Gradient Boosting

Oliver Zhao

1 Description

Gradient Boosting generates and combines weak learners in a sequential manner, attempting to minimize the loss function with each iteration. For example, consider a flawed model F_m. Our goal is to add a new weak learner h such that the new model performs better, where

$$F_{m+1}(x) = F_m(x) + h(x). \tag{1.1}$$

If the addition of the new learner h theoretically results in a perfect model, that would mean

$$F_{m+1}(x) = F_m(x) + h(x) = y, \tag{1.2}$$

Or that

$$h(x) = y - F_m(x). \tag{1.3}$$

Consequently, Gradient Boosting aims to fit the new weak learner h to the residual $y - F_m(x)$. Notice that this residual shows which data points the existing model is unable to correctly fit. The core idea is that each sequentially improved combined model results in a more accurate performance. The residuals can be interpreted as negative gradients, so the model can be updated through gradient descent, hence the name Gradient Boosting.

2 Algorithm

Description: Consider a training set $\{(x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m)\}$, where $x_i \in \mathscr{X}$ describes the features and y_i describes the class. We notate the loss function as $L(y, F(x))$.

---

1. Initialize the model $F_0(x)$ with a constant value, where

$$F_0(x) = \arg\max_{\gamma} \sum_{i=1}^{n} L(y_i, \gamma). \tag{2.1}$$

2. For t = 1 to T:

   (a) Compute the pseudo-residuals, where

$$r_{im} = -\left[\frac{\delta L(y_i, F(x_i))}{\delta F(x_i)}\right]_{F(x)=F_{m-1}(x)} \quad \text{for } i = 1, \ldots, n \tag{2.2}$$

   (b) Fit the weak learner $h_m(x)$ to the pseudo-residual with the training set $\{(x_i, r_{im})\}^n + i = 1$.

   (c) Compute the multiplier γ_m

$$y_m = \arg\max_{\gamma} \sum_{i=1}^{n} L(y_i, F_{m-1}(x_i) + \gamma h_m(x_i)). \tag{2.3}$$

   (d) Update the model, where

$$F_m(x) = F_{m-1}(x) + \gamma_m h_m(x). \tag{2.4}$$

3. Output the final model $F_M(x)$.
