Exercise of Supervised Learning: Boosting Part 2

Yawei Li

yawei.li@stat.uni-muenchen.de

January 12, 2024

Exercise 1: Gradient Boosting

In the following, you assume that your outcome follows a $\log_2$-normal distribution with density function

$$p(y|f) = \frac{1}{y\sigma\sqrt{2\pi}} \exp\left(-\frac{(\log_2(y) - f)^2}{2\sigma^2}\right) \qquad \lhd$$

where $\sigma = 1$. In other words, $\log_2(Y)$ follows a normal distribution. You observe $n = 3$ data points $\mathbf{y}$ and want to model f using features $\mathbf{X} \in \mathbb{R}^{n \times p}$. You choose to use a gradient boosting tree algorithm.

(a) Derive the pseudo residuals based on the negative log-likelihood for the given distribution assumption.

Solution to Exercise 1 (a)

(a) Derive the pseudo residuals based on the negative log-likelihood for the given distribution assumption.

- The loss is calulated by the NLL by:

$$L(y, f) = -\ell(f) = -(const. - (\log_2(y) - f)^2/2).$$

- The pseudo residuals are:

$$\tilde{r}(f) = \partial L(y, f)/\partial f = (\log_2(y) - f).$$

Solution to Exercise 1 (a)

(a) Derive the pseudo residuals based on the negative log-likelihood for the given distribution assumption.

- The loss is calulated by the NLL by:

$$L(y, f) = -\ell(f) = -(const. - (\log_2(y) - f)^2/2).$$

- The pseudo residuals are:

$$\tilde{r}(f) = \partial L(y, f)/\partial f = (\log_2(y) - f).$$

Solution to Exercise 1 (a)

(a) Derive the pseudo residuals based on the negative log-likelihood for the given distribution assumption.

- The loss is calulated by the NLL by:

$$L(y, f) = -\ell(f) = -(const. - (\log_2(y) - f)^2/2).$$

- The pseudo residuals are:

$$\tilde{r}(f) = \partial L(y, f)/\partial f = (\log_2(y) - f).$$

Exercise 1 (b)

(b) Given only the 3 samples $\mathbf{y} = (1, 2, 4)^T$ $\quad \triangleleft$
and two features

$$\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2) = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 0 \end{pmatrix} \qquad \triangleleft$$

explicitly derive or state with explanation

(i) the loss-optimal initial boosting model $\hat{f}^{[0]}(\mathbf{x})$,

(ii) the pseudo residual $\tilde{r}^{[1]}$,

(iii) the regression stump $R_t^{[1]}$, $t = 1, 2$,

(iv) the boosting model $\hat{f}^{[1]}(\mathbf{x})$ as well as

(v) the pseduo residual $\tilde{r}^{[2]}$

for tree base learners with depth $d = 1$ (stumps) and a learning rate of $\alpha = 1$.

Solution to Exercise 1 (b)

(b) (i) Derive the loss-optimal intial boosting model $\hat{f}^{[1]}(\mathbf{x})$.

- We initialize $\hat{f}^{[0]}(\mathbf{x}) = \arg\min_{f^{[0]}} \sum_{i=1}^{n} L(y^{(i)}, f^{[0]}(\mathbf{x}^{(i)}))$.
- It can be easily seen that $\hat{f}^{[0]}(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n} \log_2(y^{(i)}) = 1$, as it miminizes the squared error.

Solution to Exercise 1 (b)

(b) (i) Derive the loss-optimal intial boosting model $\hat{f}^{[1]}(\mathbf{x})$.

- We initialize $\hat{f}^{[0]}(\mathbf{x}) = \arg\min_{f^{[0]}} \sum_{i=1}^{n} L(y^{(i)}, f^{[0]}(\mathbf{x}^{(i)}))$.
- It can be easily seen that $\hat{f}^{[0]}(\mathbf{x}) = \frac{1}{n} \sum_{i=1}^{n} \log_2(y^{(i)}) = 1$, as it miminizes the squared error.

Solution to Exercise 1 (b)

(b) (i) Derive the loss-optimal intial boosting model $\hat{f}^{[1]}(\mathbf{x})$.

- We initialize $\hat{f}^{[0]}(\mathbf{x}) = \arg\min_{f^{[0]}} \sum\limits_{i=1}^{n} L(y^{(i)}, f^{[0]}(\mathbf{x}^{(i)}))$.
- It can be easily seen that $\hat{f}^{[0]}(\mathbf{x}) = \frac{1}{n} \sum\limits_{i=1}^{n} \log_2(y^{(i)}) = 1$, as it miminizes the squared error.

Solution to Exercise 1 (b): Continued

(b) (ii) Derive the pseudo residual $\tilde{r}^{[1]}$.

- From (a) we know $\tilde{r}(f) = \partial L(y, f)/\partial f = (\log_2(y) - f)$.
- Denote $\tilde{\boldsymbol{f}}^{[0]} = \left(\hat{f}^{[0]}(\mathbf{x}^{(1)}), \hat{f}^{[0]}(\mathbf{x}^{(2)}), \hat{f}^{[0]}(\mathbf{x}^{(3)})\right)^T = (1, 1, 1)^T$
- So

$$\begin{aligned}\tilde{r}^{[1]} &= \left(\log_2(y^{(1)}), \log_2(y^{(2)}), \log_2(y^{(3)})\right)^T - \tilde{\boldsymbol{f}}^{[0]} \\ &= (0, 1, 2)^T - (1, 1, 1)^T \\ &= (-1, 0, 1)^T.\end{aligned}$$

Solution to Exercise 1 (b): Continued

(b) (ii) Derive the pseudo residual $\tilde{r}^{[1]}$.

- From (a) we know $\tilde{r}(f) = \partial L(y, f)/\partial f = (\log_2(y) - f)$.
- Denote $\tilde{\boldsymbol{f}}^{[0]} = \left(\hat{f}^{[0]}(\mathbf{x}^{(1)}), \hat{f}^{[0]}(\mathbf{x}^{(2)}), \hat{f}^{[0]}(\mathbf{x}^{(3)})\right)^T = (1, 1, 1)^T$
- So

$$\begin{aligned}\tilde{r}^{[1]} &= \left(\log_2(y^{(1)}), \log_2(y^{(2)}), \log_2(y^{(3)})\right)^T - \tilde{\boldsymbol{f}}^{[0]} \\ &= (0, 1, 2)^T - (1, 1, 1)^T \\ &= (-1, 0, 1)^T.\end{aligned}$$

Solution to Exercise 1 (b): Continued

(b) (ii) Derive the pseudo residual $\tilde{r}^{[1]}$.

- From (a) we know $\tilde{r}(f) = \partial L(y, f)/\partial f = (\log_2(y) - f)$.
- Denote $\tilde{\boldsymbol{f}}^{[0]} = \left(\hat{f}^{[0]}(\mathbf{x}^{(1)}), \hat{f}^{[0]}(\mathbf{x}^{(2)}), \hat{f}^{[0]}(\mathbf{x}^{(3)})\right)^T = (1, 1, 1)^T$
- So

$$\begin{aligned}\tilde{r}^{[1]} &= \left(\log_2(y^{(1)}), \log_2(y^{(2)}), \log_2(y^{(3)})\right)^T - \tilde{\boldsymbol{f}}^{[0]} \\ &= (0, 1, 2)^T - (1, 1, 1)^T \\ &= (-1, 0, 1)^T.\end{aligned}$$

Solution to Exercise 1 (b): Continued

(b) (ii) Derive the pseudo residual $\tilde{r}^{[1]}$.

- From (a) we know $\tilde{r}(f) = \partial L(y, f)/\partial f = (\log_2(y) - f)$.
- Denote $\tilde{\boldsymbol{f}}^{[0]} = \left(\hat{f}^{[0]}(\mathbf{x}^{(1)}), \hat{f}^{[0]}(\mathbf{x}^{(2)}), \hat{f}^{[0]}(\mathbf{x}^{(3)})\right)^T = (1, 1, 1)^T$
- So

$$
\begin{aligned}
\tilde{r}^{[1]} &= \left(\log_2(y^{(1)}), \log_2(y^{(2)}), \log_2(y^{(3)})\right)^T - \tilde{\boldsymbol{f}}^{[0]} \\
&= (0, 1, 2)^T - (1, 1, 1)^T \\
&= (-1, 0, 1)^T.
\end{aligned}
$$

Solution to Exercise 1 (b): Continued

(b) (iii) Derive the regression stump $R_t^{[1]}, t = 1, 2$.

- $R_t^{[1]}, t = 1, 2$ will split using $\mathbf{x}_1$, as $\mathbf{x}_2$ carries no information.
- Note that $x_1^{(1)} = x_1^{(2)}$.
- Recall that $\tilde{r}^{[1]} = (-1, 0, 1)^T$, and $R_t^{[1]}, t = 1, 2$ aim to fit this pseudo residual.
- $R_1 = -0.5 \cdot \boldsymbol{I}_{x_1 \geq 0.5}$, for which -0.5 stems from $\frac{1}{2}(\tilde{r}^{[1](1)} + \tilde{r}^{[1](2)})$ because $\mathbf{x}_1^{(1)}, \mathbf{x}_1^{(2)} \geq 0.5$.
- $R_2 = 1 \cdot \boldsymbol{I}_{x_1 < 0.5}$, for which 1 stems from $\tilde{r}^{[1](3)}$ because only $\mathbf{x}_1^{(3)} < 0.5$.

Solution to Exercise 1 (b): Continued

(b) (iii) Derive the regression stump $R_t^{[1]}, t = 1, 2$.

- $R_t^{[1]}, t = 1, 2$ will split using $\mathbf{x}_1$, as $\mathbf{x}_2$ carries no information.
- Note that $x_1^{(1)} = x_1^{(2)}$.
- Recall that $\tilde{r}^{[1]} = (-1, 0, 1)^T$, and $R_t^{[1]}, t = 1, 2$ aim to fit this pseudo residual.
- $R_1 = -0.5 \cdot \boldsymbol{I}_{x_1 \geq 0.5}$, for which -0.5 stems from $\frac{1}{2}(\tilde{r}^{[1](1)} + \tilde{r}^{[1](2)})$ because $\mathbf{x}_1^{(1)}, \mathbf{x}_1^{(2)} \geq 0.5$.
- $R_2 = 1 \cdot \boldsymbol{I}_{x_1 < 0.5}$, for which 1 stems from $\tilde{r}^{[1](3)}$ because only $\mathbf{x}_1^{(3)} < 0.5$.

Solution to Exercise 1 (b): Continued

(b) (iii) Derive the regression stump $R_t^{[1]}$, $t = 1, 2$.

- $R_t^{[1]}$, $t = 1, 2$ will split using $\mathbf{x}_1$, as $\mathbf{x}_2$ carries no information.
- Note that $x_1^{(1)} = x_1^{(2)}$.
- Recall that $\tilde{r}^{[1]} = (-1, 0, 1)^T$, and $R_t^{[1]}$, $t = 1, 2$ aim to fit this pseudo residual.
- $R_1 = -0.5 \cdot \mathbb{I}_{x_1 \geq 0.5}$, for which -0.5 stems from $\frac{1}{2}(\tilde{r}^{[1](1)} + \tilde{r}^{[1](2)})$ because $\mathbf{x}_1^{(1)}, \mathbf{x}_1^{(2)} \geq 0.5$.
- $R_2 = 1 \cdot \mathbb{I}_{x_1 < 0.5}$, for which 1 stems from $\tilde{r}^{[1](3)}$ because only $\mathbf{x}_1^{(3)} < 0.5$.

Solution to Exercise 1 (b): Continued

(b) (iii) Derive the regression stump $R_t^{[1]}, t = 1, 2$.

- $R_t^{[1]}, t = 1, 2$ will split using $\mathbf{x}_1$, as $\mathbf{x}_2$ carries no information.
- Note that $x_1^{(1)} = x_1^{(2)}$.
- Recall that $\tilde{r}^{[1]} = (-1, 0, 1)^T$, and $R_t^{[1]}, t = 1, 2$ aim to fit this pseudo residual.
- $R_1 = -0.5 \cdot \boldsymbol{I}_{x_1 \geq 0.5}$, for which -0.5 stems from $\frac{1}{2}(\tilde{r}^{[1](1)} + \tilde{r}^{[1](2)})$ because $\mathbf{x}_1^{(1)}, \mathbf{x}_1^{(2)} \geq 0.5$.
- $R_2 = 1 \cdot \boldsymbol{I}_{x_1 < 0.5}$, for which 1 stems from $\tilde{r}^{[1](3)}$ because only $\mathbf{x}_1^{(3)} < 0.5$.

Solution to Exercise 1 (b): Continued

(b) (iii) Derive the regression stump $R_t^{[1]}, t = 1, 2$.

- $R_t^{[1]}, t = 1, 2$ will split using $\mathbf{x}_1$, as $\mathbf{x}_2$ carries no information.
- Note that $x_1^{(1)} = x_1^{(2)}$.
- Recall that $\tilde{r}^{[1]} = (-1, 0, 1)^T$, and $R_t^{[1]}, t = 1, 2$ aim to fit this pseudo residual.
- $R_1 = -0.5 \cdot \boldsymbol{I}_{x_1 \geq 0.5}$, for which -0.5 stems from $\frac{1}{2}(\tilde{r}^{[1](1)} + \tilde{r}^{[1](2)})$ because $\mathbf{x}_1^{(1)}, \mathbf{x}_1^{(2)} \geq 0.5$.
- $R_2 = 1 \cdot \boldsymbol{I}_{x_1 < 0.5}$, for which 1 stems from $\tilde{r}^{[1](3)}$ because only $\mathbf{x}_1^{(3)} < 0.5$.

Solution to Exercise 1 (b): Continued

(b) (iii) Derive the regression stump $R_t^{[1]}, t = 1, 2$.

- $R_t^{[1]}, t = 1, 2$ will split using $\mathbf{x}_1$, as $\mathbf{x}_2$ carries no information.
- Note that $x_1^{(1)} = x_1^{(2)}$.
- Recall that $\tilde{r}^{[1]} = (-1, 0, 1)^T$, and $R_t^{[1]}, t = 1, 2$ aim to fit this pseudo residual.
- $R_1 = -0.5 \cdot \boldsymbol{I}_{x_1 \geq 0.5}$, for which -0.5 stems from $\frac{1}{2}(\tilde{r}^{[1](1)} + \tilde{r}^{[1](2)})$ because $\mathbf{x}_1^{(1)}, \mathbf{x}_1^{(2)} \geq 0.5$.
- $R_2 = 1 \cdot \boldsymbol{I}_{x_1 < 0.5}$, for which 1 stems from $\tilde{r}^{[1](3)}$ because only $\mathbf{x}_1^{(3)} < 0.5$.

Solution to Exercise 1 (b): Continued

(b) (iv) Derive the boosting model $\hat{f}^{[1]}(\mathbf{x})$ (i.e., $\tilde{\boldsymbol{f}}^{[1]}$).

- Recall that $R_1^{[1]} = -0.5\boldsymbol{I}_{x_1 \geq 0.5}$ and $R_2^{[1]} = 1 \cdot \boldsymbol{I}_{x_2 < 0.5}$, and learning rate $\alpha = 1$.
- So the update direction given by the regression stump is $(-0.5, -0.5, 1)^T$.
- Therefore,

$$
\begin{aligned}
\tilde{\boldsymbol{f}}^{[1]} &= \tilde{\boldsymbol{f}}^{[0]} + 1 \cdot (-0.5, -0.5, 1)^T \\
&= (1, 1, 1)^t + (-0.5, -0.5, 1)^T \\
&= (0.5, 0.5, 2)^T
\end{aligned}
$$

Solution to Exercise 1 (b): Continued

(b) (iv) Derive the boosting model $\hat{f}^{[1]}(\mathbf{x})$ (i.e., $\tilde{\boldsymbol{f}}^{[1]}$).

- Recall that $R_1^{[1]} = -0.5\boldsymbol{I}_{x_1 \geq 0.5}$ and $R_2^{[1]} = 1 \cdot \boldsymbol{I}_{x_2 < 0.5}$, and learning rate $\alpha = 1$.
- So the update direction given by the regression stump is $(-0.5, -0.5, 1)^T$.
- Therefore,

$$\begin{aligned}\tilde{\boldsymbol{f}}^{[1]} &= \tilde{\boldsymbol{f}}^{[0]} + 1 \cdot (-0.5, -0.5, 1)^T \\ &= (1, 1, 1)^t + (-0.5, -0.5, 1)^T \\ &= (0.5, 0.5, 2)^T\end{aligned}$$

Solution to Exercise 1 (b): Continued

(b) (iv) Derive the boosting model $\hat{f}^{[1]}(\mathbf{x})$ (i.e., $\tilde{\boldsymbol{f}}^{[1]}$).

- Recall that $R_1^{[1]} = -0.5\boldsymbol{I}_{x_1 \geq 0.5}$ and $R_2^{[1]} = 1 \cdot \boldsymbol{I}_{x_2 < 0.5}$, and learning rate $\alpha = 1$.
- So the update direction given by the regression stump is $(-0.5, -0.5, 1)^T$.
- Therefore,

$$\begin{aligned}\tilde{\boldsymbol{f}}^{[1]} &= \tilde{\boldsymbol{f}}^{[0]} + 1 \cdot (-0.5, -0.5, 1)^T \\ &= (1, 1, 1)^t + (-0.5, -0.5, 1)^T \\ &= (0.5, 0.5, 2)^T\end{aligned}$$

Solution to Exercise 1 (b): Continued

(b) (iv) Derive the boosting model $\hat{f}^{[1]}(\mathbf{x})$ (i.e., $\tilde{\boldsymbol{f}}^{[1]}$).

- Recall that $R_1^{[1]} = -0.5\boldsymbol{I}_{x_1 \geq 0.5}$ and $R_2^{[1]} = 1 \cdot \boldsymbol{I}_{x_2 < 0.5}$, and learning rate $\alpha = 1$.
- So the update direction given by the regression stump is $(-0.5, -0.5, 1)^T$.
- Therefore,

$$\begin{aligned}
\tilde{\boldsymbol{f}}^{[1]} &= \tilde{\boldsymbol{f}}^{[0]} + 1 \cdot (-0.5, -0.5, 1)^T \\
&= (1, 1, 1)^t + (-0.5, -0.5, 1)^T \\
&= (0.5, 0.5, 2)^T
\end{aligned}$$

Solution to Exercise 1 (b): Continued

(b) (v) Derive the pseudo residual $\tilde{r}^{[2]}$.

- Similar as the previous step,

$$\begin{aligned}\tilde{r}^{[2]} &= \left(log_2(y^{(1)}), \log_2(y^{(2)}), \log_2(y^{(3)})\right)^T - \tilde{\boldsymbol{f}}^{[1]} \\ &= (0, 1, 2)^T - (0.5, 0.5, 2)^T \\ &= (-0.5, 0.5, 0)^T\end{aligned}$$

Exercise 1 (c)

(c) What would happend in the second iteration of the previous boosting algorithm?

Solution to Exercise 1 (c)

(c) What would happend in the second iteration of the previous boosting algorithm?

- Recall that $\tilde{r}^{[2]} = (-0.5, 0.5, 0)^T$, which needs to be fit by $R_1^{[2]}$ and $R_2^{[2]}$.
- Similar as before, we split based on $\mathbf{x}_1$, and $\mathbf{x}^{(1)}, \mathbf{x}^{(2)}$ goes to $R_1^{[2]}$, while $\mathbf{x}^{(3)}$ goes to $R_2^{[2]}$.
- Therefore $R_1^{[2]} = \frac{-0.5+0.5}{2} \cdot \boldsymbol{I}_{x_1 \geq 0.5} = 0$, and $R_2^{[2]} = 0 \cdot \boldsymbol{I}_{x_2 < 0} = 0$.
- So the update direction is $(0, 0, 0)^T$, implying that no information can be used to improve the model.

Solution to Exercise 1 (c)

(c) What would happend in the second iteration of the previous boosting algorithm?

- Recall that $\tilde{r}^{[2]} = (-0.5, 0.5, 0)^T$, which needs to be fit by $R_1^{[2]}$ and $R_2^{[2]}$.
- Similar as before, we split based on $\mathbf{x}_1$, and $\mathbf{x}^{(1)}, \mathbf{x}^{(2)}$ goes to $R_1^{[2]}$, while $\mathbf{x}^{(3)}$ goes to $R_2^{[2]}$.
- Therefore $R_1^{[2]} = \frac{-0.5+0.5}{2} \cdot \boldsymbol{I}_{x_1 \geq 0.5} = 0$, and $R_2^{[2]} = 0 \cdot \boldsymbol{I}_{x_2 < 0} = 0$.
- So the update direction is $(0, 0, 0)^T$, implying that no information can be used to improve the model.

Solution to Exercise 1 (c)

(c) What would happend in the second iteration of the previous boosting algorithm?

- Recall that $\tilde{r}^{[2]} = (-0.5, 0.5, 0)^T$, which needs to be fit by $R_1^{[2]}$ and $R_2^{[2]}$.
- Similar as before, we split based on $\mathbf{x}_1$, and $\mathbf{x}^{(1)}, \mathbf{x}^{(2)}$ goes to $R_1^{[2]}$, while $\mathbf{x}^{(3)}$ goes to $R_2^{[2]}$.
- Therefore $R_1^{[2]} = \frac{-0.5+0.5}{2} \cdot \boldsymbol{I}_{x_1 \geq 0.5} = 0$, and $R_2^{[2]} = 0 \cdot \boldsymbol{I}_{x_2 < 0} = 0$.
- So the update direction is $(0, 0, 0)^T$, implying that no information can be used to improve the model.

Solution to Exercise 1 (c)

(c) What would happend in the second iteration of the previous boosting algorithm?

- Recall that $\tilde{r}^{[2]} = (-0.5, 0.5, 0)^T$, which needs to be fit by $R_1^{[2]}$ and $R_2^{[2]}$.
- Similar as before, we split based on $\mathbf{x}_1$, and $\mathbf{x}^{(1)}, \mathbf{x}^{(2)}$ goes to $R_1^{[2]}$, while $\mathbf{x}^{(3)}$ goes to $R_2^{[2]}$.
- Therefore $R_1^{[2]} = \frac{-0.5+0.5}{2} \cdot \boldsymbol{I}_{x_1 \geq 0.5} = 0$, and $R_2^{[2]} = 0 \cdot \boldsymbol{I}_{x_2 < 0} = 0$.
- So the update direction is $(0, 0, 0)^T$, implying that no information can be used to improve the model.

Solution to Exercise 1 (c)

(c) What would happend in the second iteration of the previous boosting algorithm?

- Recall that $\tilde{r}^{[2]} = (-0.5, 0.5, 0)^T$, which needs to be fit by $R_1^{[2]}$ and $R_2^{[2]}$.
- Similar as before, we split based on $\mathbf{x}_1$, and $\mathbf{x}^{(1)}, \mathbf{x}^{(2)}$ goes to $R_1^{[2]}$, while $\mathbf{x}^{(3)}$ goes to $R_2^{[2]}$.
- Therefore $R_1^{[2]} = \frac{-0.5+0.5}{2} \cdot \boldsymbol{I}_{x_1 \geq 0.5} = 0$, and $R_2^{[2]} = 0 \cdot \boldsymbol{I}_{x_2 < 0} = 0$.
- So the update direction is $(0, 0, 0)^T$, implying that no information can be used to improve the model.

Exercise 1 (d)

(d) If you are given more data points, but still the two binary feature vectors $\mathbf{x}_1$ and $\mathbf{x}_2$, what will happen as

(i) M grows

(ii) n grows

in terms of model capacity (if d is kept fixed)?

Solution to Exercise 1 (d)

(i) M grows: capacity will increase and the algorithm may overfit.

(ii) n grows: capacity will stay the same and the algorithm may underfit.