Contents:

Lecture 17

1 Random forest and its statistical inference

Random forest is considered to be an important tool for nonlinear regression and classification. However, its statistical inference has long been an issue until recently. This section introduces the construction of random forest and its application in causal inference study. Especially, we discuss how to estimate the variance of a random forest.

§1 回归树

1.1 Growing a regression tree

1. 模型设定

Suppose the i.i.d. observations are $(X_i, Y_i) \in \mathbf{R}^d \times \mathbf{R}$ that satisfy

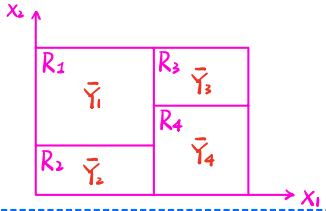
$$Y_i = f(X_i) + \epsilon_i \text{ where } \epsilon_i \perp X_i, \text{ in such case } f(X_i) = \mathbf{E}Y_i|X_i,$$

2. 模型预测

we want to estimate $f(\cdot)$. The regression tree uses a set of "rectangles" to make prediction. Roughly speaking, the estimation follows two steps:

1. We divide the feature space into J distinct and non-overlapping regions $R_1, \cdots, R_J$.
2. For every observation that falls into the region R_j , we make the same prediction, which is simply the mean of the response values for the training observations in R_j, i.e.,

$$\widehat{f}(x) = \frac{1}{N_j} \sum_{X_i \in R_j} Y_i \text{ if } x \in R_j.$$



3. 模型拟合

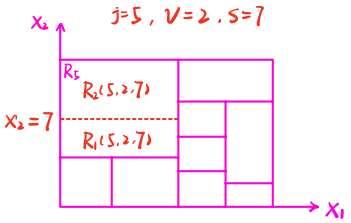
The growing of a tree relies on the recursive binary splitting. We want to minimize the residual sum of square

$$\ell = \sum_{j=1}^{J} \sum_{i \in R_j} (y_i - \widehat{f}(X_i))^2. \quad \text{(RSS)}$$

In general, this is too hard if we consider all possible R_j. However, we can do the following greedy algorithm to decrease the loss:

① Suppose we have rectangles $R_1, R_2, \cdots, R_{J_0}$ with loss

$$\ell = \sum_{j=1}^{J_0} \sum_{i \in R_j} (y_i - \widehat{f}(X_i))^2.$$



② For a tuple $(j, v, s) \in \{1, 2, \cdots, J_0\} \times \{1, 2, \cdots, d\} \times \mathbf{R}$, we consider the half-planes

$$R_1(j, v, s) = \{X_i : X_i \in \mathbf{R}_j, X_{iv} < s\} \text{ and } R_2(j, v, s) = \{X_i : X_i \in \mathbf{R}_j, X_{iv} \geq s\}.$$

③ After that, we calculate the new loss （将 R_j 根据维度 v 和阈值 s 分割后，loss 的变化量）

$$\ell_1 = \Big(\sum_{i \in R_1(j,v,s)} (y_i - \widehat{f}(X_i))^2 + \sum_{i \in R_2(j,v,s)} (y_i - \widehat{f}(X_i))^2 \Big) - \sum_{i \in R_j} (y_i - \widehat{f}(X_i))^2,$$

and find the minimizer of ℓ_1. （调整 (j, v, s) 的选取，使 loss 的下降量最大）

We separate the rectangle if $\ell_1 < -\alpha < 0$. （若 loss 的最大下降量小于 α，停止迭代）

Remark If the outcome Y_i is a binary data, then the regression tree becomes decision tree. The difference here is,

1. In each leaf, we assign "yes" or "no" by checking the probability / ratio $\frac{1}{N_j} \sum_{X_i \in R_j} Y_i$.
2. For classification tree, the RSS is not a good criterion, instead, we grow the tree via the Gini index or entropy

$$G = \sum_{j=1}^{J} \widehat{p}_j (1 - \widehat{p}_j) \text{ or } D = -\sum_{j=1}^{J} \widehat{p}_j \log(\widehat{p}_j).$$

In other words, we want the Y_i in each rectangle be in the same class.

4. causal tree

In causal inference (e.g., Wager and Athey, 2017), suppose we observe a set of data $(X_i, Y_i, W_i) \in \mathbf{R}^p \times \mathbf{R} \times \mathbf{R}$, and for each test point x, we want to estimate the conditional average treatment effect

$$\tau(x) = \mathbf{E} Y_i(1) | X_i = x - \mathbf{E} Y_i(0) | X_i = x.$$

To achieve the goal, we consider what happens in ℓ. Notice that

$$\ell = \sum_{j=1}^{J} \sum_{i \in R_j} (y_i - \widehat{f}(X_i))^2 = \sum_{j=1}^{J} \sum_{i \in R_j} (y_i^2 + \widehat{f}(X_i)^2 - 2 y_i \widehat{f}(X_i)) = \sum_{j=1}^{J} \sum_{i \in R_j} y_i^2 - \sum_{j=1}^{J} \sum_{i \in R_j} \widehat{f}(X_i)^2.$$

Therefore, minimizing ℓ is equivalent to maximizing the sum of square

$$\sum_{j=1}^{J} \sum_{i \in R_j} \widehat{f}(X_i)^2.$$ （一个重要的 observation）

This observation is the key to the validity of the causal tree. We consider the similar criterion

$$\ell = \sum_{j=1}^{J} \sum_{i \in R_j} (Y_i(1) - Y_i(0) - \widehat{\tau}(X_i))^2,$$

where

$$\widehat{\tau}(x) = \frac{1}{\sum_{W_i=1, X_i \in R(x)} 1} \sum_{W_i=1, X_i \in R(x)} Y_i - \frac{1}{\sum_{W_i=0, X_i \in R(x)} 1} \sum_{W_i=0, X_i \in R(x)} Y_i.$$

Unfortunately, this idea is infeasible because we cannot simultaneously observe $Y_i(0)$ and $Y_i(1)$. However, if we simulate the above procedure, this can be transformed to maximizing the sum of square

$$\sum_{j=1}^{J} \sum_{i \in R_j} \widehat{\tau}(X_i)^2.$$

Remark There are some issues for this criterion, especially if the leaf contains too few samples. Therefore, in practice, we may consider restricting by forcing each node containing at least k nodes.

1.2 Sampling splitting and kernel representation of random forest

One of the issue in analyzing the regression tree is the dependency between the tree-growing process and the prediction. Roughly speaking, after growing a tree, the predictor for a feature x is actually

$$\widehat{f}(x) = \sum_{i=1}^{n} K(x, X_i) Y_i,$$

where $K(x, X_i) = \frac{1}{\sum_{X_i \in R_j} 1}$ if x in R_j and 0 otherwise. Therefore, if the tree-growing process is independent of Y_i, then the analysis can be taken conditional on X, which motivates the following algorithm.

Algorithm 1: Double sample trees

1. Divide the samples into two disjoint sets I and J.
2. Grow a tree via recursive partitioning. The splits are chosen using any data from I samples, and X_i from J samples, but do not use Y_i in J samples when calculating the residual sum of squares.
3. Estimate the leaf-wise nodes using data in J samples.

2 Bootstrapping aggregation

A decision tree suffers from large variance. To settle the issue, statisticians may consider leveraging the bootstrap aggregation, whose algorithm is as follows:

Algorithm 2: Bagging

The input of data $Z_i = (X_i, Y_i) \in \mathbf{R}^d \times \mathbf{R}, i = 1, 2, \cdots, n$, predictor $x \in \mathbf{R}^d$.

1. Draw with replacement from $Z_1, \cdots, Z_n$ and form data $Z_1^*, \cdots, Z_n^*$.
2. Calculate the estimator $Y_b^* = f(x, Z_1^*, \cdots, Z_n^*)$.
3. Repeat experiment for B times and derive the estimator

$$\widehat{Y} = \frac{1}{B} \sum_{b=1}^{B} Y_b^*.$$

The problem we are interested in is, how to analyze $\widehat{Y}$.

First let $B \to \infty$, in such case, from law of large number, we have

$$\frac{1}{B}\sum_{b=1}^{B} Y_b^* \to \mathbf{E}f(x, Z_1^*, \cdots, Z_n^*)|Z_1, \cdots, Z_n. = \frac{1}{n^n}\sum_{i_1=1}^{n}\sum_{i_2=1}^{n}\cdots\sum_{i_n=1}^{n} f(x, Z_{i1}, \cdots, Z_{im}).$$

In other words, for sufficiently large B, the bagging approximately estimates the sample average across all possible combinations.

2. Infinitesimal jackknife (估计 bagging estimator 的 variance)

The next problem involves estimating the variance of the bagging estimator. Efron(2014) established an estimator for the variance of $\mathbf{E}f(x, Z_1^*, \cdots, Z_n^*)|Z_1, \cdots, Z_n$.

Theorem 2.1 (Infinitesimal jackknife). *For bth bootstrapped sample $Z_{b1}^*, Z_{b2}^*, \cdots, Z_{bn}^*$, define the number $S_{bj}^* = \#\{Z_{bk}^* = Z_j\}$ for $b = 1, \cdots, B$ and $j = 1, \cdots, n$. An estimator for the variance of $\widehat{Y}$ is*

每个 sample 在 resampling 时被(重复)抽出的次数

$$\widehat{V} = \sum_{j=1}^{n} \widehat{cov}_j^2, \text{ where } \widehat{cov}_j = \frac{1}{B}\sum_{b=1}^{B}(S_{bj}^* - \overline{S}_{\cdot j}^*) \times (f(x, Z_{b1}^*, \cdots, Z_{bn}^*) - \widehat{Y}_\cdot) \quad (1)$$

Remark If we assess the distribution of S_{bj}^*, we have

$$Prob(S_{bj}^* = s|Z_1, \cdots, Z_n) = \binom{n}{s} \times (1/n)^s \times (1 - 1/n)^{n-s}$$

and

$$Prob(S_{b1}^* = s_1, \cdots, S_{bn}^* = s_n|Z_1, \cdots, Z_n) = \frac{n!}{s_1! \cdots s_n!}(\frac{1}{n})^{s_1+\cdots+s_n}.$$

In other words, S_{bj}^* in such case satisfies a binomial distribution. $S_{bj}^* \sim Binomial(n, \frac{1}{n})$

$\Rightarrow (S_{b1}^*, \cdots, S_{bn}^*) \sim multinomial(n, \frac{1}{n}, \cdots, \frac{1}{n})$

Again, suppose $B \to \infty$, then

$$\widehat{cov}_j \to Cov(S_{bj}^*, f(x, Z_{b1}^*, \cdots, Z_{bn}^*)|Z_1, \cdots, Z_n).$$

With such technique, suppose we want to quantify the precision of the estimator, we may consider building the confidence interval by

$$[\widehat{Y} - q_{\alpha/2} \times \sqrt{\widehat{V}},\ \widehat{Y} + q_{\alpha/2} \times \sqrt{\widehat{V}}].$$

Notably, this case is for general bagging estimator. If we choose $f()$ to be a decision / regression tree, then the algorithm can be used to estimate the variance of the bagged trees.

Remark [Bias-corrected variance estimation] In practice, for B is finite, we can consider the bias-corrected version

$$\widehat{V}_{debias} = \widehat{V} - \frac{n}{B^2}\sum_{b=1}^{B}(f(x, Z_{b1}^*, \cdots, Z_{bn}^*) - \widehat{Y})^2$$

to decrease the bias. If we do not do that, then a $B \approx n^{1.5}$ is needed to eliminate the bias.

Remark Suppose a random vector $Q \sim multinominal(n, p_1, \cdots, p_n)$, define function

$$S(p_1, \cdots, p_n) = \mathbf{E} f(Q), \text{ with } (p_1^\dagger, \cdots, p_n^\dagger) = (1/n, 1/n, \cdots, 1/n).$$

Define the directional derivative

$$A_j = \lim_{\epsilon \to 0} \frac{S(p_1^\dagger - \frac{\epsilon}{n}, p_2^\dagger - \frac{\epsilon}{n}, \cdots, p_j^\dagger + \epsilon - \frac{\epsilon}{n}, \cdots, p_n^\dagger - \frac{\epsilon}{n}) - S(p_1^\dagger, \cdots, S_n^\dagger)}{\epsilon},$$

then the variance is given by

$$\frac{1}{n^2} \sum_{j=1}^{n} A_j^2.$$

To illustrate what goes on. Suppose the estimator we are interested is

$$\widehat{T} = X_1.$$

For the variance and covariances of multinomial distribution is

$$Var(X_i) = np_i(1 - p_i) \text{ and } Cov(X_i, X_j) = -np_i p_j,$$

we have

$$Var(\widehat{T}) = 1 - \frac{1}{n}.$$

On the other hand, $\mathbf{E}\widehat{T} = np_1$. Therefore,

$$A_1 = \lim_{\epsilon \to 0} \frac{n(p_1^\dagger + \epsilon - \frac{\epsilon}{n}) - np_1^\dagger}{\epsilon} = n - 1, \; A_j = \lim_{\epsilon \to 0} \frac{-\epsilon}{\epsilon} = -1.$$

Therefore,

$$\frac{1}{n^2}(n-1)^2 + \frac{1}{n^2} \sum_{j=2}^{n} 1 = 1 - 1/n.$$

3. 使用 subsample The idea of variance estimation in Wager and Athey (2017) for causal tree is similar. Suppose the subsample size is s and we draw without replacement, then the variance of the conditional average treatment effect can be estimated by

$$\widehat{V} = \frac{n-1}{n} \left(\frac{n}{n-s} \right)^2 \sum_{i=1}^{n} \widehat{cov}(\widehat{\tau}_b(x)^*, N_{ib}^*)^2, \text{ where } \widehat{cov}(\widehat{\tau}_b(x)^*, N_{ib}^*) = \frac{1}{B} \sum_{b=1}^{B} (N_{ib}^* - \overline{N}_{i\cdot}^*) \times (\widehat{\tau}_b(x)^* - \widehat{Y}).$$

Here $N_{ib}^* \in \{0, 1\}$ indicate whether or not the i-th training example was used for the b-th tree. Further-

more, we have

$$(\widehat{\tau}(x) - \tau(x))/\sqrt{\widehat{V}} \to_d N(0,1),$$

so the confidence interval for the conditional average treatment effect can be

$$[\widehat{\tau}(x) - q_{\alpha/2} \times \sqrt{\widehat{V}},\ \widehat{\tau}(x) + q_{\alpha/2} \times \sqrt{\widehat{V}}].$$