The gradient boosting decision tree approach – Mathematical illustration

Model estimation

The GBDT approach was originally proposed by Friedman (2001, 2002) and developed into R package by Ridgeway (2020) and Greenwell et al. (2020). Following the notations by Ridgeway (2020), $(\boldsymbol{X}_i, y_i)$ is the ith observation in the sample space S. $\boldsymbol{X}_i$ indicates the vector of independent variables and y_i is the dependent variable. $\boldsymbol{X}$ is the matrix of independent variables of all observations. S has N observations. Ψ is the loss function. In this study, Ψ is squared error loss.

Three important parameters need to be set before the initial step: tree depth K, learning rate λ, and the number of iterations T. $\hat{f}(\boldsymbol{X})$ is initialized as a constant, $\hat{f}(\boldsymbol{X}) = argmin_\rho \sum_{i=1}^{N} \Psi(y_i, \rho)$. The steps below will be iterated for T times. For the tth time,

Step 1. Calculate the negative gradient z_i of the ith observation with Equation A1 below.

$$z_i = -\frac{\partial}{\partial f(\boldsymbol{X}_i)} \Psi(y_i, f(\boldsymbol{x}_i))|_{f(\boldsymbol{X}_i)=\hat{f}(\boldsymbol{X}_i)} \quad \text{(A1)}$$

Step 2. Friedman (2002) found that fitting decision trees with randomly selected subsample from the original sample provides better performance. In addition, he suggested this proportion to be 0.5. Therefore, $0.5 \times N$ observations are randomly selected from the original sample space.

Step 3. Fit a decision tree with K terminal nodes using the randomly selected observations from Step 2.

Step 4. Calculate the optimal prediction ρ_k for the kth terminal node of the fitted decision tree in Step 3 using Equation A2.

$$\rho_k = argmin_\rho \sum_{\boldsymbol{x}_i \in S_k} \Psi(y_i, \hat{f}(\boldsymbol{X}_i) + \rho) \quad \text{(A2)}$$

where S_k is the set of observations categorized in the kth terminal node of the fitted decision tree.

Step 5. Update $\hat{f}(\boldsymbol{X})$ with Equation A3.

$$\hat{f}(\boldsymbol{X}) \leftarrow \hat{f}(\boldsymbol{X}) + \lambda \rho_{k(\boldsymbol{X})} \quad \text{(A3)}$$

where $k(\boldsymbol{X})$ indicates the index of the terminal node of the fitted decision tree where the observations are located.

Relative importance

Equation A4 below is to calculate the total variance reduction $\hat{J}_j^2$ by the independent variable x_j (Ridgeway 2020).

$$\hat{J}_j^2 = \sum_{splits\ on\ x_j} I_t^2 \tag{A4}$$

where I_t^2 is the variance reduction by splitting with x_j. Relative importance is calculated with Equation A5 below.

$$RI_j = \frac{\hat{J}_j^2}{\sum_j^M \hat{J}_j^2} \times 100\% \tag{A5}$$

where M is number of independent variables.

ALE estimation

Following the notations by Molnar (2020), x_j is the jth independent variable and is split into $k_j(\boldsymbol{x})$ intervals. We can use Equation A6 to estimate uncentered ALE for x_j.

$$\hat{\tilde{f}}_{j,ALE}(\boldsymbol{X}) = \sum_{k=1}^{k_j(\boldsymbol{X})} \frac{1}{n_j(k)} \sum_{i:x_j^{(i)} \in N_j(k)} \left[f\left(z_{k,j}, \boldsymbol{x}_{\setminus j}^{(i)}\right) - f\left(z_{k-1,j}, \boldsymbol{x}_{\setminus j}^{(i)}\right)\right] \tag{A6}$$

where $N_j(k)$ is the space of the kth interval. $n_j(k)$ is the number of observations in the kth interval. $z_{k,j}$ is the maximum value of x_j in $N_j(k)$. $z_{k-1,j}$ is the minimum value of x_j in $N_j(k)$. $\boldsymbol{x}_{\setminus j}^{(i)}$ is the ith observation without x_j in $N_j(k)$. $f(.,.)$ is the estimated GBDT model.

We, then, can use Equation A7 to calculate centered ALE for x_j. Note that the ALE calculated in this study is centered ALE.

$$\hat{f}_{j,ALE}(\boldsymbol{x}) = \hat{\tilde{f}}_{j,ALE}(\boldsymbol{x}) - \frac{1}{n}\sum_{i=1}^{n} \hat{\tilde{f}}_{j,ALE}\left(\boldsymbol{x}_j^{(i)}\right) \tag{A7}$$

where n is the number of observations in the sample.

Reference

Friedman, Jerome H. 2001. "Greedy function approximation: A gradient boosting machine." *The Annals of Statistics* 29 (5):1189-1232. doi: 10.1214/aos/1013203451.

Friedman, Jerome H. 2002. "Stochastic gradient boosting." *Computational Statistics & Data Analysis* 38 (4):367-378. doi: 10.1016/s0167-9473(01)00065-2.

Greenwell, Brandon, Bradley Boehmke, and Jay Cunningham. 2020. "gbm: Generalized Boosted Regression Models." https://cran.r-project.org/package=gbm.

Molnar, Christoph. 2020. "Interpretable Machine Learning - A Guide for Making Black Box Models Explainable." In: lulu.com. https://christophm.github.io/interpretable-ml-book/.

Ridgeway, Greg. 2020. Generalized Boosted Models: A guide to the gbm package. 2007.