Equivalence of Gradient BDT and Adaptive BDT

Ligang Xia
Warwick University

October 5, 2018

Introduction

- Here is the comparison between Adaptive and Gradient BDT methods.
- These slides are to show the equivalence between AdaBDT and GradBDT methods.

Quantity	AdaBDT value	GradBDT value
input variables	$\vec{x} = (x_1, x_2, \cdots)$	same
true value Y	-1, 1	same
guess value k_m	-1, 1	none
guess weight α_m/w_m	$\alpha_m = \frac{1}{2} \ln \frac{1-\epsilon_m}{\epsilon_m}$	negative gradient, $w_m = -\frac{\partial L_m}{\partial y_{m-1}}$
tree update	apply a weight ofe^{α_m} to wrong guess	fit the residues
final BDT score y_m	$y_m = y_{m-1} + \alpha_m k_m$	$y_m = y_{m-1} + w_m$
loss function	$L_m(\vec{x}, y_m) = \sum_{\vec{x}} e^{-Y(\vec{x}) y_m(\vec{x})}$	any form

- Here ϵ_m is the miscalssification rate for the $m-$th tree.
- w_m is the weight at the m-th tree. It is roughly negative of the derivative of the loss function evaluated at y_{m-1}.
- For GradBDT, the loss function could be in any form. For example, $L_m(\vec{x}, y_m) = \sum_{\vec{x}} \frac{1}{2}(y_m(\vec{x}) - Y(\vec{x}))^2$.

Review of Gradient BDT

Suppose we have m trees, the BDT score is

$$y_m(\vec{x}_i) = y_{m-1}(\vec{x}_i) + w_m(\vec{x}_i)\ . \tag{1}$$

Taking $w_m(\vec{x}_i)$ as a small quantity and expanding the m-th loss fuction around $y_{m-1}(\vec{x}_i)$, we have

$$\begin{aligned} L_m \equiv & \sum_{\vec{x}_i} l(y_m(\vec{x}_i)) = \sum_{\vec{x}_i} l(y_{m-1}(\vec{x}_i) + w_m(\vec{x}_i)) && (2) \\ \approx & \sum_{\vec{x}_i} l(y_{m-1}(\vec{x}_i)) + d_{m-1}(\vec{x}_i) w_m(\vec{x}_i) + \frac{1}{2} h_{m-1}(\vec{x}_i) w_m^2(\vec{x}_i) && (3) \\ & && (4) \end{aligned}$$

with

$$d_{m-1} \equiv \frac{\partial l(y)}{\partial y}|_{y=y_{m-1}}\ , \quad h_{m-1} \equiv \frac{\partial^2 l(y)}{\partial y^2}|_{y=y_{m-1}}\ . \tag{5}$$

In practice, each tree will only have limited number of terminal nodes (denoted by J). The events falling into the same terminal node (denoted by $R_j, j = 1, 2, \cdots, J$) will be given the same weight, $w_m(R_j)$. The loss function at the m-th three then becomes

$$\begin{aligned} L_m \approx & \sum_{\vec{x}_i} l(y_{m-1}(\vec{x}_i)) + d_{m-1}(\vec{x}_i) w_m(\vec{x}_i) + \frac{1}{2} h_{m-1}(\vec{x}_i) w_m^2(\vec{x}_i) && (6) \\ = & L_{m-1} + \sum_{j=1}^{J} \left(\sum_{\vec{x}_i \in R_j} d_{m-1}(\vec{x}_i) \right) w_m(R_j) + \left(\sum_{\vec{x}_i \in R_j} h_{m-1}(\vec{x}_i) \right) w_m^2(R_j) && (7) \end{aligned}$$

Review of Gradient BDT

$$L_m \approx \quad L_{m-1} + \sum_{j=1}^{J} \left(\sum_{\vec{x}_i \in R_j} d_{m-1}(\vec{x}_i) \right) w_m(R_j) + \frac{1}{2} \left(\sum_{\vec{x}_i \in R_j} h_{m-1}(\vec{x}_i) \right) w_m^2(R_j) \tag{8}$$

Minimizing the loss function gives

$$w_m(R_j) = -\frac{\sum_{\vec{x}_i \in R_j} d_{m-1}(\vec{x}_i)}{\sum_{\vec{x}_i \in R_j} h_{m-1}(\vec{x}_i)} \ . \tag{9}$$

and the reduction of the loss function at this point is

$$\Delta L_m \equiv L_m - L_{m-1} = -\frac{1}{2} \sum_{j=1}^{J} \frac{\left(\sum_{\vec{x}_i \in R_j} d_{m-1}(\vec{x}_i)\right)^2}{\sum_{\vec{x}_i \in R_j} h_{m-1}(\vec{x}_i)} \ . \tag{10}$$

This is used to determine the splitting in building a tree, i.e., to maximize (Let $R = R_l + R_r$ denote a node and its daughter nodes, left node R_l and right node R_r.)

$$\frac{1}{2} \frac{\left(\sum_{\vec{x}_i \in R_l} g(\vec{x}_i)\right)^2}{\sum_{\vec{x}_i \in R_l} h(\vec{x}_i)} + \frac{1}{2} \frac{\left(\sum_{\vec{x}_i \in R_r} g(\vec{x}_i)\right)^2}{\sum_{\vec{x}_i \in R_r} h(\vec{x}_i)} - \frac{1}{2} \frac{\left(\sum_{\vec{x}_i \in R} g(\vec{x}_i)\right)^2}{\sum_{\vec{x}_i \in R} h(\vec{x}_i)} \tag{11}$$

Probability distribution function of the Gradient BDT score

Let $g_m(y_m)$ be the probability distribution function of the BDT score after m trees.

$$\int_{y}^{y+\delta} g_m(y_m)dy_m = \int_{y<y_m(\vec{x})<y+\delta} f(\vec{x})d\vec{x} \tag{12}$$

This is difficult as $y(\vec{x})$ is unknown. We use the iteration relation from $(m-1)$-th tree to the m-th tree

$$y_m(\vec{x}_i) = y_{m-1}(\vec{x}_i) + \sum_{j=1}^{J} \delta_{\vec{x}_i,R_j} w_m(R_j) \tag{13}$$

Here $\delta_{\vec{x}_i,R_j}$ is 1 if $\vec{x}_i$ falls into the node R_j and 0 othrwise. Note that all terminal nodes R_j do not overlap.

$$w_m(R_j) = -\frac{\sum_{\vec{x}_i\in R_j} d_{m-1}(\vec{x}_i)}{\sum_{\vec{x}_i\in R_j} h_{m-1}(\vec{x}_i)} \ . \tag{14}$$

For simplification, let us use the loss fucntion $l(y(\vec{x}_i)) = \frac{1}{2}(y(\vec{x}_i) - Y(\vec{x}_i))^2$ where Y is the true value (1 for signal events and -1 for background events). Then (let N_{R_j} denote number of events in R_j)

$$d_{m-1}(\vec{x}_i) = y_{m-1}(\vec{x}_i) - Y(\vec{x}_i) \ , \quad h_{m-1}(\vec{x}_i) = 1 , \quad w_m(R_j) = -\frac{1}{N_{R_j}} \sum_{\vec{x}_i\in R_j} y_{m-1}(\vec{x}_i) - Y(\vec{x}_i) . \tag{15}$$

Probability distribution function of the Gradient BDT score

$$y_m(\vec{x}_i) = y_{m-1}(\vec{x}_i) - \sum_{j=1}^{J} \delta_{\vec{x}_i, R_j} \frac{1}{N_{R_j}} \left(\sum_{\vec{x}_i \in R_j} y_{m-1}(\vec{x}_i) - Y(\vec{x}_i) \right) \tag{16}$$

Let p_{m,R_j} denote the signal fraction in the node R_j (the background fraction is $1 - p_{m,R_j}$. Let f_{m,R_j} denote the fraction of total number of events in the node R_j. Let $S \cap R_j$ and $B \cap R_j$ denote the set of signal and background events in the node R_j. Then we have $p_{m,R_j} = N_{S\cap R_j}/N_{R_j}$, $\sum_j f_{m,R_j} = 1$, $\sum_j f_{m,R_j} p_{m,R_j} = \frac{1}{2}$ and $\sum_j f_{m,R_j}(1 - p_{m,R_j}) = \frac{1}{2}$(this is because we have renormalized all events to 1 for signal and background by definition).

$$\begin{aligned}
-w_m(\vec{x}_i) &= \sum_{j=1}^{J} \delta_{\vec{x}_i, R_j} \frac{1}{N_{R_j}} \left(\sum_{\vec{x}_i \in R_j} y_{m-1}(\vec{x}_i) - Y(\vec{x}_i) \right) && (17)\\
&= \sum_{j=1}^{J} \delta_{\vec{x}_i, R_j} \frac{1}{N_{R_j}} \left(N_{S\cap R_j} \frac{\sum_{\vec{x}_i \in S\cap R_j} y_{m-1}(\vec{x}_i) - 1}{N_{S\cap R_j}} + N_{B\cap R_j} \frac{\sum_{\vec{x}_i \in B\cap R_j} y_{m-1}(\vec{x}_i) + 1}{N_{B\cap R_j}} \right) && (18)\\
&= \sum_{j=1}^{J} \delta_{\vec{x}_i, R_j} \left[p_{m,R_j}(z^S_{m-1,R_j} - 1) + (1 - p_{m,R_j})(z^B_{m-1,R_j} + 1) \right] , && (19)
\end{aligned}$$

where

$$z^S_{m-1,R_j} \equiv \frac{\sum_{\vec{x}_i \in S\cap R_j} y_{m-1}(\vec{x}_i)}{N_{S\cap R_j}} , \quad z^B_{m-1,R_j} \equiv \frac{\sum_{\vec{x}_i \in B\cap R_j} y_{m-1}(\vec{x}_i)}{N_{B\cap R_j}} . \tag{20}$$

Probability distribution function of the Gradient BDT score

$$z^S_{m-1,R_j} \equiv \frac{\sum_{\vec{x}_i \in S \cap R_j} y_{m-1}(\vec{x}_i)}{N_{S\cap R_j}}\ , \quad z^B_{m-1,R_j} \equiv \frac{\sum_{\vec{x}_i \in B \cap R_j} y_{m-1}(\vec{x}_i)}{N_{B\cap R_j}}\ . \tag{21}$$

Let μ_{m-1} and σ_{m-1} denote the expectation value and variance of the distribution of y_{m-1}. They are different between signal and background generally. We assume they exist and also the Central Limit Theorem (CLT) applies here. z_{m-1,R_j} will abide by a Gaussian distribution (let $G(x|\mu,\sigma) \equiv \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$).

$$z^S_{m-1,R_j} \sim G(\mu^S_{m-1}, \frac{\sigma^S_{m-1}}{\sqrt{N_{S\cap R_j}}})\ , \quad z^B_{m-1,R_j} \sim G(\mu^B_{m-1}, \frac{\sigma^B_{m-1}}{\sqrt{N_{B\cap R_j}}}) \tag{22}$$

In the limit of large sample size, $z^{S/B}_{m-1,R_j}$ would be very peaky around the mean value $\mu^{S/B}_{m-1}$.

$$w_m(\vec{x}_i) \quad \approx \quad -\sum_{j=1}^{J} \delta_{\vec{x}_i,R_j} \left[p_{m,R_j}(\mu^S_{m-1} - 1) + (1 - p_{m,R_j})(\mu^B_{m-1} + 1)\right] \tag{23}$$

Let us consider the possible values for $w_m(\vec{x}_i)$ and the corresponding probilities. The probability of a signal event falling to the node R_j should be proportional to the fraction of signal events in R_j. This is $f_{m,R_j} p_{m,R_j} / \sum_{j=1}^{J} f_{m,R_j} p_{m,R_j} = 2 f_{m,R_j} p_{m,R_j}$ (similar argument applies to background).

Probability distribution function of the Gradient BDT score

Let us only consider two ndoes, namely, $J = 2$ (it will be shown to be equivalent to the Adaptive BDT). $w_m(\vec{x}_i)$ takes only two possible values. We have

$$
\begin{aligned}
& -w_m(\vec{x}_i) \\
\approx\quad & \delta_{\vec{x}_i,R_1}\left[p_{m,R_1}(\mu^S_{m-1}-1)+(1-p_{m,R_1})(\mu^B_{m-1}+1)\right] && (24)\\
& +\delta_{\vec{x}_i,R_2}\left[p_{m,R_2}(\mu^S_{m-1}-1)+(1-p_{m,R_2})(\mu^B_{m-1}+1)\right] && (25)\\
=\quad & \delta_{\vec{x}_i,R_1}\left[p_{m,R_1}(\mu^S_{m-1}-1)+(1-p_{m,R_1})(\mu^B_{m-1}+1)\right] && (26)\\
& +\delta_{\vec{x}_i,R_2}\left[\frac{1-2f_{m,R_1}p_{m,R_1}}{2(1-f_{m,R_1})}(\mu^S_{m-1}-1)+\frac{1-2f_{m,R_1}+2f_{m,R_1}p_{m,R_1}}{2(1-f_{m,R_1})}(\mu^B_{m-1}+1)\right] && (27)
\end{aligned}
$$

and the probability for a signal event falling in R_j

$$
\begin{aligned}
\mathrm{Prob}(\vec{x}_i\in R_1) &= 2f_{m,R_1}p_{m,R_1} && (28)\\
\mathrm{Prob}(\vec{x}_i\in R_2) &= 1-2f_{m,R_1}p_{m,R_1}\ . && (29)
\end{aligned}
$$

For convenience in the case of $J = 2$, we can drop the subscripts, namely, letting $f \equiv f_{m,R_1}$ and $p \equiv p_{m,R_1}$.

$$
\begin{aligned}
-w_m(\vec{x}_i)\approx\quad & \delta_{\vec{x}_i,R_1}\left[p(\mu^S-1)+(1-p)(\mu^B+1)\right] && (30)\\
& +\delta_{\vec{x}_i,R_2}\left[\frac{1-2fp}{2(1-f)}(\mu^S-1)+\frac{1-2f+2fp}{2(1-f)}(\mu^B+1)\right] && (31)
\end{aligned}
$$

Probability distribution function of the Gradient BDT score

$$
\begin{aligned}
-w_m(\vec{x}_i) \approx\quad & \delta_{\vec{x}_i,R_1}\left[p(\mu^S-1)+(1-p)(\mu^B+1)\right] && (32)\\
& +\delta_{\vec{x}_i,R_2}\left[\frac{1-2fp}{2(1-f)}(\mu^S-1)+\frac{1-2f+2fp}{2(1-f)}(\mu^B+1)\right] && (33)\\
\approx\quad & \delta_{\vec{x}_i,R_1}\left[p(\mu^S-1)+(1-p)(-\mu^S+1)\right] && (34)\\
& +\delta_{\vec{x}_i,R_2}\left[\frac{1-2fp}{2(1-f)}(\mu^S-1)+\frac{1-2f+2fp}{2(1-f)}(-\mu^S+1)\right] && (35)\\
\approx\quad & \delta_{\vec{x}_i,R_1}(2p-1)(\mu^S-1)+\delta_{\vec{x}_i,R_2}\frac{-f}{1-f}(2p-1)(\mu^S-1) && (36)
\end{aligned}
$$

Here we used $\mu^S = -\mu^B$, which is expected as signal and background play an equal role. Keeping in mind that in Gradient BDT, the split into two nodes (this affects f and p) is determined by maximizing the reduction of the loss function, this is to maximize (from Eq.(11))

$$
\begin{aligned}
& \frac{1}{2}fw_m(R_1)^2+\frac{1}{2}(1-f)w_m(R_2)^2 && (37)\\
\approx\quad & \frac{1}{2}f(2p-1)^2(\mu^S-1)^2+\frac{1}{2}(1-f)\left(\frac{-f}{1-f}\right)^2(2p-1)^2(\mu^S-1)^2 && (38)\\
=\quad & \frac{1}{2}\frac{f}{1-f}(2p-1)^2(\mu^S-1)^2 && (39)
\end{aligned}
$$

Probability distribution function of the Gradient BDT score

Noting that the split affects both f and p, we can take p as a function of f and $p(1) = \frac{1}{2}$ (if a node has all the events, then the signal fraction in that node is $\frac{1}{2}$ due to the initial renormalization). We assume all trees are weak learners. Under this assumption, we expect that p should be around $\frac{1}{2}$ and has little dependence upon f actually. Expanding $p(f) \approx p(1) + \frac{dp}{df}(f-1) = \frac{1}{2} + \frac{dp}{df}(f-1)$, we have

$$
\begin{aligned}
& \frac{1}{2} f w_m(R_1)^2 + \frac{1}{2}(1-f) w_m(R_2)^2 & (40)\\
\approx \quad & \frac{1}{2}\frac{f}{1-f}(2p(f)-1)^2(\mu^S-1)^2 & (41)\\
\approx \quad & \frac{1}{2}\left(\frac{dp}{df}\right)^2 f(1-f)(\mu^S-1)^2 \,. & (42)
\end{aligned}
$$

We now see that the maximization gives $f = \frac{1}{2}$, which value is also consistent with our intuitive understanding. Ok, let us summarize all keys below (recovering the subscripts). For signal, we have

$$y_m(\vec{x}_i) = y_{m-1}(\vec{x}_i) + w_m(\vec{x}_i) \tag{43}$$

$$w_m(\vec{x}_i) = \begin{cases} -(2p_{m,R_1}-1)(\mu^S_{m-1}-1) & \text{Prob}(\vec{x}_i \in R_1) = p_{m,R_1} \\ (2p_{m,R_1}-1)(\mu^S_{m-1}-1) & \text{Prob}(\vec{x}_i \in R_1) = 1 - p_{m,R_1} \end{cases} . \tag{44}$$

Probability distribution function of the Gradient BDT score

For signal, we have

$$y_m(\vec{x}_i) = y_{m-1}(\vec{x}_i) + w_m(\vec{x}_i) \tag{45}$$

$$w_m(\vec{x}_i) = \begin{cases} -(2p_{m,R_1} - 1)(\mu^S_{m-1} - 1) & \text{Prob}(\vec{x}_i \in R_1) = p_{m,R_1} \\ +(2p_{m,R_1} - 1)(\mu^S_{m-1} - 1) & \text{Prob}(\vec{x}_i \in R_1) = 1 - p_{m,R_1} \end{cases} . \tag{46}$$

Similarly, for background, we have

$$y_m(\vec{x}_i) = y_{m-1}(\vec{x}_i) + w_m(\vec{x}_i) \tag{47}$$

$$w_m(\vec{x}_i) = \begin{cases} -(2p_{m,R_1} - 1)(\mu^B_{m-1} + 1) & \text{Prob}(\vec{x}_i \in R_1) = 1 - p_{m,R_1} \\ +(2p_{m,R_1} - 1)(\mu^B_{m-1} + 1) & \text{Prob}(\vec{x}_i \in R_1) = p_{m,R_1} \end{cases} . \tag{48}$$

Let us consider only the signal BDT score distribution, $g_m(y_m)$, for the moment. Taking y_{m-1} and w_m as random variales and assuming they are independent, we have

$$\int_y^{y+\delta} g_m(y_m)dy_m = \int_{y<y_m<y+\delta} g_{m-1}(y_{m-1})f(w_m)dy_{m-1}dw_m \tag{49}$$

$$= \int_y^{y+\delta} \left[g_{m-1}(y_m - w_m(R_1))p_{m,R_1} + g_{m-1}(y_m - w_m(R_2))(1 - p_{m,R_1})\right] \tag{50}$$

From the equation above, we derive that

$$g_m(y) = g_{m-1}(y+(2p_{m,R_1}-1)(\mu^S_{m-1}-1))p_{m,R_1} + g_{m-1}(y-(2p_{m,R_1}-1)(\mu^S_{m-1}-1))(1-p_{m,R_1}) . \tag{51}$$

Probability distribution function of the Gradient BDT score

$$g_m(y) = g_{m-1}(y+(2p_{m,R_1}-1)(\mu^S_{m-1}-1))p_{m,R_1} + g_{m-1}(y-(2p_{m,R_1}-1)(\mu^S_{m-1}-1))(1-p_{m,R_1}) \,. \tag{52}$$

This is actually equivalent to the iteration formula below in the Adaptive BDT method.

$$g_m(y) = g_{m-1}(y-\alpha_m)(1-\epsilon_m) + g_{m-1}(y+\alpha_m)\epsilon_m \tag{53}$$

The corresponding relation is

$$p_{m,R_1} = \epsilon_m \,, \tag{54}$$

$$(2p_{m,R_1}-1)(\mu^S_{m-1}-1) = \alpha_m \,. \tag{55}$$

This shows the equivalence between the Adaptive BDT and the Gradient BDT. It is not difficult to derive the PDF for the GBDT score, $g_m(y)$. Let us investigate the mean value in the first place (because it has different behavior compared to AdaBDT).

$$\mu_m = \int_{-\infty}^{+\infty} y g_m(y) dy \tag{56}$$

Using the iteration formula, it is easy to find that

$$\mu^S_m = \mu^S_{m-1} - (2p_{m,R_1}-1)^2(\mu^S_{m-1}-1) \tag{57}$$

Probability distribution function of the Gradient BDT score

$$
\begin{aligned}
\mu_m^S &= \mu_{m-1}^S - (2p_{m,R_1} - 1)^2(\mu_{m-1}^S - 1) && (58)\\
\mu_m^S - 1 &= \mu_{m-1}^S - 1 - (2p_{m,R_1} - 1)^2(\mu_{m-1}^S - 1) && (59)\\
\mu_m^S - 1 &= [1 - (2p_{m,R_1} - 1)^2](\mu_{m-1}^S - 1) && (60)\\
\mu_m^S - 1 &= 4p_{m,R_1}(1 - p_{m,R_1})(\mu_{m-1}^S - 1) && (61)\\
&= \Pi_{i=1}^m 4p_{i,R_1}(1 - p_{i,R_1})(\mu_0^S - 1) && (62)\\
& && (63)
\end{aligned}
$$

Therefore, we have

$$
\mu_m^S = 1 + \Pi_{i=1}^m 4p_{i,R_1}(1 - p_{i,R_1})(\mu_0^S - 1)\ . \tag{64}
$$

Noting that $4p_{i,R_1}(1 - p_{i,R_1}) \leq 1$ (the equal sign holds only if $p_{i,R_1} = \frac{1}{2}$. If it happens, the training will stop because the loss function cannot be reduced further.), we thus have

$$
\lim_{m\to+\infty} \mu_m^S = 1\ , \tag{65}
$$

which is independent upon the choice of the intial value. Similarly, we can show that $\lim_{m\to+\infty} \mu_m^B = -1$.

Probability distribution function of the Gradient BDT score

$$(\sigma_m^S)^2 = \int_{-\infty}^{+\infty} (y - \mu_m^S) g_m(y) dy \tag{66}$$

We can obtain

$$
\begin{aligned}
(\sigma_m^S)^2 &= (\sigma_{m-1}^S)^2 + 4(p_{m,R_1} - 1)^2 \alpha_m^2 &(67)\\
&= \sum_{i=1}^{m} 4(p_{i,R_1} - 1)^2 \alpha_i^2 &(68)\\
&\approx \sum_{i=1}^{m} \alpha_i^2 \ . &(69)
\end{aligned}
$$

BACK UP

BACK UP