Tree-based models: From CART to random forests

Dongjae Son

Yonsei University

Department of Applied Statistics

February 25, 2020

Before we go on

- These slides rely heavily on [Hastie et al., 2009], [Bishop, 2006] and [Murphy, 2012]

Introduction: Adaptive basis function models

- Linear regression models:
  - Models conditional mean given covariates
  - Easy to handle and interpret
  - Various regularization methods exist: Ridge, Lasso, Elastic net, etc.
- Disadvatages of the linear models:
  - Relationship between target variable and predictors are not always linear
  - Not easy to incorporate interaction effect

$\therefore$ Bad prediction performance on nonlinear patterns

Introduction: Generalized Additive Models(GAM)

- Generalzied Additive Models(GAM) by [Hastie and Tibshirani, 1987]
  - $g\left(E\left[Y|X_1, X_2, \cdots, X_p\right]\right) = \alpha + \sum_{j=1}^{p} f_j\left(X_j\right)$
  - Choice of g can be:
    1. $g(\mu) = \mu$ is the identity link for continuous response
    2. $g(\mu) = \log\left(\frac{\mu}{1-\mu}\right)$ is the logit link for binary response
    3. $g(\mu) = \log\left(\mu\right)$ is for log-linear models for Poisson count data

    $\rightarrow$ Same with the choice of link functions in GLM

Introduction: Generalized Additive Models(GAM, cont'd)

- Fitting additive models: use scatterplot smoother, usually smoothing splines
  - $y = \alpha + \sum_{j=1}^{p} f_j\left(X_j\right) + \epsilon$ where $E\left[\epsilon\right] = 0$
  - Minimize the penalized residual sum of squares(PRSS)

$$J = \sum_{i=1}^{N}\left(y_i - \alpha - \sum_{j=1}^{p} f_j\left(X_j\right)\right)^2 + \sum_{j=1}^{p}\lambda_j \int f_j''(x_j)dx_j$$

    where $\lambda_j \geq 0$ are tuning parameters
  - α is not uniquely determined; add the constraint $\sum_{i=1}^{N} f_j\left(x_{ij}\right) = 0$ for all $j \longrightarrow \hat{\alpha} = \bar{y}$ and it never changes

Introduction: Generalized Additive Models(GAM, cont'd)

- Solving the PRSS: the backfitting algorithm
  1. Initialize: $\hat{\alpha} = \frac{1}{N}\sum_{i=1}^{N} y_i$ and $\hat{f}_j = 0$ for all j
  2. Cycle: $j = 1, 2, \cdots, p, \cdots, 1, 2, \cdots, p, \cdots$

$$\hat{f}_j \leftarrow S_j \left[\{y_i - \hat{\alpha} - \sum_{k \neq j} \hat{f}_k (x_{ik})\}_1^N \right]$$

$$\hat{f}_j \leftarrow \hat{f}_j - \frac{1}{N} \sum_{i=1}^{N} \hat{f}_j (x_{ij})$$

until the functions $\hat{f}_j$ change less than a prespecified threshold, say, 10^{-3}. The second line of the cycle process ensures the constraint $\sum_{i=1}^{N} f_j (x_{ij}) = 0$ for all j.

Introduction: Generalized Additive Models(GAM, cont'd)

Although the GAM looks fascinating, it has certain disadvantages:

1. Too much computational costs if we introduce interaction effects

$$g\left(\mu(X)\right) = \alpha + \sum_{j=1}^{p} f_j\left(X_j\right) + \sum_{j,k} f_{jk}(X_j, X_k) + \sum_{j,k,l} f_{jkl}(X_j, X_k, X_l) + \cdots$$

2. Not suitable for large p problems: COSCO procedures([Lin et al., 2006]) or SpAM(Sparse Additive Model, [Ravikumar et al., 2009]) approach

Classification and Regression Trees

- Tree-based models predict y with features $\boldsymbol{X} \in \mathbb{R}^p$ by dividing the feature space into disjoint rectangles R_m, or leaves of a tree.
  - $\hat{f}(\boldsymbol{X}) = \sum_{m=1}^{M} c_m \mathbb{I}\{\boldsymbol{X} \in R_m\}$ where $\mathbb{I}$ is an indicator function
  - $\mathcal{F} = \cup_{m=1}^{M} R_m$ such that $R_m \cap R_l = \emptyset$ for $m \neq l$
  - R_m can be constructed by binary splitting of predictors, i.e. $\{X_j \leq s\}$
- Suggested by [Breiman et al., 1984]

Classification and Regression Trees(cont'd)

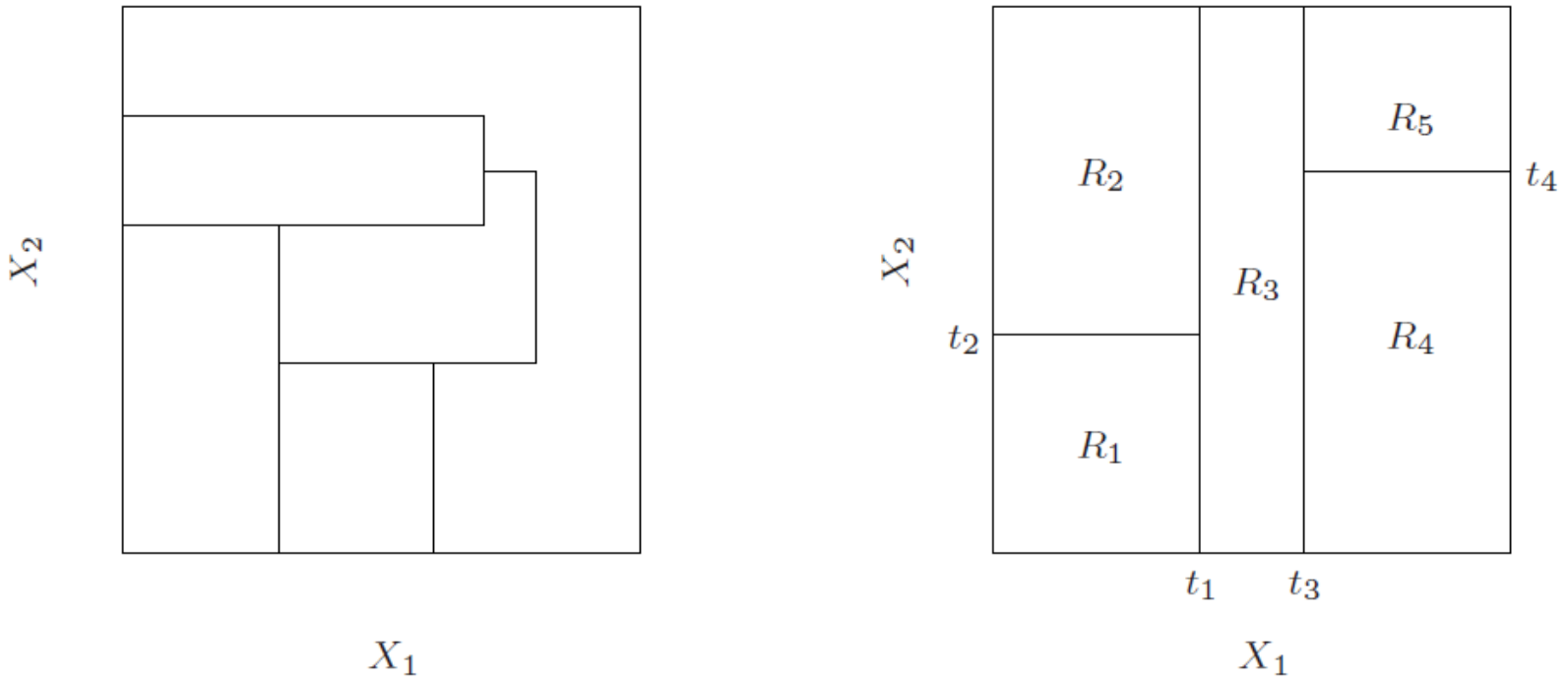

Figure: Partitioning and CART from [Hastie et al., 2009]

Classification and Regression Trees(cont'd)

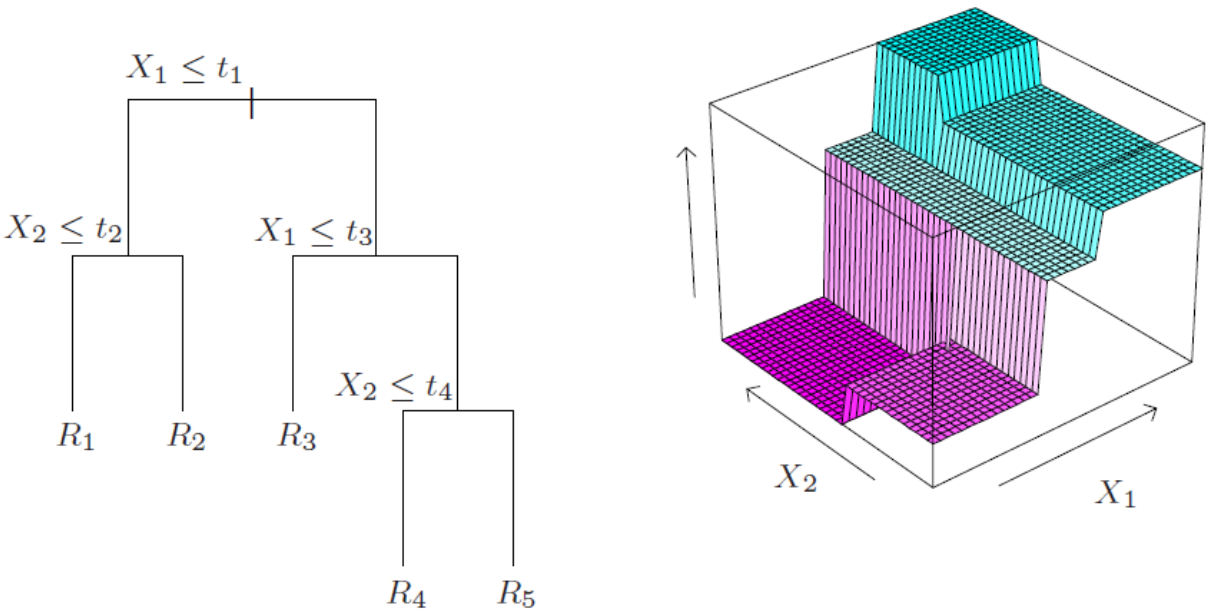

Figure: Partitioning and CART from [Hastie et al., 2009]

Regression Trees

- For regression trees, $c_m = \bar{y}_{R_m}$, just sample mean at rectangle R_m
- Thus, starting with all of the data, we seek the splitting variable j and split point s such that

$$\min_{j,s} \left[\sum_{x_i \in R_1(j,s)} \left(y_i - \bar{y}_{R_1(j,s)}\right)^2 + \sum_{x_i \in R_2(j,s)} \left(y_i - \bar{y}_{R_2(j,s)}\right)^2 \right]$$

  where $R_1(j,s) = \{X | X_j \leq s\}$ and $R_2(j,s) = \{X | X_j > s\}$
- Then, repeat this process on every resulting region and terminate if every leaf contains five or less observations

Regression Trees(cont'd)

- It can be rewritten as minimizing the loss function

$$R(T) = \sum_{m=1}^{|T|} \sum_{x_i \in R_m} (y_i - \bar{y}_{R_m})^2$$

  where $|T|$ is the tree size, or the number of leaves

- Tree size $|T|$ is a tuning parameter controlling the complexity of tree model, which should be controlled for generalization

Cost complexity pruning

- To control the tree size, we introduce the penalized loss

$$R_\alpha(T) = R(T) + \alpha|T|$$

  where $\alpha \geq 0$ is the tuning parameter that controls the trade-off between tree size and goodness-of-fit
- Idea: Find the subtree $T(\alpha) \subseteq T^0$ for each α that minimizes $R_\alpha(T)$ where T^0 is the fully grown tree

Cost complexity pruning(cont'd)

- For each α there exists a sequence of trees $T^0 \supseteq T^1 \supseteq \cdots \supseteq T^n$ where T^n is the null tree
- The sequence T^s can be generated by replacing a subtree T_t with root node t with a leaf

$$
\begin{aligned}
&[R_\alpha(T - T_t) - R_\alpha(T)] \\
&= R(T - T_t) - R(T) + \alpha(|T - T_t| - |T|) \\
&= R(T) - R(T_t) + R(t) - R(T) + \alpha(|T| - |T_t| + 1 - |T|) \\
&= R(t) - R(T_t) + \alpha(1 - |T_t|)
\end{aligned}
$$

Cost complexity pruning(cont'd)

- Solving $R(T) - R(T_t) + \alpha(1 - |T_t|) = 0$ yields $\alpha = \frac{R(t) - R(T_t)}{|T_t| - 1}$
- Starting with the full tree T^0 (and $\alpha_0 = 0$) in each step $s = 1, 2, \cdots$ the algorithm goes:
  1. Select the node t which minimizes $\frac{R(t) - R(T_t^{s-1})}{|T_t^{s-1}| - 1}$
  2. Set $T^s = T^{s-1} - T_t^{s-1}$ and $\alpha_s = \frac{R(t) - R(T_t^{s-1})}{|T_t^{s-1}| - 1}$
  3. Keep this process until we get the null tree
- Hence, we get $T^0 \supseteq T^1 \supseteq \cdots \supseteq T^n$ simultaneously with $\alpha_0 \leq \alpha_1 \leq \cdots \leq \alpha_n$. Determine the optimal α_* by 5- or 10-fold cross-validation.

V-fold cross-validation for optimal α_*

1. First, grow T_0 using the whole data $\mathcal{D}$ and get $T^0 \supseteq T^1 \supseteq \cdots \supseteq T^n$ simultaneously with $\alpha_0 \leq \alpha_1 \leq \cdots \leq \alpha_n = \infty$
2. Define $\alpha'_k = \sqrt{\alpha_k \alpha_{k+1}}$, a geometric mean
3. For each $v \in \{1, 2, \cdots, V\}$ and $\alpha \in \{\alpha'_1, \alpha'_2, \cdots\}$ use $\mathcal{D}^{(v)} = \mathcal{D} - \mathcal{D}_v$ to grow $T^{(v,0)}$ and find $T^{(v,k')}$ with corresponding sequence of α'_k
4. Choose α'_k that produces the minimum CV error then set $\alpha_* = \alpha_k$

V-fold cross-validation for optimal α_*(cont'd)

- We can use the sequence of geometric means since $T(\alpha) = T(\alpha_k)$ for $\alpha_k \leq \alpha < \alpha_{k+1}$
- Detailed proofs are omitted here. Ask me if you are curious how it works...!

Classification Trees

- For classification trees, c_m, is the modal class in the rectangle R_m among the class $\{1, 2, \cdots, K\}$
- The only difference between classification trees and regression trees can be found in *loss criteria*; no more squared-loss!
- Define
  - N_m: the number of samples in rectangle R_m
  - p_{mk}: the proportion of class k in rectangle R_m

Classification Trees(cont'd)

- Misclassification rate, or 0-1 loss:

$$R(T) = \sum_{m=1}^{|T|} \sum_{x_i \in R_m} \mathbb{I}\{y_i \neq c_m\}$$

- Gini index:

$$R(T) = \sum_{m=1}^{|T|} N_m \sum_{k=1}^{K} p_{mk}(1 - p_{mk})$$

- Cross-entropy:

$$R(T) = -\sum_{m=1}^{|T|} N_m \sum_{k=1}^{K} p_{mk} \log p_{mk}$$

Classification Trees(cont'd)

- Consider a node t with size N_t and loss criterion $R(t)$
- For some variable j and split point s, we split t into two nodes, t_R and t_L with size N_{t_R} and N_{t_L} and loss $R(t_R)$ and $R(t_L)$
- The mean decrease in loss can be defined as:

$$\triangle(j,s) = R(t) - \left(\frac{N_{t_R}}{N_t} R\left(t_R\right) + \frac{N_{t_L}}{N_t} R\left(t_L\right)\right)$$

- Find (j_*, s_*) such that

$$(j_*, s_*) = \arg\min_{j,s} \triangle(j,s)$$

Categorical predictors?

- For an un-ordered predictor variable with q levels, $2^{q-1} - 1$ splits are possible (Why?)
- `rpart` function in **R** deals with factors, but tree-based methods in **scikit-learn** do not.. need for dummification?

Further applications of CART

- Multivariate Adaptive Regression Splines(MARS): [Friedman, 1991]
- Hierarchical Mixture of Experts(HME): [Jordan and Jacobs, 1994]
- And many other researaches exist.. search for them if you find it interesting!

Bootstrap

- Before going into bagging methods, let's have a grasp on 'bootstrap' method
- The bootstrap is one of the key resampling methods in statistics
- It was mainly used to approximately estimate the variance of an estimator $\hat{\theta} = g\left(U_1, \cdots, U_N\right)$
- Key idea: Use empirical distribution $\hat{F}_N$ as a substitute for the true, unknown distrution F

Bootstrap(cont'd)

- For iid sample $U_1, \cdots, U_N \sim F$, the empirical distribution can be written as

$$\hat{F}_N(A) = \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}(U_i \in A)$$

- For univariate case, the Glivenko-Cantelli theorem says that

$$\left\| \hat{F}_N - F \right\|_\infty = \sup_{x \in \mathbb{R}} |\hat{F}_N(u) - F(u)| \to$$

almost surely as $N \to \infty$

Bootstrap(cont'd)

- Bootstrap variance estimator:
  1. Draw $\{U_i^b\}_{i=1}^n \sim \hat{F}_N$ randomly with replacement and compute $\hat{\theta}_b = g\left(U_1^b, \cdots, U_n^b\right)$
  2. Repeat this for B times
  3. Compute $\widehat{Var}\left(\hat{\theta}\right) = \frac{1}{B-1}\sum_{b=1}^{B}\left(\hat{\theta}_b - \bar{\theta}\right)^2$ where $\bar{\theta} = \frac{1}{B}\sum_{b=1}^{B}\hat{\theta}_b$
- Detailed explanations of boostrap(constructing confidence intervals or asymptotic results, etc) are omitted hereafter

Bootstrap Aggregation(Bagging)

- Idea: Use the bootstrap to enhance the prediction performance
- For training data $\boldsymbol{Z} = \{\left(\boldsymbol{X}^{(1)}, Y^{(1)}\right), \cdots, \left(\boldsymbol{X}^{(N)}, Y^{(N)}\right)\}$,
  1. Draw bootstrap samples $\boldsymbol{Z}^b$ for $b = 1, 2, \cdots, B$
  2. Obtain the bagging estimate

$$\hat{f}_{bag}(x) = \frac{1}{B} \sum_{b=1}^{B} \hat{f}^b(x)$$

where $\hat{f}^b(\cdot) = \hat{f}\left(\cdot, \boldsymbol{Z}^b\right)$ is an estimator trained with $\boldsymbol{Z}^b$ for $b = 1, 2, \cdots, B$

Bagging(cont'd)

- Basically, it works well with 'unstable' estimator, i.e. an estimator with low bias and high variance
- Intutively illustrating, if the variables $U_1, \cdots, U_B$ with positive correlation ρ are sampled from an identical distribution, its average has variance

$$Var\left(\bar{U}_B\right) = \rho\sigma^2 + \frac{1-\rho}{B}\sigma^2 \to \rho\sigma^2$$

  as $B \to \infty$
- For details, see [Breiman, 1996]

Random Forests

- First suggested by [Breiman, 2001]
- The key idea of random forests is to lower the variance of bagging estimate via reduction in correlation
- This is achieved by random selection of inputs at each split of each bootstrapped tree

Random Forests(cont'd)

1. For $b = 1, \cdots, B$:
   (a) Draw a bootstrap sample $\boldsymbol{Z}^b$ from the training data $\boldsymbol{Z}$
   (b) Build a tree $\hat{f}^b$ to the bootstrapped data $\boldsymbol{Z}^b$ without pruning, while choosing $m < p$ variables at each node (*Rule of thumb*: $m = p/3$ for regression and $m = \sqrt{p}$ for classification)
2. For the output $\{\hat{f}^b\}_{b=1}^B$:
   (a) Classification: $\hat{f}_{RF}(\boldsymbol{x}) = \arg\max_{j \in \{1, \cdots, K\}} \frac{1}{B} \sum_{b=1}^B \mathbb{I}\left(\hat{f}^b(\boldsymbol{x}) = j\right)$
   (b) Regression: $\hat{f}_{RF}(\boldsymbol{x}) = \frac{1}{B} \sum_{b=1}^B \hat{f}^b(\boldsymbol{x})$

How random forests decorrelates

- We can express $\hat{f}^b(\boldsymbol{x}) = \hat{f}(\boldsymbol{x}, \Theta_b)$ where Θ_b characterizes the b-th tree in terms of split variable, splitting points, terminal node values and $\boldsymbol{Z}^b$ itself
- From that point of view, we can consider Θ_b as a random variable
- However, distribution of Θ_b or Θ need not be specified since it would be only used theoretically

How random forests decorrelates(classification)

- By [Breiman, 2001], almost surely for all Θ

$$\frac{1}{B}\sum_{b=1}^{B}\mathbb{I}\left(\hat{f}\left(\boldsymbol{x},\Theta\right)=j\right)\rightarrow P_{\Theta}\left(\hat{f}\left(\boldsymbol{x},\Theta\right)=j\right)$$

- Define a *margin function*

$$\begin{aligned} mr\left(\boldsymbol{X},Y\right) &= P_{\Theta}\left(\hat{f}\left(\boldsymbol{x},\Theta\right)=Y\right)-\max_{j\neq Y}P_{\Theta}\left(\hat{f}\left(\boldsymbol{x},\Theta\right)=j\right) \\ &= P_{\Theta}\left(\hat{f}\left(\boldsymbol{x},\Theta\right)=Y\right)-P_{\Theta}\left(\hat{f}\left(\boldsymbol{x},\Theta\right)=\hat{j}\left(\boldsymbol{X},Y\right)\right) \end{aligned}$$

where $\hat{j}\left(\boldsymbol{X},Y\right)=\arg\max_{j\neq Y}P_{\Theta}\left(\hat{f}\left(\boldsymbol{x},\Theta\right)=j\right)$

How random forests decorrelates(classification, cont'd)

- Furthermore, define a *raw margin function* by

$$rmg\left(\boldsymbol{X}, Y, \Theta\right) = \mathbb{I}\left(\hat{f}\left(\boldsymbol{x}, \Theta\right) = Y\right) - \mathbb{I}\left(\hat{f}\left(\boldsymbol{x}, \Theta\right) = \hat{j}\left(\boldsymbol{X}, Y\right)\right)$$

so that we can rewrite a margin function as

$$mg\left(\boldsymbol{X}, Y\right) = E_{\Theta} rmg\left(\boldsymbol{X}, Y, \Theta\right)$$

How random forests decorrelates(classification, cont'd)

- Lastly, define the generalization error by

$$PE^* = P_{\boldsymbol{X},Y}\left(mr\left(\boldsymbol{X},Y\right) < 0\right)$$

- By the Chebyshev inequality, we can easily show that

$$PE^* \leq \frac{Var\left(mr\left(\boldsymbol{X},Y\right)\right)}{s^2}$$

where $s = E_{\boldsymbol{X},Y} mr\left(\boldsymbol{X},Y\right)$, strength of a classifier $\hat{f}\left(\boldsymbol{X},\Theta\right)$

How random forests decorrelates(classification, cont'd)

Theorem

Assume that $s > 0$. *Then*

$$PE^* \leq \bar{\rho}\frac{1-s^2}{s^2}$$

where $\bar{\rho} = E_\Theta E_{\Theta'}\left[\rho\left(\Theta,\Theta'\right) sd\left(\Theta\right) sd\left(\Theta'\right)\right]/\left(E_\Theta sd\left(\Theta\right)\right)^2$, *the average correlation between* $rmg\left(\boldsymbol{X},Y,\Theta\right)$ *and* $rmg\left(\boldsymbol{X},Y,\Theta'\right)$.

How random forests decorrelates(classification, cont'd)

Proof.

$$
\begin{aligned}
Var\left(mr\left(\boldsymbol{X}, Y\right)\right) &= E\left[E_{\Theta} rmg\left(\boldsymbol{X}, Y, \Theta\right)\right]^2 - \left[E E_{\Theta} rmg\left(\boldsymbol{X}, Y, \Theta\right)\right]^2 \\
&= E_{\Theta, \Theta'} Cov\left(rmg\left(\boldsymbol{X}, Y, \Theta\right), rmg\left(\boldsymbol{X}, Y, \Theta'\right)\right)
\end{aligned}
$$

which can be implied from the fact that

$$
\left[E_{\Theta} f\left(\Theta\right)\right]^2 = E_{\Theta, \Theta'} f\left(\Theta\right) f\left(\Theta'\right)
$$

for independent Θ, Θ' and any f.

How random forests decorrelates(classification, cont'd)

Proof (cont'd).

For any random variables U and V, $\rho(U,V) = \frac{Cov(U,V)}{\sqrt{Var(U)}\sqrt{Var(V)}}$. Thus

$$Var\left(mr\left(\boldsymbol{X},Y\right)\right) = E_{\Theta,\Theta'}\rho\left(\Theta,\Theta'\right) sd\left(\Theta\right) sd\left(\Theta'\right)$$

where $\rho(\Theta,\Theta')$ is the correlation btw $rmg(\boldsymbol{X},Y,\Theta)$ and $rmg(\boldsymbol{X},Y,\Theta')$ holding Θ and Θ' fixed and $sd(\Theta)$ is the standard dev. of $rmg(\boldsymbol{X},Y,\Theta)$ for fixed Θ.

How random forests decorrelates(classification, cont'd)

Proof (cont'd).

By the definition of $\bar{\rho}$, we have

$$Var\left(mr\left(\boldsymbol{X}, Y\right)\right) = \bar{\rho}\{E_{\Theta} sd\left(\Theta\right)\}^2 \leq \bar{\rho} E_{\Theta} sd\left(\Theta\right)^2$$

by the Jensen's inequality. Now consider

$$E_{\Theta} sd\left(\Theta\right)^2 = E_{\Theta}\left[E_{\boldsymbol{X},Y} rmg\left(\boldsymbol{X}, Y, \Theta\right)^2 - \left(E_{\boldsymbol{X},Y} rmg\left(\boldsymbol{X}, Y, \Theta\right)\right)^2\right]$$

How random forests decorrelates(classification, cont'd)

Proof.

Since

$$s^2 = [E_{\boldsymbol{X},Y} mr(\boldsymbol{X}, Y)]^2 = [E_{\boldsymbol{X},Y} E_\Theta rmg(\boldsymbol{X}, Y)]^2$$
$$= [E_\Theta E_{\boldsymbol{X},Y} rmg(\boldsymbol{X}, Y)]^2 \leq E_\Theta [E_{\boldsymbol{X},Y} rmg(\boldsymbol{X}, Y)]^2$$

and

$$E_\Theta E_{\boldsymbol{X},Y} rmg(\boldsymbol{X}, Y, \Theta)^2 \leq 1.$$

Therefore $E_\Theta sd(\Theta)^2 \leq 1 - s^2$ and this completes the proof. □

How random forests decorrelate(regression, cont'd)

- Also by [Breiman, 2001], we have that almost surely for Θ

$$\hat{f}_{RF}(\boldsymbol{x}) = \frac{1}{B}\sum_{b=1}^{B}\hat{f}(\boldsymbol{x},\Theta_b) \to E_{\Theta}\hat{f}(\boldsymbol{x},\Theta)$$

- Define generalization error of forests and each tree by

$$PE^*(forest) = E_{\boldsymbol{X},Y}\left(Y - E_{\Theta}\hat{f}(\boldsymbol{x},\Theta)\right)^2$$

and

$$PE^*(tree) = E_{\Theta}E_{\boldsymbol{X},Y}\left(Y - \hat{f}(\boldsymbol{x},\Theta)\right)^2$$

How random forests decorrelate(regression, cont'd)

Theorem

Assume for all Θ *that* $E_{\boldsymbol{X},Y}\left[Y - \hat{f}\left(\boldsymbol{x},\Theta\right)\right]$. *Then*

$$PE^{*}\left(forest\right) \leq \bar{\rho} PE^{*}\left(tree\right)$$

where $\bar{\rho} = E_{\Theta}E_{\Theta'}\left[\rho\left(\Theta,\Theta'\right) sd\left(\Theta\right) sd\left(\Theta'\right)\right]/\left(E_{\Theta} sd\left(\Theta\right)\right)^{2}$, *the average correlation between* $Y - \hat{f}\left(\boldsymbol{X},\Theta\right)$ *and* $Y - \hat{f}\left(\boldsymbol{X},\Theta'\right)$.

How random forests decorrelate(regression, cont'd)

Proof.

$$
\begin{aligned}
PE^{*}\left(forest\right) &= E_{\boldsymbol{X},Y}\left(Y - E_{\Theta}\hat{f}\left(\boldsymbol{x},\Theta\right)\right)^{2} \\
&= E_{\boldsymbol{X},Y}\left[E_{\Theta}\left(Y - \hat{f}\left(\boldsymbol{x},\Theta\right)\right)\right]^{2} \\
&= E_{\boldsymbol{X},Y}\left[E_{\Theta}\left(Y - \hat{f}\left(\boldsymbol{x},\Theta\right)\right)\cdot E_{\Theta'}\left(Y - \hat{f}\left(\boldsymbol{x},\Theta'\right)\right)\right] \\
&= E_{\Theta,\Theta'}\left[E_{\boldsymbol{X},Y}\left(Y - \hat{f}\left(\boldsymbol{x},\Theta\right)\right)\cdot\left(Y - \hat{f}\left(\boldsymbol{x},\Theta'\right)\right)\right] \\
&= E_{\Theta,\Theta'}Cov\left(Y - \hat{f}\left(\boldsymbol{x},\Theta\right), Y - \hat{f}\left(\boldsymbol{x},\Theta'\right)\right) \\
&= E_{\Theta,\Theta'}\left[\rho\left(\Theta,\Theta'\right)sd\left(\Theta\right)sd\left(\Theta'\right)\right]
\end{aligned}
$$

How random forests decorrelate(regression, cont'd)

Proof (cont'd).

By the definition of average correlation $\bar{\rho}$, we have

$$\begin{aligned} PE^*(forest) &= \bar{\rho} \cdot (E_\Theta sd(\Theta))^2 \\ &\leq \bar{\rho} \cdot E_\Theta sd(\Theta)^2 \\ &= \bar{\rho} \cdot E_\Theta E_{\boldsymbol{X},Y} \left(Y - \hat{f}(\boldsymbol{x}, \Theta)\right)^2 \\ &= \bar{\rho} \cdot PE^*(tree) \end{aligned}$$

which completes the proof. □

Out-of-bag(OOB) estimates

- By nature of the bootstrap, some of observations may not be included in b-th bootstrap sample, i.e.

$$P\left(i \notin \boldsymbol{Z}^b\right) = \left(1 - \frac{1}{N}\right)^N \approx 0.367$$

- Denote the set of such samples $OOB(b)$, which we can use as a *test set*
- Thus, in principle, random forests do not require cross-validation!

Out-of-bag(OOB) estimates(cont'd)

- Define

$$Q\left(\boldsymbol{X}, j\right) = \frac{\sum_{b=1}^{B} \mathbb{I}\left(\hat{f}\left(\boldsymbol{X}, \Theta_b\right) = j : (\boldsymbol{X}, Y) \in OOB(b)\right)}{\sum_{b=1}^{B} \mathbb{I}\left((\boldsymbol{X}, Y) \in OOB(b)\right)}$$

  the out-of-bag proportion of votes cast at $\boldsymbol{X}$ for class j, which is an estimator of $P_{\Theta}\left(\hat{f}\left(\boldsymbol{X}, \Theta\right) = j\right)$

- Then an estimator of $mr\left(\boldsymbol{X}, Y\right)$ is $Q\left(\boldsymbol{X}, Y\right) - \max_{j \neq Y} Q\left(\boldsymbol{X}, j\right)$

Out-of-bag(OOB) estimates(cont'd)

- Now, an estimate of generalization error can be written as

$$\frac{1}{N}\sum_{i=1}^{N}\mathbb{I}\left(Q\left(\boldsymbol{X}^{(i)},Y^{(i)}\right)-\max_{j\neq Y^{(i)}}Q\left(\boldsymbol{X}^{(i)},j\right)<0\right)$$

- OOB estimates of strength and correlation can be calculated in similar ways, where details can be found in [Breiman, 2001]

Variable importance

- One of the advantages of decision tree model is easy interpretability
- Such an advantage disappears in random forests due to bootstrap and random selection of features
- To explain which predictors contribute to forests more than the others, we introduce the *variable importance*

Variable importance(cont'd)

- [Breiman, 2001] suggests a *permutation importance*
- The idea is that if the variable $\boldsymbol{X}_j$ is important, the error using randomly shuffled data (permutation) should be much worse than that using the correct one
- The permutation importance by OOB set measures how much the accuracy deteriorates due to permutation in OOB sets

Variable importance(cont'd)

- Let π_b^j be a permutation on the j-th variable in b-th OOB set. Define

$$VI\left(j,b\right)=\frac{\sum_{i\in OOB(b)}\left[\mathbb{I}\left(Y^{(i)}=\hat{f}^b\left(\boldsymbol{X}^{(i)}\right)\right)-\mathbb{I}\left(Y^{(i)}=\hat{f}^b\left(\boldsymbol{X}_{\pi_b^j}^{(i)}\right)\right)\right]}{|OOB(b)|}$$

  where $\boldsymbol{X}_{\pi_b^j}^{(i)}$ is the value of permuted j-th variable for observation i.
- Now the permutation variable importance is written as

$$VI\left(j\right)=\frac{1}{B}\sum_{b=1}^{B}VI\left(j,b\right)$$

Asymptotic properties of random forests

- Simplified proof given by [Breiman, 2004]
- Further researches from [Lin et al., 2006], [Biau, 2012], [Wager and Athey, 2018] or etc

Further applications of random forests

- Survival data: [Ishwaran et al., 2008]
- Causal inference: [Wager and Athey, 2018]
- Generalized random forests: [Athey et al., 2019]

Homeworks

1. By using Chebyshev's inequality, prove that

$$PE^* \leq \frac{Var\left(mr\left(\boldsymbol{X}, Y\right)\right)}{s^2}.$$

2. Show that

$$P\left(i \notin \boldsymbol{Z}^b\right) = \left(1 - \frac{1}{N}\right)^N \approx 0.367$$

holds. Why does the approximation hold?

3. Are there any methods for quantifying variable importances other than permutation? If then, suggest briefly how you can!

References I

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

Biau, G. (2012).
Analysis of a random forests model.
Journal of Machine Learning Research, 13(Apr):1063–1095.

Bishop, C. M. (2006).
Pattern recognition and machine learning.
springer.

References II

Breiman, L. (1996).

Bagging predictors.

Machine learning, 24(2):123–140.

Breiman, L. (2001).

Random forests.

Machine learning, 45(1):5–32.

Breiman, L. (2004).

Consistency for a simple model of random forests.

References III

Breiman, L., Friedman, J., Stone, C. J., and Olshen, R. A. (1984).
Classification and regression trees.
CRC press.

Friedman, J. H. (1991).
Multivariate adaptive regression splines.
The annals of statistics, pages 1–67.

Hastie, T. and Tibshirani, R. (1987).
Generalized additive models: some applications.
Journal of the American Statistical Association, 82(398):371–386.

References IV

Hastie, T., Tibshirani, R., and Friedman, J. (2009).
The elements of statistical learning: data mining, inference, and prediction.
Springer Science & Business Media.

Ishwaran, H., Kogalur, U. B., Blackstone, E. H., Lauer, M. S., et al. (2008).
Random survival forests.
The annals of applied statistics, 2(3):841–860.

References V

Jordan, M. I. and Jacobs, R. A. (1994).
Hierarchical mixtures of experts and the em algorithm.
Neural computation, 6(2):181–214.

Lin, Y., Zhang, H. H., et al. (2006).
Component selection and smoothing in multivariate nonparametric regression.
The Annals of Statistics, 34(5):2272–2297.

Murphy, K. P. (2012).
Machine learning: a probabilistic perspective.
MIT press.

References VI

Ravikumar, P., Lafferty, J., Liu, H., and Wasserman, L. (2009).
Sparse additive models.
Journal of the Royal Statistical Society: Series B (Statistical Methodology), 71(5):1009–1030.

Wager, S. and Athey, S. (2018).
Estimation and inference of heterogeneous treatment effects using random forests.
Journal of the American Statistical Association, 113(523):1228–1242.