Introduction to Machine Learning, Fall 2023
Homework 3
(Due Tuesday Nov. 30 at 11:59pm (CST))

November 26, 2023

1. [15 points] [Expectation Maximization Algorithm] Consider a probabilistic model in which we collectively denote the observed variables by $\boldsymbol{X}$ and all of the hidden variables by $\boldsymbol{Z}$. The joint distribution $p(\boldsymbol{X}, \boldsymbol{Z}|\theta)$ is parameterized by θ. Our goal is to maximize the likelihood function given by

$$p(\boldsymbol{X}|\theta). \tag{1}$$

(a) Given an arbitrary distribution q, show that the log-likelihood of $\boldsymbol{X}$ is [5 points]

$$\log p(\boldsymbol{X}|\theta) = \mathbb{E}_{\boldsymbol{Z}\sim q}\left[\log \frac{p(\boldsymbol{X}, \boldsymbol{Z}|\theta)}{q(\boldsymbol{Z})}\right] + KL\left(q(\boldsymbol{Z})\|p(\boldsymbol{Z}|\boldsymbol{X}, \theta)\right). \tag{2}$$

(b) Next let's consider the expectation step. First show the evidence lower bound (ELBO) is a lower bound of the log-likelihood, namely [5 points]

$$\log p(\boldsymbol{X}|\theta) \geq \mathbb{E}_{\boldsymbol{Z}|\boldsymbol{X},\theta^{(t-1)}}\left[\log \frac{p(\boldsymbol{X}, \boldsymbol{Z}|\theta)}{p(\boldsymbol{Z}|\boldsymbol{X}, \theta^{(t-1)})}\right], \tag{3}$$

where $\theta^{(t-1)}$ is the parameter estimated in the previous iteration.

(c) We want to maximize the ELBO, $\mathbb{E}_{\boldsymbol{Z}|\boldsymbol{X},\theta^{(t-1)}}\left[\log \frac{p(\boldsymbol{X},\boldsymbol{Z}|\theta)}{p(\boldsymbol{Z}|\boldsymbol{X},\theta^{(t-1)})}\right]$ since maximizing $p(\boldsymbol{X}|\theta)$ is hard. EM algorithm defines $Q(\theta|\theta^{(t-1)}) := \mathbb{E}_{\boldsymbol{Z}|\boldsymbol{X},\theta^{(t-1)}}\left[\log p(\boldsymbol{X}, \boldsymbol{Z}|\theta)\right]$. The M-step is given by:

$$\theta^{(t)} \leftarrow \arg\max_{\theta} Q(\theta|\theta^{(t-1)}). \tag{4}$$

Show that maximizing $Q(\theta|\theta^{(t-1)})$ and maximizing the ELBO is equivalent. [5 points] Formally,

$$\arg\max_{\theta} Q(\theta|\theta^{(t-1)}) = \arg\max_{\theta} \mathbb{E}_{\boldsymbol{Z}|\boldsymbol{X},\theta^{(t-1)}}\left[\log \frac{p(\boldsymbol{X}, \boldsymbol{Z}|\theta)}{p(\boldsymbol{Z}|\boldsymbol{X}, \theta^{(t-1)})}\right] \tag{5}$$

Solution:

(a) With Bayes' Rule, we can get that

$$p(\boldsymbol{X}|\theta) = \frac{p(\boldsymbol{X}, \boldsymbol{Z}|\theta)}{p(\boldsymbol{Z}|\boldsymbol{X}, \theta)} = \frac{p(\boldsymbol{X}, \boldsymbol{Z}|\theta)}{q(\boldsymbol{Z})}\frac{q(\boldsymbol{Z})}{p(\boldsymbol{Z}|\boldsymbol{X}, \theta)}$$

So the log-likelihood of $\boldsymbol{X}$ is

$$\log p(\boldsymbol{X}|\theta) = \log\left[\frac{p(\boldsymbol{X}, \boldsymbol{Z}|\theta)}{q(\boldsymbol{Z})}\frac{q(\boldsymbol{Z})}{p(\boldsymbol{Z}|\boldsymbol{X}, \theta)}\right] = \log \frac{p(\boldsymbol{X}, \boldsymbol{Z}|\theta)}{q(\boldsymbol{Z})} + \log \frac{q(\boldsymbol{Z})}{p(\boldsymbol{Z}|\boldsymbol{X}, \theta)}$$

Take the expectation of $\boldsymbol{Z}$ with respect to $q(\boldsymbol{Z})$ to the both side, we can get that

$$\mathbb{E}_{\boldsymbol{Z}\sim q}\left[\log p(\boldsymbol{X}|\theta)\right] = \mathbb{E}_{\boldsymbol{Z}\sim q}\left[\log \frac{p(\boldsymbol{X}, \boldsymbol{Z}|\theta)}{q(\boldsymbol{Z})} + \log \frac{q(\boldsymbol{Z})}{p(\boldsymbol{Z}|\boldsymbol{X}, \theta)}\right]$$

With the linearty of expectation:

$$\mathbb{E}_{\boldsymbol{Z}\sim q}\left[\log p(\boldsymbol{X}|\theta)\right] = \mathbb{E}_{\boldsymbol{Z}\sim q}\left[\log \frac{p(\boldsymbol{X}, \boldsymbol{Z}|\theta)}{q(\boldsymbol{Z})}\right] + \mathbb{E}_{\boldsymbol{Z}\sim q}\left[\log \frac{q(\boldsymbol{Z})}{p(\boldsymbol{Z}|\boldsymbol{X}, \theta)}\right]$$

For $\mathbb{E}_{\boldsymbol{Z}\sim q}[\log p(\boldsymbol{X}|\theta)]$, we can get that it has nothing with $\boldsymbol{Z}$, so

$$\mathbb{E}_{\boldsymbol{Z}\sim q}[\log p(\boldsymbol{X}|\theta)] = \int q(\boldsymbol{z}) \log p(\boldsymbol{X}|\theta) d\boldsymbol{z} = \log p(\boldsymbol{X}|\theta) \int q(\boldsymbol{z}) d\boldsymbol{z} = \log p(\boldsymbol{X}|\theta)$$

For $\mathbb{E}_{\boldsymbol{Z}\sim q}\left[\log \frac{q(\boldsymbol{Z})}{p(\boldsymbol{Z}|\boldsymbol{X},\theta)}\right]$, accoding to the definition of KL divergence : $KL(p\|q) = \int p(\boldsymbol{z}) \log \frac{p(\boldsymbol{z})}{q(\boldsymbol{z})} d\boldsymbol{z}$, we can get that

$$\mathbb{E}_{\boldsymbol{Z}\sim q}\left[\log \frac{q(\boldsymbol{Z})}{p(\boldsymbol{Z}|\boldsymbol{X},\theta)}\right] = \int q(\boldsymbol{z}) \log \frac{q(\boldsymbol{z})}{p(\boldsymbol{z}|\boldsymbol{X},\theta)} d\boldsymbol{z} = KL(q(\boldsymbol{Z})\|p(\boldsymbol{Z}|\boldsymbol{X},\theta))$$

So above all, we have proved that

$$\log p(\boldsymbol{X}|\theta) = \mathbb{E}_{\boldsymbol{Z}\sim q}\left[\log \frac{p(\boldsymbol{X},\boldsymbol{Z}|\theta)}{q(\boldsymbol{Z})}\right] + KL(q(\boldsymbol{Z})\|p(\boldsymbol{Z}|\boldsymbol{X},\theta))$$

(b) For the log-likelihood:

$$\begin{aligned}
\log p(\boldsymbol{X}|\theta) &= \log \int p(\boldsymbol{X},\boldsymbol{z}|\theta) d\boldsymbol{z} \\
&= \log \int p(\boldsymbol{z}|\boldsymbol{X},\theta^{(t-1)}) \frac{p(\boldsymbol{X},\boldsymbol{z}|\theta)}{p(\boldsymbol{z}|\boldsymbol{X},\theta^{(t-1)})} d\boldsymbol{z} \\
&= \log \mathbb{E}_{\boldsymbol{Z}|\boldsymbol{X},\theta^{(t-1)}}\left[\frac{p(\boldsymbol{X},\boldsymbol{Z}|\theta)}{p(\boldsymbol{Z}|\boldsymbol{X},\theta^{(t-1)})}\right]
\end{aligned}$$

Since log is a concave function, with Jensen's inequality, we have $\log \mathbb{E}(X) \geq \mathbb{E}(\log X)$, so

$$\log \mathbb{E}_{\boldsymbol{Z}|\boldsymbol{X},\theta^{(t-1)}}\left[\frac{p(\boldsymbol{X},\boldsymbol{Z}|\theta)}{p(\boldsymbol{Z}|\boldsymbol{X},\theta^{(t-1)})}\right] \geq \mathbb{E}_{\boldsymbol{Z}|\boldsymbol{X},\theta^{(t-1)}}\left[\log \frac{p(\boldsymbol{X},\boldsymbol{Z}|\theta)}{p(\boldsymbol{Z}|\boldsymbol{X},\theta^{(t-1)})}\right]$$

So above all, we have proved that the ELBO is that

$$\log p(\boldsymbol{X}|\theta) \geq \mathbb{E}_{\boldsymbol{Z}|\boldsymbol{X},\theta^{(t-1)}}\left[\log \frac{p(\boldsymbol{X},\boldsymbol{Z}|\theta)}{p(\boldsymbol{Z}|\boldsymbol{X},\theta^{(t-1)})}\right]$$

(c)

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{Z}|\boldsymbol{X},\theta^{(t-1)}}\left[\log \frac{p(\boldsymbol{X},\boldsymbol{Z}|\theta)}{p(\boldsymbol{Z}|\boldsymbol{X},\theta^{(t-1)})}\right] &= \int p(\boldsymbol{z}|\boldsymbol{X},\theta^{(t-1)}) \log \frac{p(\boldsymbol{X},\boldsymbol{z}|\theta)}{p(\boldsymbol{z}|\boldsymbol{X},\theta^{(t-1)})} d\boldsymbol{z} \\
&= \int p(\boldsymbol{z}|\boldsymbol{X},\theta^{(t-1)})(\log p(\boldsymbol{X},\boldsymbol{z}|\theta) - \log p(\boldsymbol{z}|\boldsymbol{X},\theta^{(t-1)})) d\boldsymbol{z} \\
= \int p(\boldsymbol{z}|\boldsymbol{X},\theta^{(t-1)}) \log p(\boldsymbol{X},\boldsymbol{z}|\theta) d\boldsymbol{z} &- \int p(\boldsymbol{z}|\boldsymbol{X},\theta^{(t-1)}) \log p(\boldsymbol{z}|\boldsymbol{X},\theta^{(t-1)}) d\boldsymbol{z}
\end{aligned}$$

Since $Q(\theta|\theta^{(t-1)}) := \mathbb{E}_{\boldsymbol{Z}|\boldsymbol{X},\theta^{(t-1)}}[\log p(\boldsymbol{X},\boldsymbol{Z}|\theta)]$.
So we have

$$\int p(\boldsymbol{z}|\boldsymbol{X},\theta^{(t-1)}) \log p(\boldsymbol{X},\boldsymbol{z}|\theta) d\boldsymbol{z} = \mathbb{E}_{\boldsymbol{Z}|\boldsymbol{X},\theta^{(t-1)}}[\log p(\boldsymbol{X},\boldsymbol{Z}|\theta)] = Q(\theta|\theta^{(t-1)})$$

And with the definition of entropy: $H(\boldsymbol{X}) = -\int p(\boldsymbol{x}) \log p(\boldsymbol{x}) d\boldsymbol{x}$, we can get that

$$-\int p(\boldsymbol{z}|\boldsymbol{X},\theta^{(t-1)}) \log p(\boldsymbol{z}|\boldsymbol{X},\theta^{(t-1)}) d\boldsymbol{z} = H(\boldsymbol{Z}|\boldsymbol{X},\theta^{(t-1)})$$

So

$$\begin{aligned}
\int p(\boldsymbol{z}|\boldsymbol{X},\theta^{(t-1)}) \log p(\boldsymbol{X},\boldsymbol{z}|\theta) d\boldsymbol{z} &- \int p(\boldsymbol{z}|\boldsymbol{X},\theta^{(t-1)}) \log p(\boldsymbol{z}|\boldsymbol{X},\theta^{(t-1)}) d\boldsymbol{z} \\
&= Q(\theta|\theta^{(t-1)}) + H(\boldsymbol{Z}|\boldsymbol{X},\theta^{(t-1)})
\end{aligned}$$

Since $H(\boldsymbol{Z}|\boldsymbol{X},\theta^{(t-1)})$ is a constant of θ, so we can get that

$$\operatorname*{argmax}_{\theta} \mathbb{E}_{\boldsymbol{Z}|\boldsymbol{X},\theta^{(t-1)}}\left[\log \frac{p(\boldsymbol{X},\boldsymbol{Z}|\theta)}{p(\boldsymbol{Z}|\boldsymbol{X},\theta^{(t-1)})}\right] = \operatorname*{argmax}_{\theta} Q(\theta|\theta^{(t-1)}) + H(\boldsymbol{Z}|\boldsymbol{X},\theta^{(t-1)}) = \operatorname*{argmax}_{\theta} Q(\theta|\theta^{(t-1)})$$

So above all, we have proved that

$$\operatorname*{argmax}_{\theta} Q(\theta|\theta^{(t-1)}) = \operatorname*{argmax}_{\theta} \mathbb{E}_{\boldsymbol{Z}|\boldsymbol{X},\theta^{(t-1)}}\left[\log \frac{p(\boldsymbol{X},\boldsymbol{Z}|\theta)}{p(\boldsymbol{Z}|\boldsymbol{X},\theta^{(t-1)})}\right]$$

Table 1: The training data in (a).

i	x_{i1}	x_{i2}	y_i
1	1.5	0.5	1
2	2.5	1.5	1
3	3.5	3.5	1
4	6.5	5.5	1
5	7.5	10.5	1
6	1.5	2.5	-1
7	3.5	1.5	-1
8	5.5	5.5	-1
9	7.5	8.5	-1
10	1.5	10.5	-1

2. [15 points] [Boosting] Suppose that we are interested in learning a classifier, such that at any turn of a game we can pose a question, like "should I attack this ant hill now?", and get an answer.That is, we want to build a classifier which we can feed some features on the current game state, and get the output "attack" or "don't attack". There are many possible ways to define what the action "attack" means, but for now let's define it as sending all friendly ants that can see the ant hill under consideration towards it.

   Let's recall the AdaBoost algorithm described in class. Its input is a dataset $\{(x_i, y_i)\}_{i=1}^n$, with x_i being the i-th sample, and $y_i \in \{-1, 1\}$ denoting the i-th label, $i = 1, 2, ..., n$. The features might be composed of a count of the number of friendly ants that can see the ant hill under consideration, and a count of the number of enemy ants these friendly ants can see. For example, if there were 10 friendly ants that could see a particular ant hill, and 5 enemy ants that the friendly ants could see, we would have:

$$x_1 = \begin{bmatrix} 10 \\ 5 \end{bmatrix}.$$

   The label of the example x_1 is $y_1 = 1$, once the friendly ants were successful in razing the enemy ant hill, and $y_1 = 0$ otherwise. We could generate such examples by running a greedy bot (or any other opponent bot) against a bot that we periodically try to attack an enemy ant hill. Each time this bot tries the attack, we record (say, after 20 turns or some other significant amount of time) whether the attack was successful or not.

   (a) Let ϵ_t denote the error of a weak classifier h_t:

$$\epsilon_t = \sum_{i=1}^{n} D_t(i) \mathbb{1}(y_i \neq h_t(x_i)). \tag{6}$$

   In the simple "attack" / "don't attack" scenario, suppose that we have implemented the following six weak classifiers:

$$h^{(1)}(x_i) = 2 * \mathbb{1}(x_{i1} \geq 2) - 1, \qquad h^{(4)}(x_i) = 2 * \mathbb{1}(x_{i2} \leq 2) - 1,$$
$$h^{(2)}(x_i) = 2 * \mathbb{1}(x_{i1} \geq 6) - 1, \qquad h^{(5)}(x_i) = 2 * \mathbb{1}(x_{i2} \leq 6) - 1,$$
$$h^{(3)}(x_i) = 2 * \mathbb{1}(x_{i1} \geq 10) - 1, \qquad h^{(6)}(x_i) = 2 * \mathbb{1}(x_{i2} \leq 10) - 1.$$

   Given ten training data points ($n = 10$) as shown in Table 1, please show that what is the minimum value of ϵ_1 and which of $h^{(1)}, ..., h^{(6)}$ achieve this value? Note that there may be multiple classifiers that all have the same ϵ_1. You should list all classifiers that achieve the minimum ϵ_1 value. [3 points]

   (b) For all the questions in the remainder of this section, let h_1 denote $h^{(1)}$ chosen in the first round of boosting. (That is, $h^{(1)}$ was the classifier that achieved the minimum ϵ_1.)

   (1) What is the value of α_1 (the weight of this first classifier h_1)? [1 points]

   (2) What should Z_t be in order to make sure the distribution D_{t+1} is normalized correctly? That is, derive the formula of Z_t in terms of ϵ_t that will ensure $\sum_{i=1}^n D_{t+1}(i) = 1$. Please also derive the formula of α_t in terms of ϵ_t. [3 points]

(3) Which points will increase in significance in the second round of boosting? That is, for which points will we have $D_1(i) < D_2(i)$? What are the values of D_2 for these points? [3 points]

(4) In the second round of boosting, the weights on the points will be different, and thus the error ϵ_2 will also be different. Which of $h^{(1)}, ..., h^{(6)}$ will minimize ϵ_2? (Which classifier will be selected as the second weak classifier h_2?) What is its value of ϵ_2? [3 points]

(5) What will the average error of the final classifier H be, if we stop after these two rounds of boosting? That is, if $H(x) = \text{sign}(\alpha_1 h_1(x) + \alpha_2 h_2(x))$, what will the training error $\epsilon = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}(y_i \neq H(x_i))$ be? Is this more, less, or the same as the error we would get, if we just used one of the weak classifiers instead of this final classifier H? [2 points]

Solution:

(a) Since D_1 is uniform on the training data, so we have $D_1(i) = \frac{1}{10}$ for $i = 1, 2, \cdots, 10$.
So for each classifer $h^{(j)}$, we can get the error $(\epsilon_1)_j$ is

$$(\epsilon_1)_j = \mathbb{E}_{D_1}[\mathbb{1}(y_i \neq h^{(j)}(x_i))] = \sum_{i=1}^{n} D_1(i)\mathbb{1}(y_i \neq h^{(j)}(x_i)) = \frac{1}{10}\sum_{i=1}^{n} \mathbb{1}(y_i \neq h^{(j)}(x_i))$$

- For $h^{(1)}$, we can get that the data x_1, x_7, x_8, x_9 are misclassified, so we have $(\epsilon_1)_1 = \frac{1}{10} \cdot 4 = 0.4$
- For $h^{(2)}$, we can get that the data x_1, x_2, x_3, x_9 are misclassified, so we have $(\epsilon_1)_2 = \frac{1}{10} \cdot 4 = 0.4$
- For $h^{(3)}$, we can get that the data x_1, x_2, x_3, x_4, x_5 are misclassified, so we have $(\epsilon_1)_3 = \frac{1}{10} \cdot 5 = 0.5$
- For $h^{(4)}$, we can get that the data x_3, x_4, x_5, x_7 are misclassified, so we have $(\epsilon_1)_4 = \frac{1}{10} \cdot 4 = 0.4$
- For $h^{(5)}$, we can get that the data x_5, x_6, x_7, x_8 are misclassified, so we have $(\epsilon_1)_5 = \frac{1}{10} \cdot 4 = 0.4$
- For $h^{(6)}$, we can get that the data x_5, x_6, x_7, x_8, x_9 are misclassified, so we have $(\epsilon_1)_6 = \frac{1}{10} \cdot 5 = 0.5$

So above all, the minimum value of ϵ_1 is 0.4, and the classifiers $h^{(1)}, h^{(2)}, h^{(4)}, h^{(5)}$ achieve this value.

(b)
(1) From (a), we can get that $\epsilon_1 = 0.4$.
So $\alpha_1 = \frac{1}{2}\log\frac{1-\epsilon_1}{\epsilon_1} = \frac{1}{2}\log\frac{1-0.4}{0.4} = \frac{1}{2}\log\frac{3}{2}$.

So above all, $\alpha_1 = \frac{1}{2}\log\frac{3}{2}$.

(2) 1. To make sure the distribution D_{t+1} is normalized correctly, we should make sure $\sum_{i=1}^{n} D_{t+1}(i) = 1$.

Since $D_{t+1}(i) = \frac{D_t(i)}{Z_t}\exp\left(-\alpha_t y_i h_t(x_i)\right)$, so we have

$$\sum_{i=1}^{n} D_{t+1}(i) = \sum_{i=1}^{n} \frac{D_t(i)}{Z_t}\exp\left(-\alpha_t y_i h_t(x_i)\right) = \frac{1}{Z_t}\sum_{i=1}^{n} D_t(i)\exp(-\alpha_t y_i h_t(x_i)) = 1$$

So we have

$$\begin{aligned} Z_t &= \sum_{i=1}^{n} D_t(i)\exp\left(-\alpha_t y_i h_t(x_i)\right) \\ &= \sum_{i:y_i \neq h_t(x_i)} D_t(i)e^{\alpha_t} + \sum_{i:y_i = h_t(x_i)} D_t(i)e^{-\alpha_t} \\ &= e^{\alpha_t}\sum_{i=1}^{n} D_t(i)\mathbb{1}(y_i \neq h_t(x_i)) + e^{-\alpha_t}\sum_{i=1}^{n} D_t(i)\mathbb{1}(y_i = h_t(x_i)) \\ &= e^{\alpha_t}\epsilon_t + e^{-\alpha_t}(1-\epsilon_t) \qquad \text{(From the definition of } \epsilon_t) \end{aligned} \tag{7}$$

2. Then we need to derive α_t in terms of ϵ_t.
Suppose that we have run the AdaBoost algorithm for total T iterations.
Let $H_{final} = \text{sign}(\sum_{t=1}^{T} \alpha_t h_t)$
So we have the final training error is that

$$\begin{aligned}
\epsilon &= \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}(y_i \neq H_{final}(x_i)) \\
&= \frac{1}{n}\sum_{i=1}^{n} \begin{cases} 1 & \text{if } y_i \neq H_{\text{final}}\ (x_i) \\ 0 & \text{otherwise} \end{cases} \\
&= \frac{1}{n}\sum_{i=1}^{n} \begin{cases} 1 & \text{if } y_i(\sum_{t=1}^{T} \alpha_t h_t) \leq 0 \\ 0 & \text{otherwise} \end{cases} \\
&\leq \frac{1}{n}\sum_{i=1}^{n} \exp\left(-y_i(\sum_{t=1}^{T} \alpha_t h_t)\right)
\end{aligned} \tag{8}$$

Since we totally have T iterations, so for each iteration, we have

$$\begin{aligned}
D_{T+1}(i) &= \frac{D_T(i)}{Z_T}\exp(-\alpha_T y_i h_T(x_i)) \\
D_T(i) &= \frac{D_{T-1}(i)}{Z_{T-1}}\exp(-\alpha_{T-1} y_i h_{T-1}(x_i)) \\
&\vdots \\
D_2(i) &= \frac{D_1(i)}{Z_2}\exp(-\alpha_1 y_i h_1(x_i)) \\
D_1(i) &= \frac{1}{n}
\end{aligned}$$

Multiply these equations, we can get that

$$D_{T+1}(i) = \frac{1}{n}\cdot\prod_{t=1}^{T}\frac{1}{Z_t}\exp(-\alpha_t y_i h_t(x_i)) = \frac{1}{n}\cdot\frac{1}{\prod_{t=1}^{T} Z_t}\cdot\exp\left(-y_i\sum_{t=1}^{T}\alpha_t h_t(x_i)\right)$$

i.e.

$$\frac{1}{n}\cdot\exp\left(-y_i\sum_{t=1}^{T}\alpha_t h_t(x_i)\right) = D_{T+1}(i)\prod_{t=1}^{T} Z_t \tag{9}$$

If we put the equation (9) into the last of the equation (8), we can get that

$$\epsilon \leq \frac{1}{n}\sum_{i=1}^{n}\exp\left(-y_i(\sum_{t=1}^{T}\alpha_t h_t)\right) = \sum_{i=1}^{n} D_{T+1}(i)\prod_{t=1}^{T} Z_t = \prod_{t=1}^{T} Z_t\left(\sum_{i=1}^{n} D_{T+1}(i)\right)$$

Since Z_t is to make sure D_{t+1} is normalized correctly, so we have $\sum_{i=1}^{n} D_{T+1}(i) = 1$, so we have

$$\epsilon \leq \prod_{t=1}^{T} Z_t$$

So if we want to minimize the final error ϵ, we should minimize $\prod_{t=1}^{T} Z_t$. i.e. we should minimize Z_t for each $t = 1, 2, \cdots, T$.
So for each $Z_t = e^{\alpha_t}\epsilon_t + e^{-\alpha_t}(1-\epsilon_t)$

$$\begin{aligned}
\frac{\partial Z_t}{\partial \alpha_t} &= \epsilon_t e^{\alpha_t} - (1-\epsilon_t)e^{-\alpha_t} \\
\frac{\partial^2 Z_t}{\partial \alpha_t^2} &= \epsilon_t e^{\alpha_t} + (1-\epsilon_t)e^{-\alpha_t} > 0
\end{aligned}$$

So we can get that Z_t is a convex function of α_t.
So to minimize Z_t, we should make $\dfrac{\partial Z_t}{\partial \alpha_t} = 0$.
i.e.

$$\begin{aligned} \epsilon_t e^{\alpha_t} &= (1-\epsilon_t)e^{-\alpha_t} \\ e^{2\alpha_t} &= \frac{1-\epsilon_t}{\epsilon_t} \\ \alpha_t &= \frac{1}{2}\log\frac{1-\epsilon_t}{\epsilon_t} \end{aligned}$$

Since $\epsilon_t = \mathbb{E}_{D_t}[\mathbb{1}(y_i \neq h_t(x_i))] = P_{D_t}(y_i \neq h_t(x_i))$, so $\epsilon_t \in (0,1)$.
So we have $\dfrac{1-\epsilon_t}{\epsilon_t} > 0$, so $\log\dfrac{1-\epsilon_t}{\epsilon_t}$ is valid.

So we have derived that $\alpha_t = \dfrac{1}{2}\log\dfrac{1-\epsilon_t}{\epsilon_t}$.
And put it into the equation (7), we can get that

$$Z_t = \epsilon_t e^{\alpha_t} + (1-\epsilon_t)e^{-\alpha_t} = \epsilon_t\sqrt{\frac{1-\epsilon_t}{\epsilon_t}} + (1-\epsilon_t)\sqrt{\frac{\epsilon_t}{1-\epsilon_t}} = 2\sqrt{\epsilon_t(1-\epsilon_t)}$$

So above all, we have derived that

$$\begin{aligned} Z_t &= 2\sqrt{\epsilon_t(1-\epsilon_t)} \\ \alpha_t &= \frac{1}{2}\log\frac{1-\epsilon_t}{\epsilon_t} \end{aligned}$$

(3) From (2), we can get that $Z_1 = 2\sqrt{\epsilon_1(1-\epsilon_1)} = 2\sqrt{0.4\cdot 0.6} = 0.4\sqrt{6}$.
And $\alpha_1 = \dfrac{1}{2}\log\dfrac{1-\epsilon_1}{\epsilon_1} = \dfrac{1}{2}\log\dfrac{1-0.4}{0.4} = \dfrac{1}{2}\log\dfrac{3}{2}$.
Since we take $h_1 = h^{(1)}$, so

$$D_2(i) = \frac{D_1(i)}{Z_1}\exp\left(-\alpha_1 y_i h_1(x_i)\right) = \frac{1}{10\cdot 0.4\sqrt{6}}\exp\left(-\frac{1}{2}\log\frac{3}{2}\cdot y_i\cdot h^{(1)}(x_i)\right)$$

From (a), we can get that for points x_1, x_7, x_8, x_9, which are misclassified, so we have $y_i \cdot h^{(1)}(x_i) = -1$.
So their weight $D_2(i) = \dfrac{1}{10\cdot 0.4\sqrt{6}}\exp\left(-\dfrac{1}{2}\log\dfrac{3}{2}\cdot(-1)\right) = \dfrac{1}{10\cdot 0.4\sqrt{6}}\cdot\sqrt{\dfrac{3}{2}} = \dfrac{1}{8} > D_1(i) = \dfrac{1}{10}$.
And for other points $x_2, x_3, x_4, x_5, x_6, x_{10}$, which are correctly classified, so we have $y_i \cdot h^{(1)}(x_i) = 1$.
So their weight $D_2(i) = \dfrac{1}{10\cdot 0.4\sqrt{6}}\exp\left(-\dfrac{1}{2}\log\dfrac{3}{2}\cdot 1\right) = \dfrac{1}{10\cdot 0.4\sqrt{6}}\cdot\sqrt{\dfrac{2}{3}} = \dfrac{1}{12} < D_1(i) = \dfrac{1}{10}$.

So above all, the misclassified points x_1, x_7, x_8, x_9 will increase in significance in the second round of boosting, and their weight $D_2(i) = \dfrac{1}{8}$.

(4) From (3), we know that $D_2(1) = D_2(7) = D_2(8) = D_2(9) = \dfrac{1}{8}$,
and $D_2(2) = D_2(3) = D_2(4) = D_2(5) = D_2(6) = D_2(10) = \dfrac{1}{12}$.
So for each classifer $h^{(j)}$, we can get the error $(\epsilon_2)_j$ is

$$(\epsilon_2)_j = \mathbb{E}_{D_2}[\mathbb{1}(y_i \neq h^{(j)}(x_i))] = \sum_{i=1}^{n} D_2(i)\mathbb{1}(y_i \neq h^{(j)}(x_i))$$

- For $h^{(1)}$, we have $(\epsilon_2)_1 = \dfrac{1}{8}\cdot 4 + \dfrac{1}{12}\cdot 0 = \dfrac{1}{2}$.
- For $h^{(2)}$, we have $(\epsilon_2)_2 = \dfrac{1}{8}\cdot 2 + \dfrac{1}{12}\cdot 2 = \dfrac{5}{12}$.
- For $h^{(3)}$, we have $(\epsilon_2)_3 = \dfrac{1}{8}\cdot 1 + \dfrac{1}{12}\cdot 4 = \dfrac{11}{24}$.

- For $h^{(4)}$, we have $(\epsilon_2)_4 = \frac{1}{8} \cdot 1 + \frac{1}{12} \cdot 3 = \frac{3}{8}$.
- For $h^{(5)}$, we have $(\epsilon_2)_5 = \frac{1}{8} \cdot 2 + \frac{1}{12} \cdot 2 = \frac{5}{12}$.
- For $h^{(6)}$, we have $(\epsilon_2)_6 = \frac{1}{8} \cdot 3 + \frac{1}{12} \cdot 2 = \frac{13}{24}$.

So above all, the minimum value of ϵ_2 is $\frac{3}{8}$, and the classifier $h^{(4)}$ achieve this value.

(5) From (1), we can get that $\alpha_1 = \frac{1}{2} \log \frac{3}{2}$.
And from (4), we can get that $\epsilon_2 = \frac{3}{8}$.
So $\alpha_2 = \frac{1}{2} \log \frac{1-\epsilon_2}{\epsilon_2} = \frac{1}{2} \log \frac{1-\frac{3}{8}}{\frac{3}{8}} = \frac{1}{2} \log \frac{5}{3}$.
And since $h_1(x) = h^{(1)}(x)$ and $h_2(x) = h^{(4)}(x)$, so

$$H(x) = \text{sign}(\alpha_1 h_1(x) + \alpha_2 h_2(x)) = \text{sign}\left(\frac{1}{2} \log \frac{3}{2} h^{(1)}(x) + \frac{1}{2} \log \frac{5}{3} h^{(4)}(x)\right)$$

There are total 4 possible combinations of $h^{(1)}(x)$ and $h^{(4)}(x)$, which are $(-1,-1), (1,-1), (-1,1), (1,1)$.
So we can get that

- For $(-1,-1)$, we have $H(x) = \text{sign}\left(\frac{1}{2} \log \frac{3}{2} \cdot (-1) + \frac{1}{2} \log \frac{5}{3} \cdot (-1)\right) = -1$.
- For $(1,-1)$, we have $H(x) = \text{sign}\left(\frac{1}{2} \log \frac{3}{2} \cdot 1 + \frac{1}{2} \log \frac{5}{3} \cdot (-1)\right) = \text{sign}(\frac{1}{2} \log \frac{9}{10}) = -1$.
- For $(-1,1)$, we have $H(x) = \text{sign}\left(\frac{1}{2} \log \frac{3}{2} \cdot (-1) + \frac{1}{2} \log \frac{5}{3} \cdot 1\right) = \text{sign}(\frac{1}{2} \log \frac{10}{9}) = 1$.
- For $(1,1)$, we have $H(x) = \text{sign}\left(\frac{1}{2} \log \frac{3}{2} \cdot 1 + \frac{1}{2} \log \frac{5}{3} \cdot 1\right) = 1$.

So we can get that x_3, x_4, x_5 are misclassified by $H(x)$, so we have

$$\epsilon = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}(y_i \neq H(x_i)) = \frac{1}{10} \cdot 3 = 0.3$$

And we have $\epsilon_1 = \min_{i=1,2,\cdots,6} (\epsilon_1)_i = \min\{0.4, 0.4, 0.5, 0.4, 0.4, 0.5\} = 0.4$, so we can get that $\epsilon < \epsilon_1$.

So above all, the average error of the final classifier H is 0.3, and it is less than the error we would get, if we just used one of the weak classifiers instead of this final classifier H.

3. [10 points] [Perceptron Learning Algorithm] Consider a binary classification problem. The input space is $\mathbb{R}^d$. The output space is $\{+1, -1\}$. For simplicity, we modified the input to be $\mathbf{x} = [x_0, x_1, \cdots, x_d]^\intercal$ with $x_0 = 1$. The output is predicted using the hypothesis:

$$h(\mathbf{x}) = \text{sign}(\mathbf{w}^\intercal \mathbf{x}), \tag{10}$$

where $\mathbf{w} = [w_0, w_1, \cdots, w_d]^\intercal$ and w_0 is the bias.

The *perceptron learning algorithm* determines $\mathbf{w}$ using a simple iterative method. Here is how it works. At iteration t, where $t = 0, 1, 2, \ldots$, there is a current value of the weight vector, call it $\mathbf{w}(t)$. The algorithm picks an example from $(\mathbf{x}_1, y_1) \cdots (\mathbf{x}_N, y_N)$ that is currently misclassified, call it $(\mathbf{x}(t), y(t))$, and uses it to update $\mathbf{w}(t)$. Since the example is misclassified, we have $y(t) \neq \text{sign}\left(\mathbf{w}^{\mathrm{T}}(t)\mathbf{x}(t)\right)$. The update rule is

$$\mathbf{w}(t+1) = \mathbf{w}(t) + y(t)\mathbf{x}(t). \tag{11}$$

(a) Show that $y(t)\mathbf{w}^{\mathrm{T}}(t)\mathbf{x}(t) < 0$. [Hint: $\mathbf{x}(t)$ is misclassified by $\mathbf{w}(t)$.] [3 points]

(b) Show that $y(t)\mathbf{w}^{\mathrm{T}}(t+1)\mathbf{x}(t) > y(t)\mathbf{w}^{\mathrm{T}}(t)\mathbf{x}(t)$. [3 points]

(c) As far as classifying $\mathbf{x}(t)$ is concerned, argue that the move from $\mathbf{w}(t)$ to $\mathbf{w}(t+1)$ is a move "in the right direction". [4 points]

Solution:

(a) Since we are considering the misclassified, so we have $y(t) \neq \text{sign}\left(\mathbf{w}^{\mathrm{T}}(t)\mathbf{x}(t)\right)$.
And since $y(t), \text{sign}\left(\mathbf{w}^{\mathrm{T}}(t)\mathbf{x}(t)\right) \in \{+1, -1\}$, so we have $y(t) \cdot \text{sign}\left(\mathbf{w}^{\mathrm{T}}(t)\mathbf{x}(t)\right) = -1 < 0$.
Suppose that $\text{sign}\left(\mathbf{w}^{\mathrm{T}}(t)\mathbf{x}(t)\right) = k \cdot \mathbf{w}^{\mathrm{T}}(t)\mathbf{x}(t)$, where $k > 0$.
So $y(t) \cdot \text{sign}\left(\mathbf{w}^{\mathrm{T}}(t)\mathbf{x}(t)\right) = y(t) \cdot k \cdot \mathbf{w}^{\mathrm{T}}(t)\mathbf{x}(t) = k \cdot y(t)\mathbf{w}^{\mathrm{T}}(t)\mathbf{x}(t) < 0$.
Since $k > 0$, so we have

$$y(t)\mathbf{w}^{\mathrm{T}}(t)\mathbf{x}(t) < 0$$

So above all, we have proved that $y(t)\mathbf{w}^{\mathrm{T}}(t)\mathbf{x}(t) < 0$.

(b) Since we are considering the misclassified, so we have $\mathbf{w}(t+1) = \mathbf{w}(t) + y(t)\mathbf{x}(t)$.
So

$$y(t)\mathbf{w}^{T}(t+1)\mathbf{x}(t) = y(t)\mathbf{w}^{T}(t)\mathbf{x}(t) + y(t)y(t)\mathbf{x}^{T}(t)\mathbf{x}(t) = y(t)\mathbf{w}^{T}(t)\mathbf{x}(t) + y^2(t)\|\mathbf{x}(t)\|^2$$

Since $y(t) \in \{+1, -1\}$, so we have $y^2(t) = 1$.
And since for the simplicity, we have the input to be $\mathbf{x} = [x_0, x_1, \cdots, x_d]^\intercal$ with $x_0 = 1$, so we have $\mathbf{x} \neq \mathbf{0}$, i.e. $\|\mathbf{x}(t)\|^2 > 0$.
So we have

$$y^2(t)\|\mathbf{x}(t)\|^2 > 0$$

So

$$y(t)\mathbf{w}^{T}(t+1)\mathbf{x}(t) = y(t)\mathbf{w}^{T}(t)\mathbf{x}(t) + y^2(t)\|\mathbf{x}(t)\|^2 > y(t)\mathbf{w}^{T}(t)\mathbf{x}(t)$$

So above all, we have proved that $y(t)\mathbf{w}^{\mathrm{T}}(t+1)\mathbf{x}(t) > y(t)\mathbf{w}^{\mathrm{T}}(t)\mathbf{x}(t)$.

(c) We only consider about the misclassified case.
From (a), we knew that

$$y(t)\mathbf{w}^{\mathrm{T}}(t)\mathbf{x}(t) < 0$$

And from (b), we knew that

$$y(t)\mathbf{w}^{\mathrm{T}}(t+1)\mathbf{x}(t) > y(t)\mathbf{w}^{\mathrm{T}}(t)\mathbf{x}(t)$$

So we could see that the move from $\mathbf{w}(t)$ to $\mathbf{w}(t+1)$ is making the $y(t)\mathbf{w}^{T}\mathbf{x}(t)$ to the more positive direction, and since if $y(t)\mathbf{w}^{T}\mathbf{x}(t) > 0$, then it is a correct classification.
And if the total input data are linearly separable, from what we have learned, we could get that with at most $M = (\frac{R}{\gamma})^2$ such misclassified's movement, where R is the radius of the smallest sphere that contains all the input data, and γ is the margin, then we could get the correct classification.

So above all, we could say that the move from $\mathbf{w}(t)$ to $\mathbf{w}(t+1)$ is a move "in the right direction".