Gradient Boosting Machine

Epoch IIT Hyderabad

Ankit Saha
AI21BTECH11004

26 Jul 2022

1. Introduction

Gradient boosting is a supervised machine learning algorithm that is used for both classification and regression. It uses a sequence of fixed sized decision trees where each tree is generated using the errors of the previous tree. The contribution of each tree is scaled down by a fixed amount. Gradient boosting machine (GBM) usually outperforms random forest.

2. Mathematical Formulation

Suppose we are given m labelled samples $(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_m, y_m)$ where $x_i \in \mathbb{R}^n, y_i \in \mathcal{D}\ \forall i \in [m]$. $\mathcal{D}$ can either be an interval (for regression) or discrete values (for classification). The goal is to learn a function $F : \mathbb{R}^n \to \mathcal{D}$ that can predict the output $\hat{y} = F(\mathbf{x})$ for a given input $\mathbf{x}$

Choose a differentiable loss function $L(y, F(\mathbf{x}))$ where $L : \mathcal{D} \times \mathcal{D} \to \mathbb{R}$
The algortihm is as follows:

i) Initialize

$$F^{(0)}(\mathbf{x}) = \arg\min_{\gamma} \sum_{i=1}^{m} L(y_i, \gamma)$$

ii) Compute the pseudo-residuals (difference between actual and predicted values)

$$r_i^{(t)} = -\frac{\partial}{\partial F(\mathbf{x}_i)} \left(L(y_i, F(\mathbf{x}_i))\right)\Big|_{F(\mathbf{x})=F^{(t-1)}(\mathbf{x})} \quad \forall i \in [m]$$

iii) Fit a decision tree to the training set $\left\{\left(\mathbf{x}_i, r_i^{(t)}\right)\right\}_{i=1}^{m}$ to get the leaves $R_j^{(t)}$ where j is the index of the leaf

iv) Compute the output value of each leaf

$$\gamma_j^{(t)} = \arg\min_{\gamma} \sum_{\mathbf{x} \in R_j^{(t)}} L(y_i, F^{(t-1)}(\mathbf{x}) + \gamma)$$

v) Update the model

$$F^{(t)}(\mathbf{x}) = F^{(t-1)}(\mathbf{x}) + \nu\gamma_k^{(t)} \text{ when } \mathbf{x} \in R_k^{(t)}$$

where $0 < \nu < 1$ is a fixed quantity known as the learning rate

3. Questions

i) What is the significance of the learning rate?
ii) Give an example of a loss function?
iii) Compute $F^{(0)}(\mathbf{x})$ for your loss function.
iv) Compute $\gamma^{(t)}(j)$ for your loss function.
v) How would you extend GBM to classification problems?

4. Answers

i) Having a low learning rate prevents overfitting of the data by reducing the variance of the model.
ii) A commonly used loss function is

$$L(y, F(\mathbf{x})) = \frac{1}{2}\left(y - F(\mathbf{x})\right)^2$$

The factor of half is present so that it cancels out the constant factor of 2 obtained upon differentiating the loss function.

iii)

$$F^{(0)}(\mathbf{x}) = \arg\min_{\gamma} \sum_{i=1}^{m} L(y_i, \gamma) \tag{4.1}$$

$$= \arg\min_{\gamma} \sum_{i=1}^{m} \frac{1}{2}(y_i - \gamma)^2 \tag{4.2}$$

On differentiating the summation with respect to γ,

$$\frac{\mathrm{d}}{\mathrm{d}\gamma} \sum_{i=1}^{m} (y_i - \gamma)^2 = -2 \sum_{i=1}^{m} (y_i - \gamma) = 0 \tag{4.3}$$

$$\implies m\hat{\gamma} = \sum_{i=1}^{m} y_i \tag{4.4}$$

$$\implies F^{(0)}(\mathbf{x}) = \hat{\gamma} = \frac{1}{m} \sum_{i=1}^{m} y_i \tag{4.5}$$

Thus, the intial guess is the average of the labels.

iv)

$$\gamma_j^{(t)} = \arg\min_{\gamma} \sum_{\mathbf{x} \in R_j^{(t)}} L(y_i, F^{(t-1)}(\mathbf{x}) + \gamma) \tag{4.6}$$

$$= \arg\min_{\gamma} \sum_{\mathbf{x} \in R_j^{(t)}} \left(y_i - F^{(t-1)}(\mathbf{x}) - \gamma\right)^2 \tag{4.7}$$

$$= \frac{1}{\left|R_j^{(t)}\right|} \sum_{\mathbf{x} \in R_j^{(t)}} \left(y_i - F^{(t-1)}(\mathbf{x})\right) \tag{4.8}$$

Thus, the output of each leaf is the average of the residuals at that leaf.

v) We can convert the categorical labels into real-valued labels by taking the logarithm of the odds of the labels and then using the sigmoid function to convert them into probabilities. Use these probabilities to calculate the residuals and fit the subsequent decision trees.