XGBoost
Epoch IIT Hyderabad

Ankit Saha
AI21BTECH11004

26 Jul 2022

1. Introduction

Extreme gradient boosting (XGBoost) is a supervised machine learning algorithm that is used for both classification and regression. It is very similar to gradient boosting but the difference is that XGBoost is analogous to Newton-Raphson in the same way GBM is analogous to gradient descent. XGBoost is a very popular algorithm because of its automatic feature selection and high accuracy.

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 gradients and hessians

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

$$h_i^{(t)} = \frac{\partial^2}{\partial F(\mathbf{x}_i)^2} \left(L(y_i, F(\mathbf{x}_i))\right)\Big|_{F(\mathbf{x})=F^{(t-1)}(\mathbf{x})} \quad \forall i \in [m]$$

Let their ratio be denoted by $r_i^{(t)}$

$$r_i^{(t)} = \frac{g_i^{(t)}}{h_i^{(t)}}\ \forall i \in [m]$$

iii) Fit an XGBoost decision tree to the training set $\left\{\left(\mathbf{x}_i, r_i^{(t)}\right)\right\}_{i=1}^{m}$

iv) Compute the incremental function

$$\phi^{(t)} = \arg\min_{\phi} \sum_{i=1}^{m} \frac{1}{2} h_i^{(t)} \left(r_i^{(t)} - \phi(\mathbf{x}_i)\right)^2 \tag{2.1}$$

v) Update the model

$$F^{(t)}(\mathbf{x}) = F^{(t-1)}(\mathbf{x}) + \nu\phi^{(t)}(\mathbf{x})$$

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

3. Questions

i) How does XGBoost perform feature selection?
ii) Give an example of a loss function.
iii) Compute $F^{(0)}(\mathbf{x})$ for your loss function.
iv) Compute the gradient of your loss function.
v) Compute the hessian of your loss function.

4. Answers

i) The model calculates a feature importance score for each of the features after training and the construction of the trees. This metric gives information about the contribution of each feature to the final model. Features with a score less than a particular threshold can be dropped.
ii) A commonly used loss function is

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

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)

$$g_i^{(t)} = -\frac{\partial}{\partial F(\mathbf{x}_i)} \left(L(y_i, F(\mathbf{x}_i))\right)\Big|_{F(\mathbf{x})=F^{(t-1)}(\mathbf{x})} \tag{4.6}$$

$$= -\frac{\partial}{\partial F(\mathbf{x}_i)} \left(\frac{1}{2}\,(y_i - F(\mathbf{x}_i))^2\right)\Big|_{F(\mathbf{x})=F^{(t-1)}(\mathbf{x})} \tag{4.7}$$

$$= (y_i - F(\mathbf{x}_i))|_{F(\mathbf{x})=F^{(t-1)}(\mathbf{x})} \tag{4.8}$$

$$= y_i - F^{(t-1)}(\mathbf{x}_i) \tag{4.9}$$

v)

$$h_i^{(t)} = \frac{\partial^2}{\partial F(\mathbf{x}_i)^2} \left(L(y_i, F(\mathbf{x}_i))\right)\Big|_{F(\mathbf{x})=F^{(t-1)}(\mathbf{x})} \tag{4.10}$$

$$= -\frac{\partial}{\partial F(\mathbf{x}_i)} \left(g_i^{(t)}\right)\Big|_{F(\mathbf{x})=F^{(t-1)}(\mathbf{x})} \tag{4.11}$$

$$= -\frac{\partial}{\partial F(\mathbf{x}_i)} \left(y_i - F(\mathbf{x}_i)\right)\Big|_{F(\mathbf{x})=F^{(t-1)}(\mathbf{x})} \tag{4.12}$$

$$= 1 \tag{4.13}$$