Mathematics behind GBM

Wenbo Ma

Oct-Nov 2016

1 Gradient Boosting

Mathematical Representation:

$\hat{y}^{(i)} = H(x^{(i)}) = \sum_{t=1}^{t=T} f_t(x^{(i)})$ where f_t is a CART

Define an observation-wise loss function $L = Loss(y^{(i)}, H(x^{(i)}))$

The parameters in this model are the structure and leaf score for each CART. CART is trained in an additive fashion which means at step t, all previous CART are fixed.

At step t

$$
\begin{aligned}
obj^{(t)} &= [\sum_{i=1}^{i=m} Loss(y^{(i)}, \hat{y}_t^{(i)}) + \lambda \sum_{i=1}^{i=t} \Omega(f_i)] \\
&= [\sum_{i=1}^{i=m} Loss(y^{(i)}, \hat{y}_{t-1}^{(i)} + f_t(x^{(i)})) + \lambda \sum_{i=1}^{i=t} \Omega(f_i)] \\
&= [\sum_{i=1}^{i=m} Loss(y^{(i)}, f_t(x^{(i)})) + \Omega(f_t)] + const
\end{aligned}
$$

Ω is the complexity for a single CART.

Approximate $Loss(y^{(i)}, f_t(x^{(i)}))$ with first order and second order Taylor expansion

$$
\begin{aligned}
Loss(y^i, \hat{y}_t^{(i)}) &= Loss(y^i, H_\theta(x^i)) \\
&= Loss(y^i, \hat{y}_{t-1}^i + f_t(x^i)) \\
&\approx Loss(y^i, \hat{y}_{t-1}^i) + \frac{\partial Loss(y^i, H_\theta(x^i))}{\partial H_\theta(x^i)}\Big|_{H_\theta(x^i)=\hat{y}_{t-1}^i} f_t(x^i) + \frac{1}{2}\frac{\partial^2 Loss(y^i, H_\theta(x^i))}{\partial H_\theta(x^i)^2}\Big|_{H_\theta(x^i)=\hat{y}_{t-1}^i} f_t^2(x^i)
\end{aligned}
$$

so the object function at step t:

$$
obj^{(t)} \approx [\sum_{i=1}^{i=m} \frac{\partial Loss(y^i, H_\theta(x^i))}{\partial H_\theta(x^i)}\Big|_{H_\theta(x^i)=\hat{y}_{t-1}^i} f_t(x^i) + \frac{1}{2}\frac{\partial^2 Loss(y^i, H_\theta(x^i))}{\partial H_\theta(x^i)^2}\Big|_{H_\theta(x^i)=\hat{y}_{t-1}^i} f_t^2(x^i)] + \Omega(f_t) + const
$$

(1) Each single CART can be defined as

$$f_t = w_{q(x^i)}$$

$w \in R^T$ and $q : R^d -> 1, 2, ..., T$.

w is a T*1 vector where each element is the score for a leaf; q is a function which projects an observation(sample) to a leaf; T is the total number of leaves.

(2) The model complexity of a single tree can be defined as:

$$\Omega(f_t) = \frac{1}{2}\lambda \sum_{i=1}^{i=T} w_i^2 + \gamma T$$

(3) let

$$g(x^i) = \frac{\partial Loss(y^i, H_\theta(x^i))}{\partial H_\theta(x^i)}\Big|_{H_\theta(x^i)=\hat{y}_{t-1}^i}$$

and

$$k(x^i) = \frac{1}{2}\frac{\partial^2 Loss(y^i, H_\theta(x^i))}{\partial H_\theta(x^i)^2}\Big|_{H_\theta(x^i)=\hat{y}_{t-1}^i}$$

So plug (1),(2) and (3) into the objective function and ignore constant, we get

$$\begin{aligned}
obj^t &\approx \sum_{i=1}^{m}[g(x^i)f_t(x^i) + k(x^i)\frac{1}{2}f_t^2(x^i)] + \frac{1}{2}\lambda\sum_{j=1}^{T} w_j^2 + \gamma T \\
&= \sum_{i=1}^{m}[g(x^i)w_{q(x^i)} + k(x^i)\frac{1}{2}w_{q(x^i)}^2] + \frac{1}{2}\lambda\sum_{j=1}^{T} w_j^2 + \gamma T \\
&= \sum_{j=1}^{T}[(\sum_{i\in I_j} g_i w_j) + (\sum_{i\in I_j} k_i \frac{1}{2}w_j^2)] + \frac{1}{2}\lambda\sum_{j=1}^{T} w_j^2 + \gamma T \\
&= \sum_{j=1}^{T}[w_j(\sum_{i\in I_j} g_i) + \frac{1}{2}w_j^2(\sum_{i\in I_j}(k_i) + \lambda)] + \gamma T
\end{aligned}$$

let $G_j = \sum_{i\in I_j} g_i$ and $K_j = \sum_{i\in I_j} k_i$, we get

$$obj^t = \sum_{j=1}^{T}[w_j G_j + \frac{1}{2}w_j^2(K_j + \lambda)] + \gamma T$$

Let us assume T is fixed at this point, so the optimum value is taken when $w_j^* = -\frac{G_j}{K_j+\lambda}$. The optimal objective value is $obj^* = -\frac{1}{2}\sum_{j=1}^{T}\frac{G_j^2}{(H_j+\lambda)}$.

As enumerate all the possible true structure is intractable, we grow the tree by splitting a leaf at a time and evaluate with the structure score obj_* above.

$$
\begin{aligned}
obj^0 &= -\frac{1}{2}\frac{G_0^2}{H_0+\lambda} + \lambda * 1 \\
obj^L &= -\frac{1}{2}\frac{G_L^2}{H_L+\lambda} + \lambda * 1 \\
obj^R &= -\frac{1}{2}\frac{G_R^2}{H_R+\lambda} + \lambda * 1
\end{aligned}
$$

As a smaller obj value means a better structure, the benefit getting from splitting can be measured as

$$
\begin{aligned}
Benefit &= obj^0 - (obj^R + obj^L) \\
&= \frac{1}{2}[\frac{G_L^2}{H_L+\lambda} + \frac{G_R^2}{H_R+\lambda} - \frac{G_0^2}{H_0+\lambda}] - \lambda
\end{aligned}
$$

Therefore, the algorithm will split until $\frac{1}{2}[\frac{G_L^2}{H_L+\lambda} + \frac{G_R^2}{H_R+\lambda} - \frac{G_0^2}{H_0+\lambda}] < \lambda$,