Gradient Boosting Machine

Aaryan
CO21BTECH11001

Gradient Boosting Machine (GBM) is a boosting algorithm which creates a sequence of (weaker) predictive models and uses them to build a (stronger) predictive model. Our model in this case is a Decision Tree, which can be used for both regression as well as classification.

Input: $\{x_i, y_i\}_{i=1}^{m}$, a loss function $L(y, F(x))$, max iterations = M

The algorithm can be divided into 3 steps.

1. Initialize the model with a constant value

$$F_0(x) = \underset{\gamma}{\operatorname{argmin}} \left(\sum_{i=1}^{m} L(y_i, \gamma) \right)$$

   which we can calculate by $\frac{dF_0(x)}{d\gamma} = 0$

2. Iterate k =1 to M:
   a. Compute pseudo residuals
      for i =1 **to** m:
      $r_{ik} = -\left[\frac{\partial L(y_i, F_{k-1}(x_i))}{\partial F_{k-1}(x_i)}\right]$
      end
   b. Fit a base learner (decision tree) $h_k(x)$ with input $\{(x_i, r_{ik})\}_{i=1}^{m}$
      Note: The maximum number of leaf nodes in $h_k(x)$ is generally between 8 to 32.
   c. Find the optimal learning rate for kth iteration.

$$\gamma_k = \underset{\gamma}{\operatorname{argmin}} \left(\sum_{i=1}^{m} L(y_i, F_{k-1}(x_i) + \gamma) \right)$$

   d. Update the model

$$F_k(x) = F_{k-1}(x) + \gamma_k h_k(x)$$

3. Output function: $F_M(x)$.

Note: For a regression problem, the loss function generally used is squared error loss, whereas for a classification problem, binary cross entropy loss is generally used.

Questions:

1. GBM is a/an:
   (a) Supervised learning algorithm
   (b) Unsupervised learning algorithm
   Ans. (a)

2. GBM can be used for:
   (a) Regression
   (b) Classification
   (c) Both of these
   (d) None of these
   Ans. (c)

3. What are pseudo residuals and how are they calculated?
   Ans. Pseudo residuals are intermediate error terms i.e., difference between the actual value and intermediate predicted value. They are calculated by taking the derivative of the loss function with previous hypothesis function ($F_{k-1}(x_i)$).

4. What learning rate is used in GBM?
   Ans. Learning rate can either be constant for all iterations/models, or it can be calculated by optimizing the loss function i.e., we can skip step 2.c if we want and just use some constant value (say 0.1) as learning rate.

5. What are disadvantages of using GBM?

Ans. It is prone to overfitting and computationally expensive.