Gradient Boosting

By: Behzad Asadi

Gradient Boosting

Gradient boosting is about constructing an ensemble learner $F_M(x) = \sum_{m=1}^{M} \alpha_m f_m(x)$ in a sequential manner from a weak learner $f_m(x)$ with a high bias. The final $F_M(x)$ is a strong learner with a low bias. This is done by optimizing a differentiable loss function,

$$\frac{1}{n}\sum_{i=1}^{n} \ell(F_j(x_i), y_i).$$

General idea considering one sample

- At stage j, the learner $F_j(x)$ incurs the loss $\ell(F_j(x), y)$
- At stage $j+1$, the weak learner $f_{j+1}(x)$ is trained to approximate the gradient

$$-\frac{\partial \ell(F_j(x), y)}{\partial F_j(x)}$$

- At stage $j+1$, the learner is updated as

$$F_{j+1}(x) = F_j(x) + \eta_{j+1} f_{j+1}(x)$$

Algorithm

Considering the training set $\{(x_i, y_i)\}_{i=1}^n$, and a differentiable loss function $\ell(F(x), y)$, the algorithm consists of the following steps.

1. Initializing the model with a constant value

$$F_0(x) = \arg\min_{\gamma} \sum_{i=1}^{n} \ell(y_i, \gamma)$$

2. For $m = 1$ to M
   a) Computing

$$r_{im} = -\frac{\partial \ell(y_i, F_{m-1}(x_i))}{\partial F_{m-1}(x_i)} \quad \text{for } i = 1, \ldots, n.$$

   b) Fitting a base (weak) model $f_m(x)$ to the training set $\{(x_i, r_{im})\}_{i=1}^n$.
   c) Computing the multiplier η_m using

$$\eta_m = \arg\min_{\eta} \sum_{i=1}^{n} \ell(y_i, F_{m-1}(x_i) + \eta h_m(x_i))$$

   d) Updating the model

$$F_m(x) = F_{m-1}(x) + \eta_m f_m(x)$$

3. Outputting $F_M(x)$

Example: Regression with Squared Loss

Considering the regression problem with squared loss $\ell(y, F(x)) = \frac{1}{2}(F(x) - y)^2$, the negative of the gradient is equal to residuals

$$r_{im} = -\frac{\partial \ell(y_i, F_{m-1}(x_i))}{\partial F_{m-1}(x_i)} = y_i - F_{m-1}(x_i) \text{ for } i = 1, \ldots, n.$$

Therefore, the labels for our training dataset at stage m are

$$\begin{gathered} y_1 - F_{m-1}(x_1) \\ y_2 - F_{m-1}(x_2) \\ \vdots \\ y_n - F_{m-1}(x_n) \end{gathered}$$

and we fit a weak learner which can be a small decision tree to these residuals.

References

1. Kevin P. Murphy, Machine Learning: A Probabilistic Perspective.
2. Trevor Hastie, Robert Tibshirani, and Jerome Friedman, The Elements of Statistical Learning.