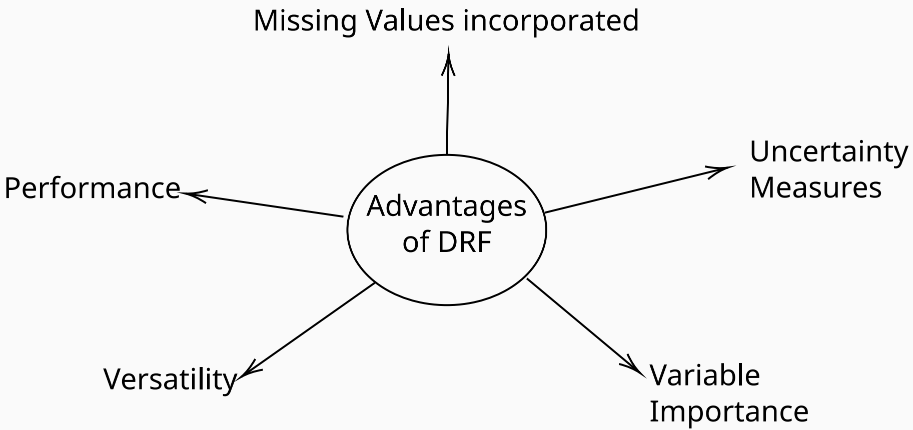
Distributional Random Forests

Jeffrey Näf (*Inria*)

January 4, 2024

DRF Features



Advantages of DRF

Missing Values incorporated

Uncertainty Measures

Variable Importance

Versatility

Performance

1. **Random Forests (RF)**

2. Distributional Random Forest (DRF)

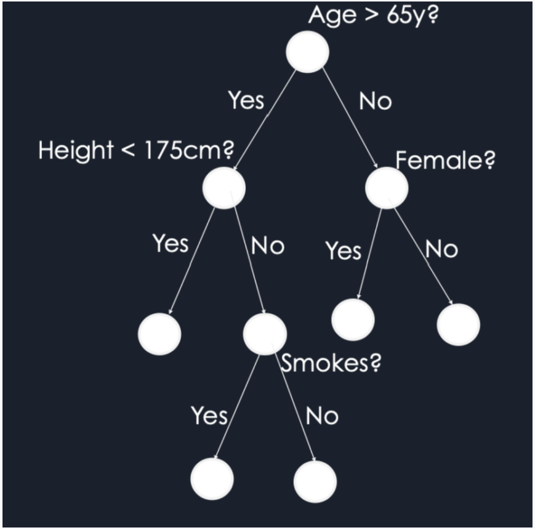
3. Example

4. Uncertainty Assessment

5. Variable Importance

6. Conclusion

Decision Tree



Random Forest (RF) of Breimann (2001)

–> Want to learn the conditional expectation of $Y \in \mathbb{R}$ given covariates $\mathbf{X} \in \mathbb{R}^p$ from i.i.d observations $(Y_1, \mathbf{X}_1), \ldots, (Y_n, \mathbf{X}_n)$

–> Two steps:

1. Construct a forest with N trees
2. Predict for a test point $\mathbf{x}$

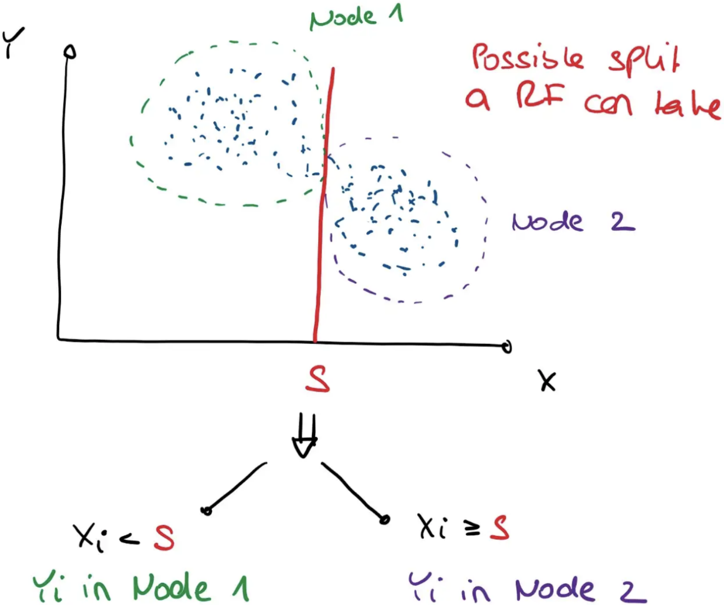
1. Forest construction

–> Fit N trees

–> Each tree splits the $Y_i's$ according to some rule depending on the covariates.

–> Conventional RF uses the CART criterion, which compares the means of Y in the two child nodes.

–> The split is taken where the squared difference in means is maximized.


Node 1
Y
Possible split
a RF can take
Node 2
S
X
$X_i < S$
$X_i \geq S$
Yi in Node 1
Yi in Node 2

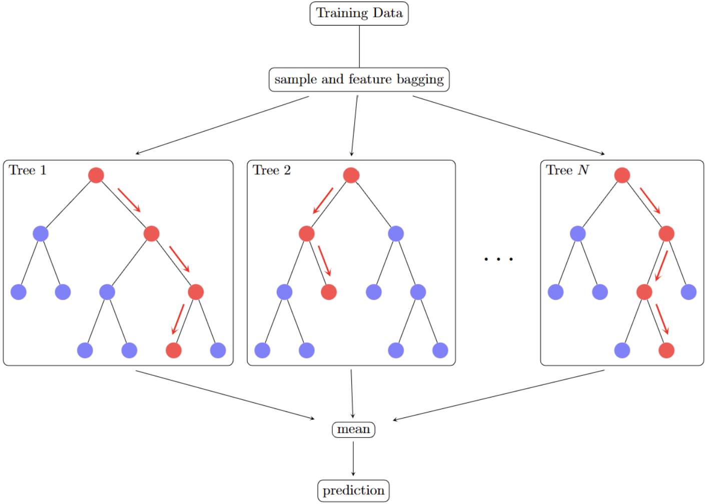
2. Prediction

–> Drop test point **x** in all trees $k = 1, \ldots, N$

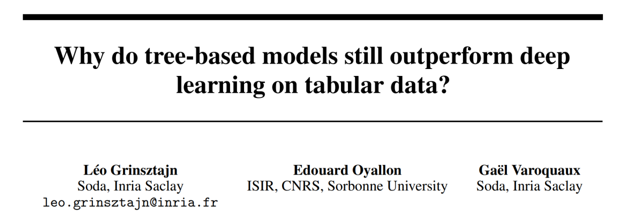
–> Let $\mathcal{L}_k(\mathbf{x})$ be the leaf where it falls.

–> Average all Y_i for $i \in \mathcal{L}_k(\mathbf{x})$ to get a prediction for each tree

–> Average over all N trees


Training Data
sample and feature bagging
Tree 1
Tree 2
. . .
Tree N
mean
prediction

State-of-the-Art performance on Tabular Data

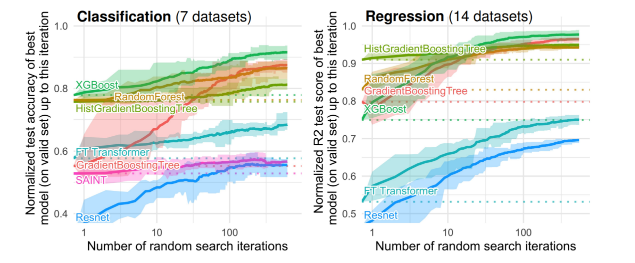


Why do tree-based models still outperform deep learning on tabular data?

Léo Grinsztajn
Soda, Inria Saclay
leo.grinsztajn@inria.fr

Edouard Oyallon
ISIR, CNRS, Sorbonne University

Gaël Varoquaux
Soda, Inria Saclay



1. Random Forests (RF)

2. Distributional Random Forest (DRF)

3. Example

4. Uncertainty Assessment

5. Variable Importance

6. Conclusion

Distributional Random Forest (DRF)

–> Let's say we want to predict at **x**

–> RF implicitly also produces weights $w_i(\mathbf{x})$, $i = 1, \ldots, n$, indicating the importance of point *i* for this prediction:

$$w_i(\mathbf{x}) = \frac{1}{N} \sum_{k=1}^{N} \frac{\mathbf{1}\{\mathbf{X}_i \in \mathcal{L}_k(\mathbf{x})\}}{\#\mathcal{L}_k(\mathbf{x})}$$

–> Can write the prediction as

$$\widehat{\mathbb{E}}[Y \mid \mathbf{X} = \mathbf{x}] = \sum_{i=1}^{n} w_i(\mathbf{x}) Y_i.$$

$\Longrightarrow$ RF is a nearest neighborhood method with a data-adaptive notion of neighborhood.

Distributional Random Forest (DRF)

–> Can use the weights to approximate other things than conditional expectations

–> Example: Conditional quantiles [Meinshausen, 2006]

–> However, doing this it might make sense to adapt the splitting criterion!

–> Generalized Random Forest (GRF) of [Athey et al., 2019]: Define an estimation target and adapt the splitting criterion by this target

–> DRF: Define one splitting criterion that makes sense for many targets.

CART

CART criterion:

$$\min_{\text{splits}} \frac{1}{n_P} \left(\sum_{i \in C_L} (Y_i - \overline{Y}_L)^2 + \sum_{i \in C_R} (Y_i - \overline{Y}_R)^2 \right) \tag{1}$$

is equivalent to

$$\max_{\text{splits}} \frac{n_L n_R}{n_P^2} \left(\frac{1}{n_L} \sum_{i \in C_L} Y_i - \frac{1}{n_R} \sum_{i \in C_R} Y_i \right)^2 . \tag{2}$$

$\Longrightarrow$ Splits are chosen to make the means in the child nodes as different as possible.

Splitting Criteria

–> RF:

$$\frac{n_L n_R}{n_P^2}\left(\bar{Y}_L - \bar{Y}_R\right)^2$$

–> GRF:

$$\frac{n_L n_R}{n_P^2}\left(\hat{\tau}_L - \hat{\tau}_R\right)^2$$

Idea of DRF: Do CART but with means in a Reproducing Kernel Hilbert space (RKHS)!

MMD Criterion

Idea of DRF: Do CART but with means in a Reproducing Kernel Hilbert space (RKHS)!

–> RKHS $\mathcal{H}$ is a Hilbert-space defined by a kernel $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$

–> Any probability measure P can be mapped to an expectation in $\mathcal{H}$!

–> For certain choices of k learning this expectation is akin to learning the distribution!

–> This is the idea of DRF: We use CART in $\mathcal{H}$ and estimate the conditional expectation in $\mathcal{H}$.

MMD Criterion

–> Let Φ be the function that takes a probability measures and maps it into $\mathcal{H}$.

–> For the dirac measure $\Phi(\delta_{\mathbf{Y}_i}) = k(\mathbf{Y}_i, \cdot)$:

$$\max_{\text{split}} \frac{n_L n_R}{n_P^2} \left\| \Phi\left(\frac{1}{|n_L|}\sum_{i\in C_L} \delta_{\mathbf{Y}_i}\right) - \Phi\left(\frac{1}{|n_R|}\sum_{i\in C_R} \delta_{\mathbf{Y}_i}\right) \right\|_{\mathcal{H}}^2 =$$

$$\max_{\text{split}} \frac{n_L n_R}{n_P^2} \left\| \frac{1}{|n_L|}\sum_{i\in C_L} k(\mathbf{Y}_i, \cdot) - \frac{1}{|n_R|}\sum_{i\in C_R} k(\mathbf{Y}_i, \cdot) \right\|_{\mathcal{H}}^2$$

$\Longrightarrow$ Splits are chosen to make the means in the child nodes as different as possible, but now in the *Hilbert Space*.

DRF Estimator

–> As a consequence, we get an estimate of the conditional mean embedding (CME)

$$\mu(\mathbf{x}) = \Phi(\mathbb{P}_{\mathbf{Y}|\mathbf{X}=\mathbf{x}}) = \mathbb{E}[k(\mathbf{Y}, \cdot) \mid \mathbf{X} = \mathbf{x}]$$

–> This has the form

$$\hat{\mu}_n(\mathbf{x}) = \sum_{i=1}^{n} w_i(\mathbf{x}) k(\mathbf{Y}_i, \cdot) \in \mathcal{H}$$

–> This can easily be translated back into the empirical distribution:

$$\hat{\mathbb{P}}_{\mathbf{Y}|\mathbf{X}=\mathbf{x}} = \sum_{i=1}^{n} w_i(\mathbf{x}) \delta_{\mathbf{Y}_i}$$

–> Access to $\hat{\mathbb{P}}_{\mathbf{Y}|\mathbf{X}=\mathbf{x}}$ is nice because a large range of targets can be calculated from it!

DRF Estimator: Summary

–> i.i.d data $(\mathbf{Y}_1, \mathbf{X}_1), \dots, (\mathbf{Y}_n, \mathbf{X}_n)$, $\mathbf{Y} \in \mathbb{R}^d$ and $\mathbf{X} \in \mathbb{R}^p$

–> Random Forest (RF) is a powerful tool to estimate $\widehat{\mathbb{E}}[Y \mid \mathbf{X} = \mathbf{x}]$, for $d = 1$

–> Idea of DRF: Use a RF in a Reproducing Kernel Hilbert space (RKHS) $\mathcal{H}$

–> Learning the conditional expectation in this space
= Learning a representation of the conditional distribution $\mathbb{P}_{\mathbf{Y}|\mathbf{X}=\mathbf{x}}$

–> Resulting estimate can be conveniently written as

$$\hat{\mathbb{P}}_{\mathbf{Y}|\mathbf{X}=\mathbf{x}} = \sum_{i=1}^{n} w_i(\mathbf{x}) \delta_{\mathbf{Y}_i}$$

with weights $w_i(\mathbf{x})$, $i = 1, \dots, n$, indicating the importance of point i

–> This also works when $\mathbf{Y}$ takes values in $\mathbb{R}^d$, for $d > 1$!

DRF Estimator for a target τ

$$\begin{tikzcd}
\mathbb{P}_{\mathbf{Y} \mid \mathbf{X}=\mathbf{x}} \arrow[r, "\text{1) get } w_{\mathbf{x}}(\cdot) \text{ with DRF}"] \arrow[d, dashed, "\text{objective}"'] & \hat{\mathbb{P}}_{\mathbf{Y} \mid \mathbf{X}=\mathbf{x}} \arrow[d, "\text{2) compute from } w_{\mathbf{x}}(\cdot)"] \\
\tau(\mathbb{P}) & \tau(\hat{\mathbb{P}}) \arrow[l, dashed, "\text{induced estimator}"']
\end{tikzcd}$$

DRF Estimator

$\mathbf{Y} = (\text{O3}, \text{SO2}, \text{PM2.5})$, $\mathbf{X} = (\text{longitude}, \text{latitude}, \text{elevation}, \text{location setting}, \ldots)$

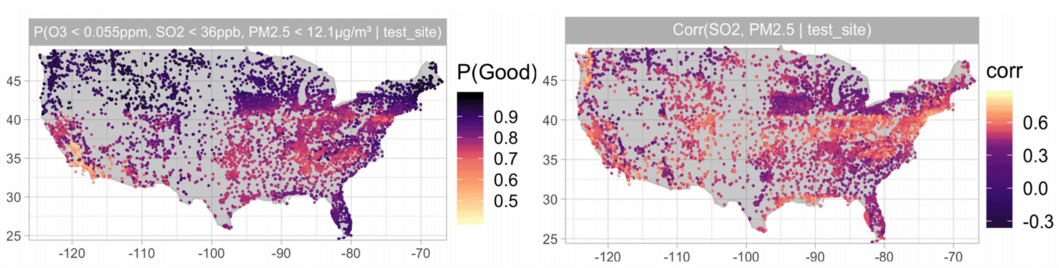

P(O3 < 0.055ppm, SO2 < 36ppb, PM2.5 < 12.1µg/m³ | test_site)
P(Good)
0.9
0.8
0.7
0.6
0.5
Corr(SO2, PM2.5 | test_site)
corr
0.6
0.3
0.0
-0.3
45
40
35
30
25
-120
-110
-100
-90
-80
-70

Figure 5: Estimates of the probability $\mathbb{P}(\text{AQI} \leq 50 \,|\, \text{test site})$ (left) and the conditional correlation (right) derived from the DRF estimate of the multivariate conditional distribution.

1. Random Forests (RF)

2. Distributional Random Forest (DRF)

3. Example

4. Uncertainty Assessment

5. Variable Importance

6. Conclusion

Example: Fairness

⚠

Warning:

The following analysis is for illustration purposes only and should not be repeated unsupervised. Side effects include overconfidence in nonexisting effects. Please refer to your local statistician for further information.

Example: Fairness

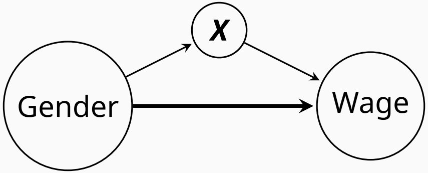
–> Is observed gender gap in wage due to gender alone, or can it be explained by other factors (e.g. different industries)?

–> Want to find the distribution of the nested counterfactual

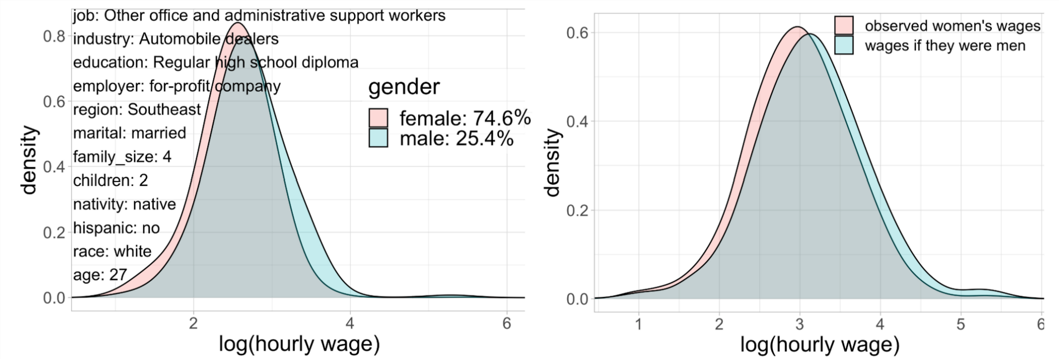
$$W(male, \mathbf{X}(female)),$$

women's wage had they been treated the same way as men for setting the wage

$$\mathbb{P}\left(W(\text{male}, \mathbf{X}(\text{female}))\right) = \int \mathbb{P}\left(W \mid G = \text{male}, \mathbf{X} = \mathbf{x}\right) \mathbb{P}(\mathbf{X} = \mathbf{x} \mid G = \text{female}) d\mathbf{x}, \tag{3}$$



Example: Fairness



1. Random Forests (RF)

2. Distributional Random Forest (DRF)

3. Example

4. **Uncertainty Assessment**

5. Variable Importance

6. Conclusion

Theoretical View

–> To recap, DRF estimates the mean $\mu(\mathbf{x}) = \mathbb{E}[k(\mathbf{Y}, \cdot) \mid \mathbf{X} = \mathbf{x}]$ in the Hilbert space, as

$$\hat{\mu}_n(\mathbf{x}) = \sum_{i=1}^{n} w_i(\mathbf{x}) k(\mathbf{Y}_i, \cdot) \in \mathcal{H}.$$

–> Though $\mathcal{H}$ is an infinite-dimensional Hilbert space for the “interesting” choices of k, $\hat{\mu}_n(\mathbf{x})$ can be thought of as a weighed mean in Euclidean space!

–> Can we have a consistency or even asymptotic normality result?

Result

Theorem

Assume a certain list of conditions holds. Then, there exists $\sigma_n > 0$, $\sigma_n \to 0$, such that

$$\frac{1}{\sigma_n}\left(\hat{\mu}_n(\mathbf{x}) - \mu(\mathbf{x})\right) \xrightarrow{D} N(0, \boldsymbol{\Sigma}_{\mathbf{x}}), \tag{4}$$

where $\boldsymbol{\Sigma}_{\mathbf{x}}$ is a self-adjoint HS operator satisfying

$$\langle \boldsymbol{\Sigma}_{\mathbf{x}} f, f \rangle = \frac{\mathbb{V}(\langle k(\mathbf{Y},\cdot), f\rangle | \mathbf{X} = \mathbf{x})}{\mathbb{V}(k(\mathbf{Y},\cdot) | \mathbf{X} = \mathbf{x})} > 0 \tag{5}$$

for all $f \in \mathcal{H}$.

Consequence

–> We thus have an idea about the asymptotic distribution of our estimator

–> In particular,

(I) For any smooth enough functional $F : \mathcal{H} \to \mathbb{R}^q$,

$$\frac{1}{\sigma_n}(F(\hat{\mu}_n(\mathbf{x})) - F(\mu(\mathbf{x})))$$

is asymptotically normal

(II) For any continuous functional $F : \mathcal{H} \to \mathbb{R}^q$,

$$F\left(\frac{1}{\sigma_n}(\hat{\mu}_n(\mathbf{x}) - \mu(\mathbf{x}))\right) \xrightarrow{D} F(N(0, \mathbf{\Sigma}_{\mathbf{x}}))$$

Consequence

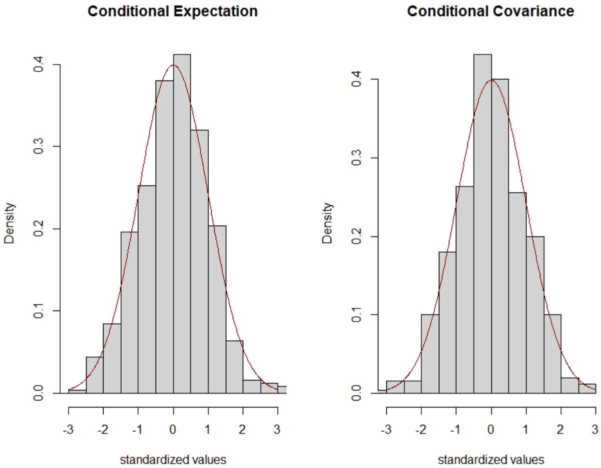
–> (I) Motivates the approximation of the sampling distribution of targets

$$\tau(\hat{\mathbb{P}}_{\mathbf{Y}|\mathbf{X}=\mathbf{x}}) = F(\hat{\mu}_n(\mathbf{x})),$$

such as

- conditional means
- conditional quantiles
- conditional variance-covariance matrices

by a normal distribution (even though depending on the kernel these might not be smooth enough functions (!))



Conditional Expectation

Conditional Covariance

Figure: Simulated Example with $d = 2$, $p = 2$. The histogram shows estimated values minus truth, standardized by the estimated standard deviation, for 500 repetitions. Left: Conditional Expectation of Y_1, Right: Conditional Covarance between Y_1, Y_2.

Consequence

–> (II) is used to approximate the distribution of

$$\left\| \frac{1}{\sigma_n} \left(\hat{\mu}_n(\mathbf{x}) - \mu(\mathbf{x}) \right) \right\|_{\mathcal{H}}^2 ,$$

which can be used to build *simultaneous* CIs around $\hat{\mu}_n(\mathbf{x})(\mathbf{y})$, see later

Problem

–> The abstract result above doesn't really help us in approximating the asymptotic distributions

–> For (I) we lack estimates of variances, for (II) there is not even a tractable way of writing the asymptotic distribution ($\|N(0, \mathbf{\Sigma}_{\mathbf{x}})\|^2_{\mathcal{H}}$)

–> Solution: Subsample the subsampling!

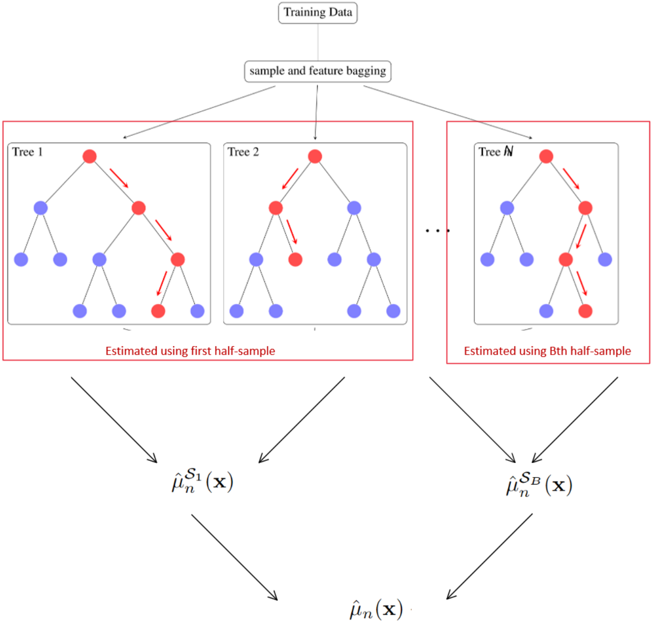
Solution

–> Instead of just fitting N trees to build our forest, we build B groups of L trees (such that $N = B * L$).

–> For each group of trees or mini forests, we subsample at random about half of the data points and then fit the forest using only this subsample.

–> Let's call this subset of samples chosen $\mathcal{S}$

–> For each drawn $\mathcal{S}$, we then get another DRF estimator in the Hilbert space denoted $\hat{\mu}_n^{\mathcal{S}}(\mathbf{x})$



Training Data

sample and feature bagging

Estimated using first half-sample

Estimated using Bth half-sample

$\hat{\mu}_n^{\mathcal{S}_1}(\mathbf{x})$

$\hat{\mu}_n^{\mathcal{S}_B}(\mathbf{x})$

$\hat{\mu}_n(\mathbf{x})$

Solution

–> Thus we have B groups of trees resulting in the estimators $\hat{\mu}_n^{\mathcal{S}_1}(\mathbf{x}), \ldots, \hat{\mu}_n^{\mathcal{S}_B}(\mathbf{x})$.

–> The overall estimate can be build out of the average of those B estimators:

$$\hat{\mu}_n(\mathbf{x}) = \frac{1}{B} \sum_{b=1}^{B} \hat{\mu}_n^{\mathcal{S}_b}(\mathbf{x})$$

–> The B draws can be used to approximate aspects of the sampling distribution.

Result

Theorem

Assume the same list of conditions hold. Then

$$\xi_n^{\mathcal{S}} = \frac{1}{\sigma_n}\left(\hat{\mu}_n^{\mathcal{S}}(\mathbf{x}) - \hat{\mu}_n(\mathbf{x})\right) \xrightarrow[W]{D} N(0, \mathbf{\Sigma}_{\mathbf{x}}) \tag{6}$$

holds.

–> Such results are typical in Bootstrap arguments

–> In words: For fixed data, only considering the randomness of $\mathcal{S}$, $\hat{\mu}_n^{\mathcal{S}}(\mathbf{x}) - \hat{\mu}_n(\mathbf{x})$, appropriately scaled, is asymptotically normal.

Result

2 main insights:

1. Given the data and only considering the randomness of $\mathcal{S}$,

$$\frac{1}{\sigma_n}\left(\hat{\mu}_n^{\mathcal{S}}(\mathbf{x}) - \hat{\mu}_n(\mathbf{x})\right)$$

   has the same distributional limit as

$$\frac{1}{\sigma_n}\left(\hat{\mu}_n(\mathbf{x}) - \mu(\mathbf{x})\right)$$

2. We can simulate from the former distribution by drawing $\mathcal{S}$!

$\Longrightarrow$ We can approximate the (asymptotic) distribution of $\hat{\mu}_n(\mathbf{x})$ by simulating $\hat{\mu}_n^{\mathcal{S}}(\mathbf{x})$ B times!

Full algorithm

Thus, with our new methodology, we get in one fell swoop:

–> The estimate $\hat{\mu}_n(\mathbf{x})$ that can be translated back into an estimate of the conditional distribution $\hat{\mathbb{P}}_{\mathbf{Y}|\mathbf{X}=\mathbf{x}}$

–> An i.i.d. sample from the sampling distribution $\hat{\mu}_n^{\mathcal{S}_1}(\mathbf{x}), \ldots, \hat{\mu}_n^{\mathcal{S}_B}(\mathbf{x})$

- Complexity of this approach is (almost) the same as for the original Random Forest: $\mathcal{O}(B \times N \times p \times n \log n)$ (!)
- However, need *a lot* of trees (large N)

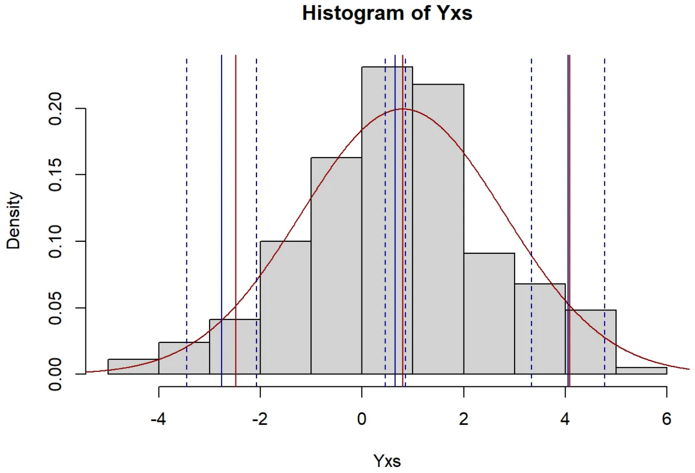


Figure: Histogram of the simulated conditional distribution overlaid with the true density (in red). Additionally, the estimated conditional expectation and the conditional (0.05, 0.95) quantiles are in blue, with true values in red. Moreover, the dashed red lines are the confidence intervals for the estimates as calculated by DRF.

1. Random Forests (RF)

2. Distributional Random Forest (DRF)

3. Example

4. Uncertainty Assessment

5. Variable Importance

6. Conclusion

Variable Importance in RF

–> Since their inception RFs were able to provide a notion of importance of the features in **X**

–> Until recently these were mostly Mean Decrease Accuracy (MDA) and Mean Decrease Impurity (MDI)

–> Recently there has been renewed interest in such variable importance measures for nonparametric models

–> In particular, [Bénard et al., 2022] demonstrate the inconsistency of the classical importance measures for RF and develop a principled variable importance measure (“Sobol-MDA”)

Importance measure for Distributional Prediction

–> We would like to define a meaningful notion of variable importance for DRF

–> Previous measures, such as (Sobol-)MDA, are designed for conditional expectation estimation

–> Idea adapted from [Da Veiga, 2021]: Measure the distance of

$\mu(\mathbf{X}^{(-j)})$, the estimate when variable X_j is removed

$\mu(\mathbf{X})$, the estimate with all variables

to obtain:

$$\mathrm{I}^{(j)} = \frac{\mathbb{E}[\|\mu(\mathbf{X}) - \mu(\mathbf{X}^{(-j)})\|^2_{\mathcal{H}}]}{\mathbb{E}[\|\mu(\mathbf{X}) - \mathbb{E}[\mu(\mathbf{X})]\|^2_{\mathcal{H}}]} \tag{7}$$

Importance measure for Distributional Prediction

Alternative Formulation: Measure the distance of

- $\mathbb{P}_{\mathbf{Y}|\mathbf{X}^{(-j)}}$, the estimate when variable X_j is removed
- $\mathbb{P}_{\mathbf{Y}|\mathbf{X}}$, the estimate with all variables

to obtain:

$$\mathrm{I}^{(j)} = \frac{\mathbb{E}[\mathsf{MMD}^2(\mathbb{P}_{\mathbf{Y}|\mathbf{X}}, \mathbb{P}_{\mathbf{Y}|\mathbf{X}^{(-j)}})]}{\mathbb{E}[\mathsf{MMD}^2(\mathbb{P}_{\mathbf{Y}}, \mathbb{P}_{\mathbf{Y}|\mathbf{X}})]} \tag{8}$$

Importance measure for Distributional Prediction

–> With DRF $\mathrm{I}^{(j)}$ can easily be estimated by the drop-and-relearn principle:

–> Let $\hat{\mu}_n(\mathbf{x}_i^{(-j)})$ be the forest retrained with the *j*th variable (X_j) removed

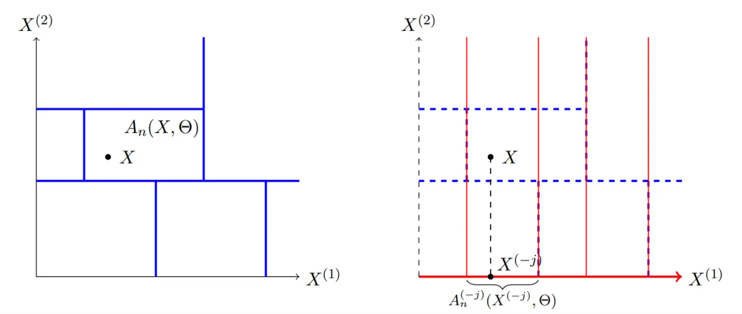
–> Using an independent sample $\mathbf{X}'_1, \ldots \mathbf{X}'_n$, we define

$$\mathrm{I}_n^{(j)} = \frac{\sum_{i=1}^n \|\hat{\mu}_n(\mathbf{X}'_i) - \hat{\mu}_n(\mathbf{X}_i'^{(-j)})\|_{\mathcal{H}}^2}{\sum_{i=1}^n \|\hat{\mu}_n(\mathbf{X}'_i) - \overline{\mu_n}\|_{\mathcal{H}}^2}. \qquad (9)$$

Importance measure for Distributional Prediction

–> Under relatively weak assumptions: $\mathrm{I}_n^{(j)} \xrightarrow{p} \mathrm{I}^{(j)}$

–> Moreover, one can implement the projection approach of [Bénard et al., 2022] to get a consistent estimator that is fast to compute, even for large *p*



Example

–> We simulate $n = 1000$ observations of $(\mathbf{Y}_i, \mathbf{X}_i)$

–> We use 10-dimensional **X**, with $X_i \sim \text{Unif}(-1, 1)$, $i = 1, 3, \ldots, 10$, $X_3 = X_1 + \text{Unif}(-1, 1)$

–> We simulate the following dependent variable

$$Y \sim N(0.8 \cdot \mathbf{1}(X_1 > 0), (1 + \mathbf{1}(X_2 > 0))^2) \tag{10}$$

- Dependence between X_1, X_3
- Only X_1 is relevant to estimate the conditional expectation
- Only X_1, X_2 are relevant to estimate the conditional distribution (quantiles (!))

Example

–> We use 10-dimensional **X**, with $X_i \sim$ Unif$(-1,1)$, $i = 1,3,\dots,10$, $X_3 = X_2 +$ Unif$(-1,1)$ and

$$Y \sim N(0.8 \cdot \mathbf{1}(X_1 > 0), (1 + \mathbf{1}(X_2 > 0))^2) \tag{11}$$

- **Conditional Expectation:** MDA of [Breiman, 2001] quantifies correctly that X_1 is most important, but considers X_3 in second place
- **Conditional Expectation:** Sobol-MDA of [Bénard et al., 2022] quantifies correctly that X_1 is most important and all others are not relevant
- **Conditional Distribution:** $\mathrm{I}_n^{(j)}$ quantifies correctly that both X_1 and X_2 are important and all others are not relevant.

1. Random Forests (RF)

2. Distributional Random Forest (DRF)

3. Example

4. Uncertainty Assessment

5. Variable Importance

6. Conclusion

Current Work

–> Implementing the Missing Values Incorporated (MIA) method results, DRF can be used even with missing values

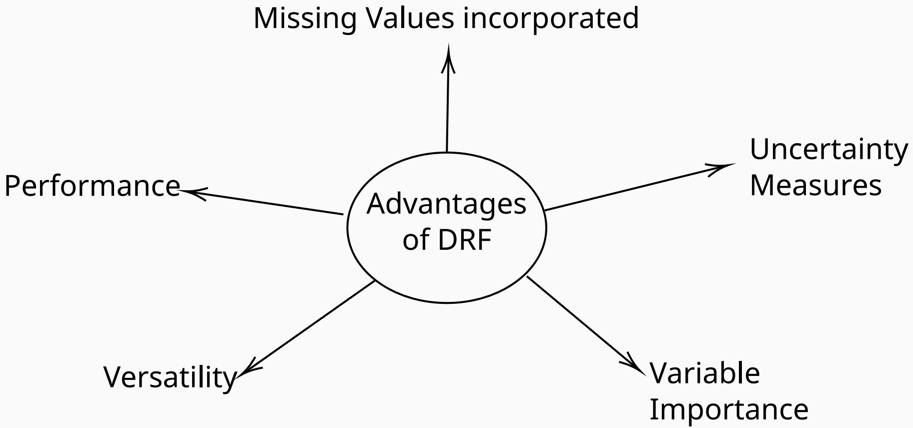
–> Trying to theoretically guarantee asymptotic normality for a wide range of targets.

–> Trying to improve sample efficiency, especially for uncertainty estimates

–> Trying new applications, such as weather prediction

–> Package needs a heavy update

DRF Features



Advantages of DRF

Missing Values incorporated

Uncertainty Measures

Variable Importance

Versatility

Performance

Software and Further reading

–> drf package on CRAN (works, but outdated)

–> Updated but poorly written code: `https://github.com/JeffNaef/drfupdate`

–> Medium Articles: `https://medium.com/@jeffrey_85949`

References I

- Athey, S., Tibshirani, J., and Wager, S. (2019).
  Generalized random forests.
  The Annals of Statistics, 47(2):1148–1178.

- Breiman, L. (2001).
  Random forests.
  Machine learning, 45(1):5–32.

- Bénard, C., Da Veiga, S., and Scornet, E. (2022).
  Mean decrease accuracy for random forests: inconsistency, and a practical solution via the Sobol-MDA.
  Biometrika, 109(4):881–900.

References II

Da Veiga, S. (2021).
Kernel-based anova decomposition and shapley effects – application to global sensitivity analysis.

Meinshausen, N. (2006).
Quantile regression forests.
Journal of Machine Learning Research, 7(Jun):983–999.