HW 1

Wednesday, September 30, 2020 9:13 PM

1(1.3)

(a) Proof:

Let $1 \le n \le N$. Consider two cases when $y_n = 1$ and $y_n = -1$. Since w^* is an optimal set of weights which correctly separates all the data points,

when $y_n = 1$, $\text{sign}(w^{*T}x_n) = y_n = 1$;

when $y_n = -1$, $\text{sign}(w^{*T}x_n) = y_n = -1$.

Therefore, $y_n w^{*T} x_n > 0 \quad \forall\, 1 \le n \le N$.

Thus, we can show that $\rho = \min_{1\le n\le N} y_n(w^{*T}x_n) > 0$

$\square$

(b) Proof:

At iteration t of PLA, we update $w(t)$ using the current misclassified data point such that

$w(t+1) = w(t) + y(t)x(t)$

Therefore,

$$w(t)w^* = (w(t-1) + y(t-1)x(t-1))\,w^*$$
$$\begin{aligned} w^T(t)w^* &= (w^T(t-1) + y(t-1)x^T(t-1))w^* \quad (y(t-1) \text{ is a constant, i.e. } +1 \text{ or } -1) \\ &= w^T(t-1)w^* + y(t-1)x^T(t-1)w^* \\ &= w^T(t-1)w^* + y(t-1)w^{*T}x(t-1) \\ &\ge w^T(t-1)w^* + \min_{1\le n\le N} y_n(w^{*T}x_n) \\ &= w^T(t-1)w^* + \rho \end{aligned}$$

Therefore, $w^T(t)w^* \ge w^T(t-1)w^* + \rho$ ①

To prove $w^T(t)w^* \ge t\rho$, we have

Base case:

when $t=0$, since $w(0)=0$, $w^T(0)w^* = 0 = t\rho$.

So the base case holds.

Inductive case:

Suppose $w(n)$ is true, then by ①,

$$\begin{aligned} w^T(n+1)w^* &\ge w^T(n)w^* + \rho \\ &\ge n\rho + \rho = (n+1)\rho \end{aligned}$$

Therefore, the inductive case holds.

So we can conclude $w^T(t)w^* \ge t\rho$.

$\square$

(c) Proof:

At iteration t of PLA, we update $w(t)$ using the current misclassified data point such that

$w(t+1) = w(t) + y(t)x(t)$

Therefore at iteration $t-1$, we have

$w(t) = w(t-1) + y(t-1)x(t-1)$

Take the square of the norm of both sides,

$$\begin{aligned} \|w(t)\|^2 &= \|w(t-1) + y(t-1)x(t-1)\|^2 \\ &= (w(t-1) + y(t-1)x(t-1))^T(w(t-1) + y(t-1)x(t-1)) \\ &= w^T(t-1)w(t-1) + y(t-1)x^T(t-1)w(t-1) + w^T(t-1)y(t-1)x(t-1) + y(t-1)x^T(t-1)y(t-1)x(t-1) \\ &= w^T(t-1)w(t-1) + 2y(t-1)w^T(t-1)x(t-1) + x^T(t-1)x(t-1) \\ &= \|w(t-1)\|^2 + \|x(t-1)\|^2 + 2y(t-1)w^T(t-1)x(t-1) \end{aligned}$$

Since $y(t-1)\cdot(w^T(t-1)x(t-1)) \le 0$ because $x(t-1)$ was misclassified by $w(t-1)$,

$\|w(t)\|^2 \le \|w(t-1)^2\| + \|x(t-1)\|^2$

$\square$

(d) Proof:

Base case: when $t=0$, $\|w(t)\|^2 = \|w(0)\|^2 = 0 = tR^2$;

when $t=1$, using the conclusion from (c),

we have $\|w(1)\|^2 \le \|w(0)\|^2 + \|x(0)\|^2$

$\le R^2$ where $R = \max_{1\le n\le N}\|x_n\|$

Therefore, the base case holds.

Inductive case: suppose $\|w(t)\|^2 \le tR^2$ is true when $t=n$,

$$\begin{aligned} \text{then } \|w(n+1)\|^2 &\le \|w(n)\|^2 + \|x(n)\|^2 \\ &\le nR^2 + R^2 \\ &= (n+1)R^2 \end{aligned}$$

Therefore, we can conclude that $\|w(t)\|^2 \le tR^2$ where $R = \max_{1\le n\le N}\|x_n\|$.

$\square$

(e) From (d), we have $\|w(t)\|^2 \le tR^2$

$\|w(t)\| \le \sqrt{t}R$

From (b), we have $w^T(t)w^* \ge t\rho$,

then $\dfrac{w^T(t)w^*}{\|w(t)\|} \ge \dfrac{t\rho}{\sqrt{t}R} = \dfrac{\sqrt{t}\rho}{R}$

Since $|w^T(t)w^*| \le \|w^T(t)\|\,\|w^*\| = \|w(t)\|\,\|w^*\|$, $\dfrac{|w^T(t)w^*|}{\|w(t)\|} \le \|w^*\|$.

Since $\|w(t)\|, \dfrac{\sqrt{t}\rho}{R} \ge 0$, $|w^T(t)w^*| \ge 0$. Thus $|w^T(t)w^*| = w^T(t)w^*$.

Then $\|w^*\| \ge \dfrac{\sqrt{t}\rho}{R}$

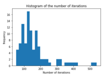
$\|w^*\|^2 \ge \dfrac{t\rho^2}{R^2}$

$t \le \dfrac{R^2\|w^*\|^2}{\rho^2}$

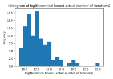
$\square$

2

(a)



(b)



From the above plot, we can see that all the actual number of iterations are much smaller than their corresponding theoretical bounds. Also we can observe that the graph is approximately normal but slightly right-skewed, which indicates that the mean of the log differences is greater than the median. In addition, the mean of the log difference is approximately between 11 and 13.

3(1.7)

(a) Given samples are generated independently. Let E denote the event that at least one coin will have $\nu = 0$.

$P(\mu=0.05, N=10, k=0) = \binom{10}{0}0.05^0(1-0.05)^{10} = 0.59874$

(i) when number of coins $=1$,

$P(E) = 0.59874$

when number of coins $=1000$,

$P(E) = 1-(1-0.59874)^{1000} = 1$

when number of coins $=1000000$,

$P(E) = 1-(1-0.59874)^{10^6} = 1$

$P(\mu=0.8, N=10, k=0) = \binom{10}{0}0.8^0(1-0.8)^{10} = 1.024\times10^{-7}$

(ii) when number of coins $=1$:

$P(E) = 1.024\times10^{-7}$

when number of coins $=1000$:

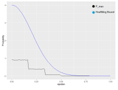
$P(E) = 1-(1-1.024\times10^{-7})^{1000} = 1.0239\times10^{-4}$

when number of coins $=1000000$:

$P(E) = 1-(1-1.024\times10^{-7})^{10^6} = 0.09733$

(b) Given $N=6$, # of coins $=2$, $\mu=0.5$, and since samples are generated independently,

$$\begin{aligned} P(\max_i|\nu_i-\mu_i|>\varepsilon) &= P((\nu_1-\mu_1>\varepsilon)\cup(\nu_2-\mu_2>\varepsilon)) \\ &= P(\nu_1-\mu_1>\varepsilon) + P(\nu_2-\mu_2>\varepsilon) - P((\nu_1-\mu_1)>\varepsilon \cap (\nu_2-\mu_2)>\varepsilon) \\ &= P(\nu_1-\mu_1>\varepsilon) + P(\nu_2-\mu_2>\varepsilon) - P(\nu_1-\mu_1>\varepsilon)\,P(\nu_2-\mu_2>\varepsilon) \\ &\le 2e^{-2\cdot6\varepsilon^2}\cdot2 \\ &= 4e^{-12\varepsilon^2} \end{aligned}$$



4(1.8)

(a) Proof: define a nonnegative random variable Y by:

$$Y = \begin{cases} \alpha, & t \ge \alpha \\ 0, & t < \alpha \end{cases}, \quad \alpha > 0$$

Then $Y \le t$, and take the expectation of both sides, $E[Y] \le E[t]$. Since $E[Y] = \alpha\cdot P(t\ge\alpha)$, we have $\alpha\cdot P(t\ge\alpha) \le E[t]$, and $P(t\ge\alpha) \le \dfrac{E[t]}{\alpha}$ follows.

$\square$

(b) Proof: let X be a random variable such that $X = (u-\mu)^2$, and let $\alpha > 0$, then from (a)

$P(X\ge\alpha) \le \dfrac{E[X]}{\alpha}$

$$\begin{aligned} E[X] &= E[(u-\mu)^2] \\ &= E[u^2] + E[\mu^2] - 2E[u\mu] \\ &= E[u^2] + \mu^2 - 2E[u]^2 \\ &= E[u^2] - E[u]^2 + \mu^2 - E[u]^2 \\ &= \text{Var}(u) \\ &= \sigma^2 \end{aligned}$$

And $P((u-\mu)^2\ge\alpha) \le \dfrac{\sigma^2}{\alpha}$ follows.

$\square$

(c) Proof: since $u = \frac{1}{N}\sum_{n=1}^{N}u_n$,

$E[u] = \frac{1}{N}\sum_{n=1}^{N}E[u_n] = \frac{1}{N}\cdot N\mu = \mu$

$\text{Var}(u) = \frac{1}{N^2}\sum_{n=1}^{N}\text{Var}(u_n) = \frac{1}{N^2}\cdot N\sigma^2 = \frac{\sigma^2}{N}$

Let $\alpha > 0$. From (b), we have

$P((u-\mu)^2\ge\alpha) \le \dfrac{\text{Var}(u)}{\alpha} = \dfrac{\sigma^2}{N\alpha}$

$\square$

5(1.12)

(a) To minimize the in sample sum of squared deviations, we take the derivative of $E_{in}(h)$ with respect to h:

$\frac{\partial}{\partial h}E_{in}(h) = \frac{\partial}{\partial h}\sum_{n=1}^{N}(h-y_n)^2 = 2\sum_{n=1}^{N}(h-y_n)$

Then we set $\frac{\partial}{\partial h}E_{in}(h)$ to 0:

$2\sum_{n=1}^{N}(h-y_n) = 0$

$Nh - \sum_{n=1}^{N}y_n = 0$

$h = \dfrac{\sum_{n=1}^{N}y_n}{N}$

To check if $h = \frac{\sum_{n=1}^{N}y_n}{N}$ is the minimum:

$\frac{\partial}{\partial h}\left(2\sum_{n=1}^{N}(h-y_n)\right) = 2N > 0$

Therefore, $h = \frac{\sum_{n=1}^{N}y_n}{N}$ minimizes $E_{in}(h) = \sum_{n=1}^{N}(h-y_n)^2$

(b) Similar to (a), we take the derivative of $E_{in}(h)$ with respect to h, we get:

$\frac{\partial}{\partial h}E_{in}(h) = \frac{\partial}{\partial h}\sum_{n=1}^{N}|h-y_n| = \sum_{n=1}^{N}\frac{h-y_n}{|h-y_n|} = \sum_{n=1}^{N}\text{sign}(h-y_n)$

Then we set $\frac{\partial}{\partial h}E_{in}(h)$ to 0:

$\sum_{n=1}^{N}\text{sign}(h-y_n) = 0$

which happens only when the number of positive terms equals the number of negative terms so that they can cancel out.

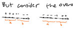
Therefore $h = \text{median}(y_1, y_2, \ldots, y_n)$

(c) Since h_{mean} is the average of the sum of all y_n's, if one of the y_n becomes $y_n+\varepsilon$, where $\varepsilon\to\infty$, then $h_{mean}\to\infty$ as well. But h_{median} would not be affected if N is large because as $\varepsilon\to\infty$, only $\max(y, y_1, \ldots, y_n)$ becomes this outlier, $\text{median}(y_1, y_2, \ldots, y_n)$ remains unchanged.

6(2.3)

(a) For positive rays and negative rays, they each have $m_{\mathcal{H}}(N) = N+1$, and therefore, if we remove the overcounting of all +'s and all −'s, $m_{\mathcal{H}}(N) = (N+1)\times2-2 = 2N$

(b) For positive intervals and negative intervals, they each have $m_{\mathcal{H}}(N) = \frac{N^2}{2}+\frac{N}{2}+1$. But consider the overcounting of the cases below:



where the points are divided into 2 parts with each part having all points being classified as all −'s or all +'s. Each case has $a+b=N$ counts, so $m_{\mathcal{H}}(N) = 2(\frac{N^2}{2}+\frac{N}{2}+1) - 2N = N^2-N+2$

(c) Since the segment between a and b are classified as +1 only, we can understand the problem as choosing two different points between d points and their edges, where the inside is +1 and the outside is −1. It results in d+1 points in total. In addition, we also need to consider the case where a=b such that all points are classified as −1.

Therefore, $m_{\mathcal{H}}(N) = \binom{N+1}{2}+1 = \frac{(N+1)!}{(N-1)!2!} = \frac{N(N+1)}{2}+1 = \frac{N^2}{2}+\frac{N}{2}+1$