Machine Learning Homework 1

B04705003 Tzu-Chuan, Lin

Problem 1. I get 100 points. See Figure 1.

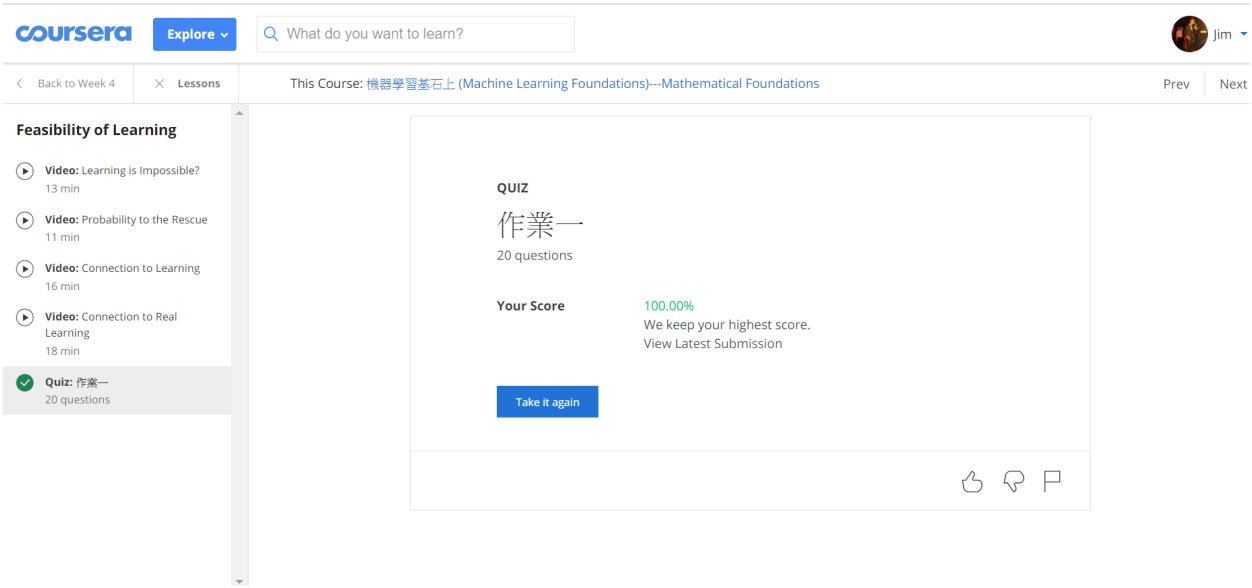

Figure 1: Coursera score

Problem 2. Consider four cases:

1. Case 1: N is odd and L is odd.

$$E_{OTS}(g,f) = \frac{1}{L}\left(\sum_{i=1}^{\frac{L+1}{2}} [\![g(\mathbf{x}_{N+2i-1}) \neq f(\mathbf{x}_{N+2i-1})]\!] + \sum_{i=1}^{\frac{L-1}{2}} [\![g(\mathbf{x}_{N+2i}) \neq f(\mathbf{x}_{N+2i})]\!]\right)$$

(separate indices into odd and even)

$$= \frac{1}{L}\left(\frac{L+1}{2} + 0\right) = \frac{L+1}{2L}$$

2. Case 2: N is odd and L is even.

$$E_{OTS}(g,f) = \frac{1}{L}\left(\sum_{i=1}^{\frac{L}{2}} [\![g(\mathbf{x}_{N+2i-1}) \neq f(\mathbf{x}_{N+2i-1})]\!] + \sum_{i=1}^{\frac{L}{2}} [\![g(\mathbf{x}_{N+2i}) \neq f(\mathbf{x}_{N+2i})]\!]\right)$$

$$= \frac{1}{L}\left(\frac{L}{2} + 0\right) = \frac{L}{2L}$$

3. Case 3: N is even and L is odd.

$$E_{OTS}(g,f) = \frac{1}{L}(\sum_{i=1}^{\frac{L+1}{2}}[\![g(\mathbf{x}_{N+2i-1}) \neq f(\mathbf{x}_{N+2i-1})]\!] + \sum_{i=1}^{\frac{L-1}{2}}[\![g(\mathbf{x}_{N+2i}) \neq f(\mathbf{x}_{N+2i})]\!])$$
$$= \frac{1}{L}(0 + \frac{L-1}{2}) = \frac{L-1}{2L}$$

4. Case 4: N is even and L is even.

$$E_{OTS}(g,f) = \frac{1}{L}(\sum_{i=1}^{\frac{L}{2}}[\![g(\mathbf{x}_{N+2i-1}) \neq f(\mathbf{x}_{N+2i-1})]\!] + \sum_{i=1}^{\frac{L}{2}}[\![g(\mathbf{x}_{N+2i}) \neq f(\mathbf{x}_{N+2i})]\!])$$
$$= \frac{1}{L}(0 + \frac{L}{2}) = \frac{L}{2L}$$

Problem 3.

Proof. Define the event E = function that can noiselessly generate $\mathcal{D}$.

Define the function set that can cause event E as $\mathcal{F}$.

We can divide F set into L disjoint sets corresponding to a function g, formally written as:

$$\mathcal{F} = \bigcup_{\ell=0}^{L} \mathcal{F}_\ell^g$$

where $\forall f \in \mathcal{F}_\ell^g, f$ causes $E_{OTS}(g,f) = \frac{\ell}{L}$

By the statement - "all those f that can generate $\mathcal{D}$ in a noiseless setting are equally likely in probability", we can know that for every g,

$$p(\mathcal{F}_i^g) : p(\mathcal{F}_j^g) = \binom{L}{i} : \binom{L}{j}$$

It means the probabilistic proportion of g getting i to j incorrect answers on **test inputs** is $\binom{L}{i}$ to $\binom{L}{j}$. You can see Table 1 for illustration. Then, let

$$p(\mathcal{F}_i^g) = \binom{L}{i} r$$

Due to

$$\sum_{\ell=0}^{L} p(\mathcal{F}_\ell^g) = 1$$

We can get:

$$\sum_{\ell=0}^{L} \binom{L}{\ell} r = 1 \Rightarrow r = \frac{1}{2^L}$$

Let $\mathcal{A}_1(D) = g_1$ and $\mathcal{A}_2(D) = g_2$. Then, expand the left hand side and the right hand side:

$$\begin{aligned}
LHS &= \mathbb{E}_{f\in\mathcal{F}}\{E_{OTS}(\mathcal{A}_1(D), f)\} \\
&= \mathbb{E}_{f\in\mathcal{F}}\{E_{OTS}(g_1, f)\} \\
&= \sum_{\ell=0}^{L} \mathbb{E}_{f\in\mathcal{F}_\ell}\{E_{OTS}(g_1, f)\} \\
&= \sum_{\ell=0}^{L} \mathbb{E}_{f\in\mathcal{F}_\ell}\{\frac{\ell}{L}\} \text{ (by the defintion we just defined)} \\
&= \sum_{\ell=0}^{L} \frac{\ell}{L} \cdot \mathbb{E}_{f\in\mathcal{F}_\ell}\{1\} \\
&= \sum_{\ell=0}^{L} \frac{\ell}{L} \cdot p(\mathcal{F}_\ell^{g_1}) \\
&= \sum_{\ell=0}^{L} \frac{\ell}{L} \cdot \frac{\binom{L}{\ell}}{2^L}
\end{aligned}$$

$$\begin{aligned}
RHS &= \mathbb{E}_{f\in\mathcal{F}}\{E_{OTS}(\mathcal{A}_2(D), f)\} \\
&= \mathbb{E}_{f\in\mathcal{F}}\{E_{OTS}(g_2, f)\} \\
&= \sum_{\ell=0}^{L} \mathbb{E}_{f\in\mathcal{F}_\ell}\{E_{OTS}(g_2, f)\} \\
&= \sum_{\ell=0}^{L} \mathbb{E}_{f\in\mathcal{F}_\ell}\{\frac{\ell}{L}\} \text{ (by the defintion we just defined)} \\
&= \sum_{\ell=0}^{L} \frac{\ell}{L} \cdot \mathbb{E}_{f\in\mathcal{F}_\ell}\{1\} \\
&= \sum_{\ell=0}^{L} \frac{\ell}{L} \cdot p(\mathcal{F}_\ell^{g_2}) \\
&= \sum_{\ell=0}^{L} \frac{\ell}{L} \cdot \frac{\binom{L}{\ell}}{2^L}
\end{aligned}$$

We get:

$$LHS = RHS$$

. □

Problem 4. Let X be the random variable of the number of orange marbles in a sample of 10 marbles.

	f_1	f_2	f_3	f_4	f_5	f_6	f_7	f_8
$\mathcal{D}$	o	o	o	o	o	o	o	o
testing set	o	o	o	o	x	x	x	x
	o	o	x	x	o	o	x	x
other examples	o	x	o	x	o	x	o	x

Table 1: Problem 3 illustration (Assume ground truth of all data are o)

1.

$$\nu \leq 0.1 \Longleftrightarrow X \leq 1$$

Compute the probability:

$$\begin{aligned} P(X \leq 1) = P(X = 0) + P(X = 1) &= \binom{10}{0}\mu^0(1-\mu)^{10} + \binom{10}{1}\mu^1(1-\mu)^9 \\ &= \frac{41}{5^{10}} \end{aligned}$$

2.

$$\nu \geq 0.9 \Longleftrightarrow X \geq 9$$

Compute the probability:

$$\begin{aligned} P(X \geq 9) = P(X = 9) + P(X = 10) &= \binom{10}{9}\mu^9(1-\mu)^1 + \binom{10}{10}\mu^{10}(1-\mu)^0 \\ &= \frac{10 * 4^9 + 4^{10}}{5^{10}} = \frac{14 * 4^9}{5^{10}} \end{aligned}$$

Problem 5. If you want to get green 1, you can only get them from A and D dices. Therefore, for each dice, the probability you get green 1 is $\frac{1}{2}$.

$$\begin{aligned} P(\text{five green 1's}) &= \binom{5}{5}(\frac{1}{2})^5 \\ &= \frac{1}{32} \end{aligned} \tag{1}$$

Problem 6. Let X be the random variable of which dice number is purely green

$$P(\text{some number purely green}) = \sum_{i=1}^{3} P(\text{exactly } i \text{ numbers have purely green}) \tag{2}$$

$$\begin{aligned} P(\text{exactly 1 number have purely green}) &= \sum_{i=1}^{6} P(\text{dice number } i) \\ &= 0 + ((\frac{1}{2})^5 - 2 \cdot (\frac{1}{4})^5) + 0 + 0 + ((\frac{1}{2})^5 - 2 \cdot (\frac{1}{4})^5) + 0 \\ &= \frac{15}{256} \end{aligned} \tag{3}$$

$$
\begin{aligned}
P(\text{exactly 2 number have purely green}) &= P((1,3)) + P((4,6)) \\
&= ((\frac{1}{2})^5 - 2 \cdot (\frac{1}{4})^5) + ((\frac{1}{2})^5 - 2 \cdot (\frac{1}{4})^5) \\
&= \frac{15}{256}
\end{aligned} \tag{4}
$$

$$
\begin{aligned}
&P(\text{exactly 3 number have purely green}) \\
&\quad = P(\text{all A dices}) + P(\text{all B dices}) + P(\text{all C dices}) + P(\text{all D dices}) \\
&\quad = ((\frac{1}{4})^5) + ((\frac{1}{4})^5) + ((\frac{1}{4})^5) + ((\frac{1}{4})^5) \\
&\quad = \frac{1}{256}
\end{aligned} \tag{5}
$$

In total,

$$
P(\text{some number purely green}) = \frac{31}{256}
$$

In fact, all numbers from 1 to 6 have same probability of getting green number(i.e. $\frac{1}{2}$) but we have high probability of getting **some** of dice numbers purely green. In general, we hope that a purely green number carries significant statistical meaning. However, in this problem setting, we actually know that each number has same probability of getting green one. Therefore, the showing of purely green may not as important as we think.

If each number represent a hypothesis function and **orange** number is analogous to **BAD** for a hypothesis function. This result just told us that we can easily think we get a great result but we actually have all hypothesis functions in the hypothesis set with same $|E_{in}(h) - E_{out}(h)|$ probability.

Problem 7.

1. Average number of updates: about 40 times.

2. Histogram: See Figure 2.

Problem 8(Bonus).

1. Prove that M exists by deriving its formula.

$$
\begin{aligned}
y_{n(t)}\mathbf{w}_{t+1}^T\mathbf{x}_{n(t)} &= y_{n(t)}(\mathbf{w}_t + y_{n(t)}\mathbf{x}_{n(t)} \cdot M)^T\mathbf{x}_{n(t)} \\
&= y_{n(t)}\mathbf{w}_t^T\mathbf{x}_{n(t)} + y_{n(t)}^2 M \left\|\mathbf{x}_{n(t)}\right\|_2^2 \\
&= y_{n(t)}\mathbf{w}_t^T\mathbf{x}_{n(t)} + M \left\|\mathbf{x}_{n(t)}\right\|_2^2
\end{aligned} \tag{6}
$$

Set (6) to be greater than zero:

$$
y_{n(t)}\mathbf{w}_{t+1}^T\mathbf{x}_{n(t)} = y_{n(t)}\mathbf{w}_t^T\mathbf{x}_{n(t)} + M \left\|\mathbf{x}_{n(t)}\right\|_2^2 > 0
$$

Then, we can find the bound of M.

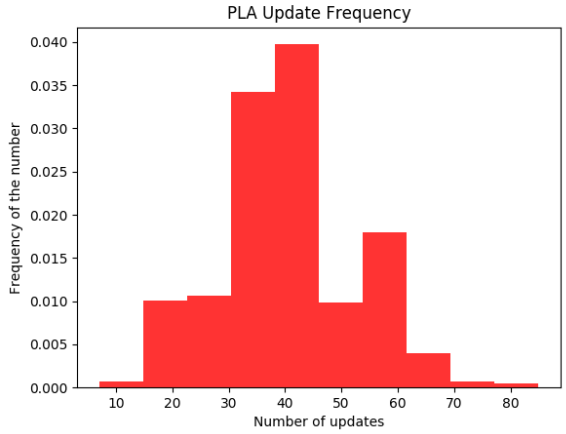

Figure 2: Pre-determined random cycles PLA number of updates histogram

Due to the bias ($\mathbf{x}_{n(t),0} = 1$), $\left\|\mathbf{x}_{n(t)}\right\|_2^2 \neq 0$. Therefore, we can derive M range.

$$M > (-y_{n(t)}\mathbf{w}_t^T\mathbf{x}_{n(t)}) / \left\|\mathbf{x}_n(t)\right\|_2^2 \quad (-y_{n(t)}\mathbf{w}_t\mathbf{x}_{n(t)} \geq 0 \because y_{n(t)} \neq sign(\mathbf{w}_t^T\mathbf{x}_{n(t)})) \tag{7}$$

$$\geq 0 \quad \text{(Note that this may be equal to zero)} \tag{8}$$

Therefore, we can always find the smallest integer (also positive) M such that $y_{n(t)}\mathbf{w}_{t+1}^T\mathbf{x}_{n(t)}$ is greater than zero.

2. Prove or disprove the claim:

Proof. Because the linear separable property, we can assume the existence of $\mathbf{w}_f$ such that $\forall n, y_n = sign(\mathbf{w}_f^T\mathbf{x}_n)$. Consider:

$$\begin{aligned}
\mathbf{w}_f^T\mathbf{w}_{t+1} &= \mathbf{w}_f^T(\mathbf{w}_t + y_{n(t)}\mathbf{x}_{n(t)} \cdot M_t) \\
&= \mathbf{w}_f^T\mathbf{w}_t + y_{n(t)}\mathbf{w}_f^T\mathbf{x}_{n(t)} \cdot M_t \\
&\geq \mathbf{w}_f^T\mathbf{w}_t + M_t \min_n y_n\mathbf{w}_f^T\mathbf{x}_n \\
&> \mathbf{w}_f^T\mathbf{w}_t + 0
\end{aligned} \tag{9}$$

where M_t denotes the smallest positive integer updated at time t.

Then, consider that $\mathbf{w}_t$ makes mistake at $\mathbf{x}_{n(t)}$, which means $y_{n(t)}\mathbf{w}_t^T\mathbf{x}_{n(t)} \leq 0$

$$\begin{aligned}
\left\|\mathbf{w}_{t+1}\right\|^2 &= \left\|\mathbf{w}_t + M_t \cdot y_{n(t)}\mathbf{x}_{n(t)}\right\|^2 \\
&= \left\|\mathbf{w}_t\right\|^2 + 2M_t \cdot y_{n(t)}\mathbf{w}_t^T\mathbf{x}_{n(t)} + \left\|M_t \cdot y_{n(t)}\mathbf{x}_{n(t)}\right\|^2 \\
&\leq \left\|\mathbf{w}_t\right\|^2 + 0 + M_t^2\left\|\mathbf{x}_{n(t)}\right\|^2 (\because M_t > 0 \text{ by inequality (8)}) \\
&\leq \left\|\mathbf{w}_t\right\|^2 + M_t^2 \max_n \left\|\mathbf{x}_n\right\|^2
\end{aligned} \tag{10}$$

Then, consider the inner product of $\mathbf{w}_f$ and $\mathbf{w}_T$:

$$\begin{aligned}
\mathbf{w}_f^T \mathbf{w}_T &\geq \mathbf{w}_f^T \mathbf{w}_{T-1} + M_{T-1} \min_n y_n \mathbf{w}_f^T \mathbf{x}_n \\
&\geq \mathbf{w}_f^T \mathbf{w}_{T-2} + (\sum_{i=1}^{2} M_{T-i}) \min_n y_n \mathbf{w}_f^T \mathbf{x}_n \text{ (plug inequality (9))} \\
&\geq \mathbf{w}_f^T \mathbf{w}_0 + (\sum_{i=1}^{T} M_{T-i}) \min_n y_n \mathbf{w}_f^T \mathbf{x}_n \\
&\geq (\sum_{i=1}^{T} M_{T-i}) \min_n y_n \mathbf{w}_f^T \mathbf{x}_n \ (\because \mathbf{w}_0 = 0) \\
&> 0
\end{aligned} \tag{11}$$

Then, consider $\|\mathbf{w}_T\|^2$:

$$\begin{aligned}
\|\mathbf{w}_T\|^2 &\leq \|\mathbf{w}_{T-1}\|^2 + (\sum_{i=1}^{1} M_{T-i}^2) \max_n \|\mathbf{x}_n\|^2 \\
&\leq \|\mathbf{w}_0\|^2 + (\sum_{i=1}^{T} M_{T-i}^2) \max_n \|\mathbf{x}_n\|^2 \\
&\leq (\sum_{i=1}^{T} M_{T-i}^2) \max_n \|\mathbf{x}_n\|^2 \ (\because \mathbf{w}_0 = 0)
\end{aligned} \tag{12}$$

Finally, consider cosine similarity and plug inequalities (11) and (12) into it. We can get:

$$\begin{aligned}
\frac{\mathbf{w}_f^T \mathbf{w}_T}{\|\mathbf{w}_f\| \|\mathbf{w}_T\|} &\geq \frac{\sum_{i=1}^{T} M_{T-i}}{\sum_{i=1}^{T} M_{T-i}^2} \cdot \frac{\min_n y_n \mathbf{w}_f^T \mathbf{x}_n}{\|\mathbf{w}_f\| \sqrt{\max_n \|\mathbf{x}_n\|^2}} \\
&= \frac{\sum_{i=1}^{T} M_{T-i}}{\sum_{i=1}^{T} M_{T-i}^2} \cdot \text{constant}
\end{aligned} \tag{13}$$

Also, we know cosine similarity's upper bound is 1 and $M_t \geq 1$. We can get:

$$\frac{\sum_{i=1}^{T} M_{T-i}}{\sum_{i=1}^{T} M_{T-i}^2} \cdot \text{constant} \leq 1 \Rightarrow \frac{\sum_{i=1}^{T} M_{T-i}}{\sum_{i=1}^{T} M_{T-i}^2} \leq \frac{1}{\text{constant}}$$

And because $M_t \geq 1$, we know that:

$$\sum_{i=1}^{T} M_{T-i}^2 \leq (\sum_{i=1}^{T} M_{T-i})^2$$

By using inequality above:

$$\sqrt{T} \leq \sqrt{\sum_{i=1}^{T} M_{T-i}} = \frac{\sum_{i=1}^{T} M_{T-i}}{(\sum_{i=1}^{T} M_{T-i})^2} \leq \frac{\sum_{i=1}^{T} M_{T-i}}{\sum_{i=1}^{T} M_{T-i}^2} \leq \frac{1}{\text{constant}}$$

And conclude that:

$$T \leq \frac{1}{\text{constant}^2}$$

Therefore, this algorithm will halt. □