An Efficient Permutation Test for Feature Significance in Random Forests

ASA Pittsburgh Chapter, Spring 2018 Banquet

Overview

- Random Forests produce highly accurate predictions with little tuning but are “black box" methods
- Valid inference (like the F-test for linear regression) remains computationally infeasible and restricted to limited hypotheses
- We utilize the *exchangeability* of tree predictions to provide a statistically valid framework for feature significance with a much lower computational cost

Variable Importance in Random Forests is Hard

- Typical measures of variable importance are neither statistically valid nor reliable heuristics (Strobl et al., 2007), but are widely used
- Closed form distributions for RF predictions have been derived, but require training enormous numbers of trees (Mentch and Hooker, 2016) or require simplifying the tree building process like “honest trees" in Wager and Athey (2017)

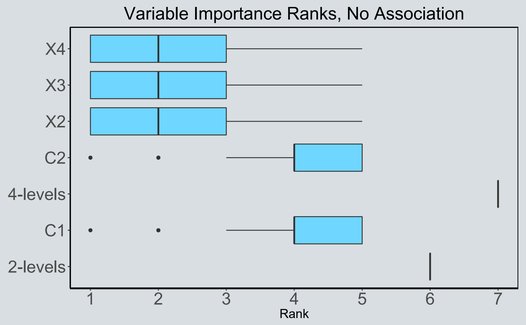


Figure: Ranking of covariates by RF importance, calculated on 500 RF's trained on independent simulations, where all predictors are independent of the response. Correlated (C_1, C_2) predictors are consistently ranked less important , as are predictors with fewer levels. Figure inspired by Strobl et al. (2007). Note: $X_2, X_3, X_4 \overset{iid}{\sim} Unif(0,1)$

The Intuition

- Standard permutation test would permute a feature of interest many times, creating a new random forest each time. We permute a feature a single time, and permute trees many times.
- Feature is insignificant $\implies$ tree predictions (and residuals) should be exchangeable between forests, p-values should be uniform.

References

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.

Friedman, J. H. (1991). Multivariate adaptive regression splines. *The Annals of Statistics*, 1–67.

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 (2017). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*

Theoretical Backing

- Permutation tests are non-parametric equivalents to the two sample t-test, i.e. given $\{X_i\}_{i=1}^n \overset{iid}{\sim} P$ and, independently, $\{Y_j\}_{j=1}^m \overset{iid}{\sim} P$, we test $H_0 : \mathbb{E}(X) = \mathbb{E}(Y)$, by shuffling the (X, Y) labels many times.
- Fundamentally, tests H_0 : data are exchangeable, which means the joint distribution of the data is invariant under permuting the order of the observations
- Fundamental result about exchangeable sequences:
  - ***DeFinetti's Theorem:*** *Data are exchangeable if and only if they are iid conditional on a latent variable*
  - Trees (and their predictions) in a random forest satisfy this; the latent variable is the observed dataset. A violation of exchangeability implies that trees are not identically distributed - correlation structure remains approximately the same.
- Present work is to prove asymptotic normality of a wide class of permutation statistics

The Procedure

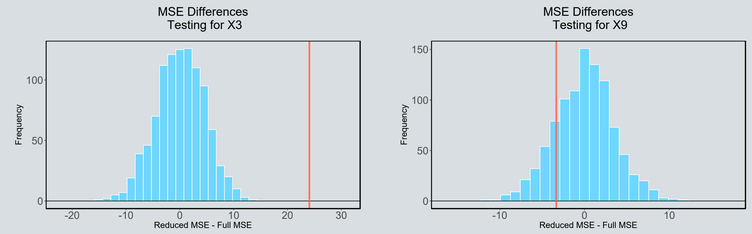
- Start with the original training data, X, and a test set, $\mathcal{T}$, and declare features of interest $X_s = \{X_{s_1}, ..., X_{s_k}\}$. Let $\alpha \in (0, 1)$.
  1. Train B trees on subsamples from original data, say $T_1, ...T_B$. Calculate: $MSE_0 = \sum_{y_t \in \mathcal{T}} \left(\frac{1}{B}\sum_{i=1}^B T_i - y_t\right)^2$
  2. Permute the features of interest X_s a single time, by row, creating X^π. Train B trees, say $T_1^\pi, ..., T_B^\pi$. Calculate: $MSE_0^\pi = \sum_{y_t \in \mathcal{T}} \left(\frac{1}{B}\sum_{i=1}^B T_i^\pi - y_t\right)^2$
  3. Now, permute the labels (i.e. switch the π labels) of the trees N_{mc} times, calculating $MSE_1^{*\pi}, ..., MSE_{N_{mc}}^{*\pi}$ and $MSE_1^*, ..., MSE_{N_{mc}}^*$
  4. Return:
$$\tilde{p} = \frac{1}{N_{mc}+1}\left[1 + \sum_{j=1}^{N_{mc}} I\Big((MSE_0^\pi - MSE_0)) > (MSE_j^{*\pi} - MSE_j^*)\Big)\right]$$
  5. Conclude that at least one predictor in X_s provides significant accuracy improvements if $\tilde{p} < \alpha$
- Under H_0, $\mathbb{E}(\tilde{p}) \approx \alpha$, and approximation becomes exact under assumption that all covariates are independent, so permuting a covariate is like adding a “knock-off"

Simulation Models

- Let $X_1, ..., X_5 \overset{iid}{\sim} Unif(0,1)$, and $X_6, ..., X_{10} \overset{iid}{\sim} Multi(1/3, 1/3, 1/3)$.
- Consider two models: 1) linear model and 2) the MARS model, from Friedman (1991).
  - Linear model: $Y = \beta(X_1 + I(X_6 = 2)) + \epsilon$
  - MARS: $Y = \beta\Big[\sin(\pi X_1 I(X_7 = 2)) + 2(X_3 - .05)^2 + X_4 + X_2 + \frac{1}{2}I(X_8 = 3)\Big] + \epsilon$
- We let $\epsilon \overset{iid}{\sim} \mathcal{N}(0, \mathbb{SD}^2(\epsilon))$, and $\beta = 10$. Each model has $N = 2000$ training examples and $|\mathcal{T}| = 100$ test examples.
- We test for a variety of $\mathbb{SD}(\epsilon)$, namely $\left[2\beta, \frac{4}{3}\beta, \beta, \frac{2}{3}\beta, \frac{1}{5}\beta\right]$.

Simulation Results

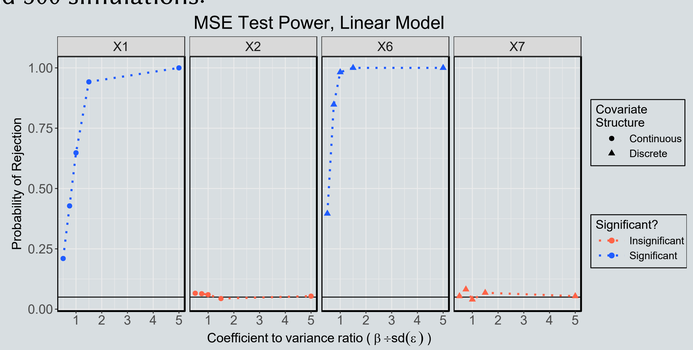
- First, we test for two features in the MARS model, X_3, which is important, and X_9 which is unimportant.



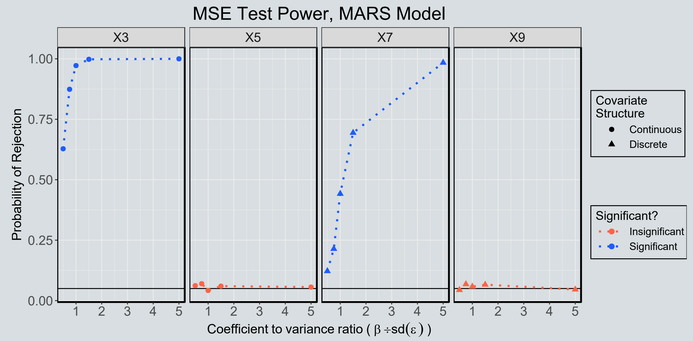
(a) Test for X_3; red line represents observed difference in MSE, $\tilde{p} = 0.00099$

(b) Test for X_9; red line represents observed difference in MSE, $\tilde{p} = 0.824$

- Now we test for X_1, X_2, X_6, X_7 in the linear model using $B = 125$ trees and 500 simulations:



- Now testing for X_3, X_7, X_5, X_9 in the MARS model:



Conclusions

- P-values appear uniform under H_0, empirical backing for theoretical asymptotic validity, and **attains high power with only 250 trees!**
- Future work: explicit importance measures based off this test

Tim Coleman
tsc35@pitt.edu
University of Pittsburgh, Department of Statistics, Adviser: Lucas Mentch