Homework - Boosting

Jianxiang Fan
jianxiang.fan@colorado.edu

1.

(a) Final distribution

$$D_{T+1}(i) = \frac{1}{m}\frac{\exp(-y_i f(x_i))}{\prod\limits_t Z_t}$$

where $f(x) = \sum\limits_t \alpha_t h_t(x)$

Compute empirical error $\hat{R}(h)$:

$$\begin{aligned}
\hat{R}(h) &= \frac{1}{m}\sum_i \mathbf{1}(y_i \neq f(x_i)) \\
&= \frac{1}{m}\sum_i \mathbf{1}(y_i f(x_i) < 0) \\
&\leq \frac{1}{m}\sum_i \exp(-y_i f(x_i)) \\
&= \sum_i D_{T+1}(i)\prod_t Z_t \\
&= \prod_t((1-\epsilon_t)e^{-\alpha} + \epsilon_t e^{\alpha}) \\
&= \prod_t(\epsilon_t(e^{\alpha} - e^{-\alpha}) + e^{-\alpha}) \\
&\leq ((\frac{1}{2}-\gamma)(e^{\alpha} - e^{-\alpha}) + e^{-\alpha})^T \\
&= ((\frac{1}{2}-\gamma)e^{\alpha} + (\frac{1}{2}+\gamma)e^{-\alpha})^T
\end{aligned} \tag{1}$$

Define $g(\alpha) = (1/2-\gamma)e^{\alpha} + (1/2+\gamma)e^{-\alpha}$, then $g'(\alpha) = (1/2-\gamma)e^{\alpha} - (1/2+\gamma)e^{-\alpha}$. Let it equal to 0, then

$$(\frac{1}{2}-\gamma)e^{\alpha} = (\frac{1}{2}+\gamma)e^{-\alpha}$$
$$\alpha = \frac{1}{2}\ln\frac{1+2\gamma}{1-2\gamma}$$

(b) At round t,

$$\begin{aligned}
p(wrong) - p(right) &= \epsilon_t e^{\alpha} - (1-\epsilon_t)e^{-\alpha} \\
&\leq (\frac{1}{2}-\gamma)\sqrt{\frac{1+2\gamma}{1-2\gamma}} - (\frac{1}{2}+\gamma)\sqrt{\frac{1-2\gamma}{1+2\gamma}} \\
&= 0
\end{aligned} \tag{2}$$

That is $p(right) \geq p(wrong)$.

(c) From question (a), we have already got:

$$\begin{aligned}
\hat{R}(h) &\leq ((\frac{1}{2}-\gamma)e^{\alpha} + (\frac{1}{2}+\gamma)e^{-\alpha})^T \\
&= ((\frac{1}{2}-\gamma)\sqrt{\frac{1+2\gamma}{1-2\gamma}} + (\frac{1}{2}+\gamma)\sqrt{\frac{1-2\gamma}{1+2\gamma}})^T \\
&= (1-4\gamma^2)^{T/2} \\
&\leq \exp(-2\gamma^2 T)
\end{aligned} \tag{3}$$

2. Suppose we select h_t for the distribution D_{t+1}, let's compute its empirical error:

$$\begin{aligned}
\hat{R}_{D_{t+1}}(h_t) &= \frac{\sum\limits_{i:y_i h_t(x_i)<0} D_t(i)e^{\alpha_t}}{Z_t} \\
&= \frac{\epsilon_t e^{\alpha_t}}{Z_t} \\
&= \frac{\epsilon_t\sqrt{\frac{1-\epsilon_t}{\epsilon_t}}}{2\sqrt{(1-\epsilon_t)\epsilon_t}} = \frac{1}{2}
\end{aligned} \tag{4}$$

This contradict with the weak learning assumption, so h_{t+1} must must be different from h_t.

3. In each round, we can use x-axis or y-axis as our weak classifier. For example, in the first round, we can choose the hypothesis that label all the points left to the y-axis as postive, right ones as negative, then the training error $\epsilon_1 = 1/4 + (1-\epsilon)/4 = 1/2 - \epsilon/4$. After making new distriduton, in the second round, we will choose label all the points above the x-axis as postive, beneath ones as negative.
As we show in last problem, in consecutive rounds, we cannot choose same hypothesis h_t. In fact, we cannot choose the opposite hypothesis $-h_t$ either, because:

$$\begin{aligned}
\hat{R}_{D_{t+1}}(-h_t) &= \frac{\sum\limits_{i:y_i h_t(x_i)>0} D_t(i)e^{-\alpha_t}}{Z_t} \\
&= \frac{(1-\epsilon_t)e^{-\alpha_t}}{Z_t} \\
&= \frac{(1-\epsilon_t)\sqrt{\frac{\epsilon_t}{1-\epsilon_t}}}{2\sqrt{(1-\epsilon_t)\epsilon_t}} = \frac{1}{2}
\end{aligned} \tag{5}$$

So at each round, we will alternately choose x-axis and y-axis as our classifier. Using these hypotheses, the points at $(1,-1)$ are always misclassified, so $\hat{R}(H_{final}) = (1-\epsilon)/4$