Problem Set 4*

Xiaowei Zhang†

Problem 1

(1) Emperical zero-one error $\hat{\epsilon}_{D_t}(h_t)$, by definition, $= \sum_{i=1}^{N} D_t(i)[[h_t(x_i) \neq y_i]]$, here [[A]] is the indicator function.

$$\begin{aligned}
\hat{\epsilon}_{D_t}(h_t) &= \sum_{i=1}^{N} D_t(i)[[h_t(x_i) \neq y_i]] \\
&= \sum_{i=1}^{N} \frac{D_{t-1}(i)\exp(-\alpha_t y_i h_t(x_i))}{\sum_{j=1}^{N} D_{t-1}(j)\exp(-\alpha_t y_j h_t(x_j))}[[h_t(x_i) \neq y_i]] \\
&= \frac{\sum_{i=1}^{N}[[h_t(x_i) \neq y_i]] D_{t-1}(i)\exp(-\alpha_t y_i h_t(x_i))}{\sum_{j=1}^{N} D_{t-1}(j)\exp(-\alpha_t y_j h_t(x_j))} \\
&= \frac{\hat{\epsilon}_{D_{t-1}}(h_t)\exp(\alpha_t)}{\sum_{j=1}^{N} D_{t-1}(j)\exp(-\alpha_t y_j h_t(x_j))} \\
&= \frac{\hat{\epsilon}_{D_{t-1}}(h_t)\exp(\alpha_t)}{\sum_{j:h_t(x_j)\neq y_j} D_{t-1}(j)\exp(\alpha_t) + \sum_{j:h_t(x_j)=y_j} D_{t-1}(j)\exp(-\alpha_t)} \\
&= \frac{\hat{\epsilon}_{D_{t-1}}(h_t)\exp(\alpha_t)}{\hat{\epsilon}_{D_{t-1}}(h_t)\exp(\alpha_t) + \left(1-\hat{\epsilon}_{D_{t-1}}(h_t)\right)\exp(-\alpha_t)} = \frac{1}{2}
\end{aligned}$$

We know that $y_i h_t(x_i) = -1$ when $y_i \neq h_t(x_i)$ because both y_i and $h_t(x_i)$ has two possible values: 1 or -1. When they are different, the product should be -1. When they are the same, the product should be 1. Therefore, $\exp(-\alpha_t y_i h_t(x_i))$ becomes $\exp(\alpha_t)$ when $y_i \neq h_t(x_i)$ and becomes $\exp(-\alpha_t)$ when $y_i = h_t(x_i)$.

We also know that $\sum_{j:h_t(x_j)=y_j} D_{t-1}(j) = \sum_{j=1}^{N} D_{t-1}(j)[[h_t(x_j) = y_j]] = 1 - \sum_{j=1}^{N} D_{t-1}(j)[[h_t(x_j) \neq y_j]] = 1 - \hat{\epsilon}_{D_{t-1}}(h_t)$ because the sum of a distribution should be 1

*Due: `Nov 23, 2021`, Student(s) worked with:

†NetID: `xz561`, Email: `xz561@scarletmail.rutgers.edu`

(2)

$$\frac{\hat{\epsilon}_{D_{t-1}}(h_t)\exp(\alpha_t)}{\hat{\epsilon}_{D_{t-1}}(h_t)\exp(\alpha_t)+\big(1-\hat{\epsilon}_{D_{t-1}}(h_t)\big)\exp(-\alpha_t)} = \frac{1}{2}$$
$$2\hat{\epsilon}_{D_{t-1}}(h_t)\exp(\alpha_t) = \hat{\epsilon}_{D_{t-1}}(h_t)\exp(\alpha_t) + \big(1-\hat{\epsilon}_{D_{t-1}}(h_t)\big)\exp(-\alpha_t)$$
$$\hat{\epsilon}_{D_{t-1}}(h_t)\exp(\alpha_t) = \big(1-\hat{\epsilon}_{D_{t-1}}(h_t)\big)\exp(-\alpha_t)$$
$$\exp(\alpha_t+\alpha_t) = \frac{1-\hat{\epsilon}_{D_{t-1}}(h_t)}{\hat{\epsilon}_{D_{t-1}}(h_t)}$$
$$2\alpha_t = \log\Big(\frac{1-\hat{\epsilon}_{D_{t-1}}(h_t)}{\hat{\epsilon}_{D_{t-1}}(h_t)}\Big)$$
$$\alpha_t = \frac{1}{2}\log\Big(\frac{1-\hat{\epsilon}_{D_{t-1}}(h_t)}{\hat{\epsilon}_{D_{t-1}}(h_t)}\Big)$$

We obtained expression (2). We see indeed that α_t results in $\hat{\epsilon}_{D_t}(h_t) = 1/2$

Problem 2

It is not possible to have $h_{t+1} = h_t$ given we assume that the base classifiers in $\mathcal{H}$ are neither too weak nor too strong. From previous problem, we discovered that adaboost will always choose α_t such that the updated data distribution yields the worst possible weighted loss for h_t, which is 1/2. By letting $h_{t+1} = h_t$, we are saying that the best base classifiers we can find in $\mathcal{H}$ to minimize $\hat{\epsilon}_{D_t}(h)$ is h such that the $\hat{\epsilon}_{D_t}(h) = 1/2$. This conclusion contradicts with our assumption that there is always some h in $\mathcal{H}$ such that $\hat{\epsilon}_D(h) < 1/2$. Therefore, we can never let $h_{t+1} = h_t$ for any t.

Problem 3

It is possible if we assume the number of base classifiers in $\mathcal{H}$ is finite and if T, the number of rounds, is larger than the number of base classifiers. If that is the case, it is deemed that there is some t and $n > 1$ such that $h_{t+n} = h_t$.

Formally, suppose the size of $\mathcal{H}$ is Q and the number of rounds of adaboost is T and $Q < T$. By pigeonhole principle, there is at least one base classifier h will be picked more than once.

Problem 4

$$\begin{aligned}
\hat{l}_{D_{t-1}}(h_t) &= \sum_{i=1}^{N} D_{t-1}(i)\exp(-\alpha_t y_i h_t(x_i)) \\
&= \sum_{i:y_i=h_t(x_i)}^{N} D_{t-1}(i)\exp(-\alpha_t) + \sum_{i:y_i\neq h_t(x_i)}^{N} D_{t-1}(i)\exp(\alpha_t) \\
&= \sum_{i=1}^{N} D_{t-1}(i)[[h_t(x_i)\neq y_i]]\exp(\alpha_t) + \sum_{i=1}^{N} D_{t-1}(i)[[h_t(x_i)=y_i]]\exp(-\alpha_t) \\
&= \hat{\epsilon}_{D_{t-1}}(h_t)\exp(\alpha_t) + \big(1-\hat{\epsilon}_{D_{t-1}}(h_t)\big)\exp(-\alpha_t)
\end{aligned}$$

Assuming h_t is selected: To minimize $\hat{l}_{D_{t-1}}(h_t)$ We take partial derivative of $\hat{l}_{D_{t-1}}(h_t)$ with respect to α_t.

$$\begin{aligned}\frac{\partial \hat{l}_{D_{t-1}}(h_t)}{\partial \alpha_t} &= \frac{\partial\big(\hat{\epsilon}_{D_{t-1}}(h_t)\exp(\alpha_t) + \big(1-\hat{\epsilon}_{D_{t-1}}(h_t)\big)\exp(-\alpha_t)\big)}{\partial \alpha_t} \\ &= \hat{\epsilon}_{D_{t-1}}(h_t)\exp(\alpha_t) - \big(\big(1-\hat{\epsilon}_{D_{t-1}}(h_t)\big)\exp(-\alpha_t)\big)\end{aligned}$$

We want minimize this, so set it to 0 .

$$\begin{aligned}0 &= \hat{\epsilon}_{D_{t-1}}(h_t)\exp(\alpha_t) - \big(\big(1-\hat{\epsilon}_{D_{t-1}}(h_t)\big)\exp(-\alpha_t)\big) \\ \hat{\epsilon}_{D_{t-1}}(h_t)\exp(\alpha_t) &= \big(1-\hat{\epsilon}_{D_{t-1}}(h_t)\big)\exp(-\alpha_t) \\ \frac{\hat{\epsilon}_{D_{t-1}}(h_t)}{1-\hat{\epsilon}_{D_{t-1}}(h_t)} &= \exp(-\alpha_t - \alpha_t) \\ \alpha_t &= -\frac{1}{2}\log\left(\frac{\hat{\epsilon}_{D_{t-1}}(h_t)}{1-\hat{\epsilon}_{D_{t-1}}(h_t)}\right) \\ \alpha_t &= \frac{1}{2}\log\left(\frac{1-\hat{\epsilon}_{D_{t-1}}(h_t)}{\hat{\epsilon}_{D_{t-1}}(h_t)}\right)\end{aligned}$$

Indecd, the choice of α_t in (2) minimizes $\hat{l}_{D_{t-1}}(h_t)$