic also approaches the same unconditional distribution

Power Simulations

- Procedure that permutes trees between a full and reduced forest attains high power and maintains Type I error rate across a variety of scenarios

Tim Coleman, Lucas Mentch

email: tsc35@pitt.edu, website: tim-coleman.github.io, LM email: lkm31@pitt.edu

University of Pittsburgh, Department of Statistics