This discussion was released **Friday, November 20**.

In this discussion section we study the decision tree-based methods. You will spend ~ 20 mins on the first problem visualizing the decision tree, random forest, and Adaboost in 2d (Use this this notebook to open the file in datahub and follow instructions therein). Then move on to the second problem to theoretically study the convergence behavior of the Adaboost algorithm.

1 Decision Boundary Visualization on Decision Tree, Random Forest, and Adaboost

In this problem, we will visualize the decision boundaries of decision tree, random forest, and adaboost with decision tree. **Please go to the Jupyter Notebook part and visualize the decision boundaries of the above approaches. You do not need to write code in the Jupyter Notebook.**

2 AdaBoost

The AdaBoost algorithm for a boosted random forest as follows: this time, we will build the trees sequentially. We will collect one sampling at a time and then we will change the weights on the data after each new tree is fit to generate more trees that focus their attention on tackling some of the more challenging data points in the training set. Suppose there are N training samples $(\mathbf{x}_i, y_i) \sim \mathcal{D}$, where $\mathbf{x}_i \in \mathcal{X}$ and $y_i \in \{+1, -1\}$. Let $\mathbf{w}^{(1)} \in \mathbb{R}^N$ denote the initial probability vector for each data point (initially, uniform), where N denotes the number of data points. More specifically,

$$\mathbf{w}^{(1)} = [w_1^{(1)}, \cdots, w_N^{(1)}]^\top = \left[\frac{1}{N}, \cdots, \frac{1}{N}\right]^\top .$$

To start off, as before, we will weight the the original training set accordingly to $\mathbf{w}$ as the training set. Fit a decision tree for this sampling, again using a randomly sampled subset of the features. Compute the weight for tree j based on its weighted accuracy:

$$a^{(t)} = \frac{1}{2} \log\left(\frac{1 - \epsilon^{(t)}}{\epsilon^{(t)}}\right) \tag{1}$$

where $\epsilon^{(t)}$ is the weighted error:

$$\epsilon^{(t)} = \frac{\sum_{i=1}^N \mathbf{1}\{h_t(\mathbf{x}_i) \neq y_i\} w_i^{(t)}}{\sum_{i=1}^N w_i^{(t)}}$$

and $\mathbf{1}\{h_t(\mathbf{x}_i) \neq y_i\}$ is an indicator for data i being *incorrectly classified by this t-th learned tree $h_t(\cdot)$*.

Then update the weights as follows:

$$w_i^{(t+1)} = \begin{cases} \frac{w_i^{(t)} \exp\left(a^{(t)}\right)}{Z^{(t)}}, & \text{if } \mathbf{1}\{h_t(\mathbf{x}_i) \neq y_i\} \\ \frac{w_i^{(t)} \exp\left(-a^{(t)}\right)}{Z^{(t)}}, & \text{otherwise.} \end{cases}$$

$Z^{(t)}$ is the normalization factor, which is defined as

$$Z^{(t)} = \sum_{i=1}^{N} w_i^{(t)} \exp\left(-a^{(t)} y_i h_t(\mathbf{x}_i)\right). \tag{2}$$

Repeat until you have M trees.

Predict by first calculating the sign of the weighted vote, $H(\mathbf{x})$, for a data sample $\mathbf{x}$:

$$H(\mathbf{x}) = \operatorname{sign}\left(\sum_{t=1}^{M} a^{(t)} h_t(\mathbf{x})\right).$$

Then, the class with the highest weighted vote is the prediction (classification result):

$$\hat{y} = H(\mathbf{x}).$$

In this problem, we will prove the training error of Adaboost algorithm will go down exponentially fast. To start with, we will make the assumption that there exist $\gamma \in (0, 1/2)$ such that

$$\epsilon^{(t)} \leq \frac{1}{2} - \gamma.$$

Our goal is to prove that the training error of H will converge to zero exponentially, i.e., there exists constant $c(\gamma) < 1$ such that

$$\widehat{\text{err}}(H) = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}\left\{H(\mathbf{x}_i) \neq y_i\right\} \leq \left[c(\gamma)\right]^M. \tag{3}$$

We will prove the above conclusion with three steps.

(a) First of all, we define $\Psi^{(t)}(\mathbf{x})$ as

$$\Psi^{(t)}(\mathbf{x}) = \sum_{j=1}^{t} a^{(j)} h_j(\mathbf{x}).$$

Then we have $H(\mathbf{x}) = \operatorname{sign}\left(\Psi^{(t)}(\mathbf{x})\right)$. **Prove that for each training data** $(\mathbf{x}_i, y_i)$:

$$\mathbf{1}\left\{H(\mathbf{x}_i) \neq y_i\right\} \leq \exp\left\{-y_i \Psi^{(M)}(\mathbf{x}_i)\right\},$$

also, **draw the 0-1 loss function and exponential function on 1d**, i.e., $\ell_{0-1}(y \cdot \hat{y}) = \operatorname{sign}(-y \cdot \hat{y})$ and $\ell_{\exp} = \exp(-y \cdot \hat{y})$.

Solution: First of all, draw the 0-1 loss function and exponential function on 1d, we can observe that the zero-one loss $\mathbf{1}\{H(\mathbf{x}_i) \neq y_i\}$ can be upper bounded by $\exp\left\{-y_i\Psi^{(M)}(\mathbf{x}_i)\right\}$, i.e.,

$$\mathbf{1}\{H(\mathbf{x}_i) \neq y_i\} \leq \exp\left\{-y_i\Psi^{(M)}(\mathbf{x}_i)\right\},$$

which is because

$$H(\mathbf{x}_i) = \text{sign}\Big(\underbrace{\sum_{t=1}^{M} a^{(t)}h_t(\mathbf{x})}_{\Psi^{(M)}(\mathbf{x}_i)}\Big) = \text{sign}\left(\Psi^{(M)}(\mathbf{x}_i)\right).$$

Then we have

$$\mathbf{1}\{H(\mathbf{x}_i) \neq y_i\} = \mathbf{1}\left\{\text{sign}\left(\Psi^{(M)}(\mathbf{x}_i)\right) \neq y_i\right\} = \mathbf{1}\left\{y_i\,\text{sign}\left(\Psi^{(M)}(\mathbf{x}_i)\right) \leq 0\right\} \leq \exp\left\{-y_i\Psi^{(M)}(\mathbf{x}_i)\right\},$$

where the last step can be verified either $\text{sign}\left(\Psi^{(M)}(\mathbf{x}_i)\right) \neq y_i$ or $\text{sign}\left(\Psi^{(M)}(\mathbf{x}_i)\right) = y_i$. This completes the proof.

(b) In this part, we first assume $\widehat{\text{err}}(H) \leq \Pi_{t=1}^{M} Z^{(t)}$ and $Z^{(t)} = 2\sqrt{\epsilon^{(t)}(1-\epsilon^{(t)})}$, then **prove there exist constant** $c(\gamma) < 1$ **such that**

$$\widehat{\text{err}}(H) \leq [c(\gamma)]^M. \tag{4}$$

Solution:

$$\begin{aligned}
\widehat{\text{err}}(H) &\leq \Pi_{t=1}^{M} Z^{(t)} \\
&= \Pi_{t=1}^{M} 2\sqrt{\epsilon^{(t)}(1-\epsilon^{(t)})} \\
&\leq \Pi_{t=1}^{M} 2\sqrt{(1/2-\gamma)(1/2+\gamma)} \\
&= \Pi_{t=1}^{M} \underbrace{\sqrt{1-4\gamma^2}}_{c(\gamma)} = [c(\gamma)]^M \\
\text{(optional)} \quad &\leq \Pi_{t=1}^{M} \exp\{-2\gamma^2\} \\
&= \exp\{-2\gamma^2 M\},
\end{aligned}$$

where the first inequality we apply the conclusion from Part (a). For the the second inequality we use the assumption that $\epsilon^{(t)} \leq 1/2 - \gamma$, and the last inequality we apply the inequality $\sqrt{1-x} \leq \exp(-x/2)$.

(c) In the last part, **prove the equation we applied in Part (b)**:

$$Z^{(t)} = 2\sqrt{\epsilon^{(t)}(1-\epsilon^{(t)})}.$$

(Hint: plug in the upper bound of $\epsilon^{(t)} \leq 1/2 - \gamma$ and then plug in the definition of $a^{(t)}$.)

Solution: According to the definition of $Z^{(t)}$ in Eq. (2),

$$Z^{(t)} = \sum_{i=1}^{N} w_i^{(t)} \exp\left\{-a^{(t)}y_ih_t(\mathbf{x}_i)\right\}$$

$$
\begin{aligned}
&= \sum_{i:\mathbf{1}\left\{h_t(\mathbf{x}_i)\neq y_i\right\}} w_i^{(t)} \exp\left\{a^{(t)}\right\} + \sum_{i:\mathbf{1}\left\{h_t(\mathbf{x}_i)= y_i\right\}} w_i^{(t)} \exp\left\{-a^{(t)}\right\} \\
&= \exp\left(a^{(t)}\right) \underbrace{\sum_{i:\mathbf{1}\left\{h_t(\mathbf{x}_i)\neq y_i\right\}} w_i^{(t)}}_{\epsilon^{(t)}} + \exp\left(-a^{(t)}\right) \underbrace{\sum_{i:\mathbf{1}\left\{h_t(\mathbf{x}_i)= y_i\right\}} w_i^{(t)}}_{1-\epsilon^{(t)}} \\
&= \exp\left(a^{(t)}\right)\epsilon^{(t)} + \exp\left(-a^{(t)}\right)(1-\epsilon^{(t)}),
\end{aligned}
$$

you could traing to minimize the above expression with respect to $a^{(t)}$, and you can find out that the optimal $a^{(t)}$ is exactly the one defined in Eq. (1)!

Plug in the $a^{(t)} = \frac{1}{2}\log\left(\frac{1-\epsilon^{(t)}}{\epsilon^{(t)}}\right)$ into the above epression, we could prove that $Z^{(t)} = 2\sqrt{\epsilon^{(t)}(1-\epsilon^{(t)})}$.

(d) In the last part, prove the equation we applied in Part (b), i.e., **first prove that the following expression holds for** $w_i^{(t+1)}$

$$
w_i^{(t+1)} = \frac{\exp\left\{-y_i \Psi^{(t)}(\mathbf{x}_i)\right\}}{N\left(\Pi_{j=1}^{t} Z^{(j)}\right)}. \tag{5}
$$

Then, **prove the training error of H can be upper bounded in terms of $Z^{(j)}$**:

$$
\widehat{\text{err}}(H) \leq \Pi_{t=1}^{M} Z^{(t)}.
$$

Solution: By definition, we could rewrote $w_i^{(t+1)}$ as

$$
\begin{aligned}
w_i^{(t+1)} &= w_i^{(t)} \frac{\exp\left\{-y_i a^{(t)} h_t(\mathbf{x}_i)\right\}}{Z^{(t)}} \\
&= \underbrace{w_i^{(1)}}_{1/N} \frac{\exp\left\{-y_i a^{(t)} h_t(\mathbf{x}_i)\right\}}{Z^{(t)}} \cdots \frac{\exp\left\{-y_i a^{(1)} h_1(\mathbf{x}_i)\right\}}{Z^{(1)}} \\
&= \frac{1}{N} \frac{\exp\left\{-y_i \sum_{j=1}^{t} a^{(j)} h_j(\mathbf{x}_i)\right\}}{\Pi_{j=1}^{t} Z^{(j)}} \\
&= \frac{\exp\left\{-y_i \Psi^{(t)}(\mathbf{x}_i)\right\}}{N\left(\Pi_{j=1}^{t} Z^{(j)}\right)},
\end{aligned}
$$

which completes the proof for the first part.

Then together with the conclusion in Part (a), we could upper bound $\widehat{\text{err}}(H)$ as

$$
\begin{aligned}
\widehat{\text{err}}(H) &= \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}\left\{H(\mathbf{x}_i) \neq y_i\right\} \\
&\leq \frac{1}{N}\sum_{i=1}^{N} \exp\left\{-y_i \Psi^{(M)}(\mathbf{x}_i)\right\}
\end{aligned}
$$

$$
\begin{aligned}
&= \frac{1}{N}\sum_{i=1}^{N} w_i^{(t+1)} N \left(\Pi_{t=1}^{M} Z^{(t)}\right) \\
&= \left(\Pi_{t=1}^{M} Z^{(t)}\right) \underbrace{\sum_{i=1}^{N} w_i^{(t+1)}}_{=1} \\
&= \Pi_{t=1}^{M} Z^{(t)},
\end{aligned}
$$

which completes the proof.

Contributors:

- Anant Sahai
- Yaodong Yu