王若飞 2021533135.

No.

Date. / /

HW3

1\. a) From Bayes rule, $P(X,Z|\theta) = \frac{P(Z|X,\theta)P(X,\theta)}{P(\theta)} = P(Z|X,\theta)P(X|\theta)$

$\ln P(X,Z|\theta) = \ln P(Z|X,\theta) + \ln P(X|\theta)$

$\ln P(X|\theta) = \ln P(X,Z|\theta) - \ln P(Z|X,\theta)$

$\ln P(X|\theta) = \ln\frac{P(X,Z|\theta)}{q(Z)} - \ln\frac{P(Z|X,\theta)}{q(Z)}$

$E_{Z\sim q}[\cdot]$

$\Rightarrow E_{Z\sim q}[\ln P(X|\theta)] = E_{Z\sim q}[\ln\frac{P(X,Z|\theta)}{q(Z)}] - E_{Z\sim q}[\ln\frac{P(Z|X,\theta)}{q(Z)}]$

$\Rightarrow \int_Z q(Z)\ln P(X|\theta)\,dZ = \ln P(X|\theta)\cdot\int_Z q(Z)\,dZ = \ln P(X|\theta)$

$\therefore \ln(P(X|\theta)) = E_{Z\sim q}[\ln\frac{P(X,Z|\theta)}{q(Z)}] \underline{- E_{Z\sim q}[\ln\frac{P(Z|X,\theta)}{q(Z)}]}$

$\Rightarrow$ From the definition of KL

$\therefore KL(q(Z)||P(Z|X,\theta)) = \int_Z q(Z)\ln\frac{q(Z)}{P(Z|X,\theta)}\,dZ = -\int_Z q(Z)\ln\frac{P(Z|X,\theta)}{q(Z)}\,dZ$

$\therefore$ Finally, we get $\ln P(X|\theta) = E_{Z\sim q}[\ln\frac{P(X,Z|\theta)}{q(Z)}] + KL(q(Z)||P(Z|X,\theta))$

$\Rightarrow \log P(X|\theta) = E_{Z\sim q}[\log\frac{P(X,Z|\theta)}{q(Z)}] + KL(q(Z)||P(Z|X,\theta))$

b) According to the non-negative property of KL, then $KL(q(Z)||P(Z|X,\theta)) \geq 0$

$\therefore \log P(X|\theta) \geq E_{Z\sim q}[\log\frac{P(X,Z|\theta)}{q(Z)}]$, when $q(Z) = P(Z|X,\theta)$. $=$ ✓

$\Rightarrow$ Then, every time we set $q(Z) = P(Z|X,\theta^{(t-1)})$, because we need to know $q(Z)$ and in this way, $KL(q(Z)||P(Z|X,\theta))$ is also $\geq 0$. so that we can get the lower bound easily.

$\therefore E_{Z\sim q}[\log\frac{P(X,Z|\theta)}{q(Z)}] = E_{Z|X,\theta^{(t-1)}}[\log\frac{P(X,Z|\theta)}{P(Z|X,\theta^{(t-1)})}]$

$\therefore \log P(X|\theta) \geq E_{Z|X,\theta^{(t-1)}}[\log\frac{P(X,Z|\theta)}{P(Z|X,\theta^{(t-1)})}]$ ✓

At the same time, for the first step, we can also use Jensen's inequality to prove

No.

Date. / /

(c) for ELBO we get in (b), $E_{Z|X,\theta^{(t-1)}}[\log \frac{P(X,Z|\theta)}{P(Z|X,\theta^{(t-1)})}]$

$= E_{Z|X,\theta^{(t-1)}}[\log P(X,Z|\theta)] - \underline{E_{Z|X,\theta^{(t-1)}}[\log P(Z|X,\theta^{(t-1)})]}$

for this term, it doesn't contain $\theta$, so we can seem it as a constant, because our goal is to find a correct $\theta$ ✓

$\therefore \Rightarrow \arg\max_\theta E_{Z|X,\theta^{(t-1)}}[\log P(X,Z|\theta)]$

$\therefore \arg\max_\theta E_{Z|X,\theta^{(t-1)}}[\log \frac{P(X,Z|\theta)}{P(Z|X,\theta^{(t-1)})}]$

& $\because Q(\theta|\theta^{(t-1)}) = E_{Z|X,\theta^{(t-1)}}[\log P(X,Z|\theta)]$

$\therefore \arg\max_\theta Q(\theta|\theta^{(t-1)}) = \arg\max_\theta E_{Z|X,\theta^{(t-1)}}[\log \frac{P(X,Z|\theta)}{P(Z|X,\theta^{(t-1)})}]$

$\therefore$ we get it.

Hw 3

3. a) ∵ output space is $\{-1, 1\}$

∴ $(x(t), y(t))$, which is misclassified, $y(t) = +1$ or $-1$

And because the example is misclassified, we have $\underbrace{y(t)}_{\text{true}} \neq \underbrace{\text{sign}(w^T(t)x(t))}_{\text{predict}}$

∴ when the sample is misclassified, $y(t)$ and $\text{sign}(w^T(t)x(t))$ will have different sign

And because sign function $\begin{cases} x>0, & 1 \\ x=0, & 0 \\ x<0, & -1 \end{cases}$ ∴ $y(t)$ and $w^T(t)x(t)$ will have different sign

∴ $y(t)w^T(t)x(t) < 0$, we get it.

b) $y(t)w^T(t+1)x(t) = y(t)(w(t) + y(t)x(t))^T x(t)$

$= y(t)w^T(t)x(t) + \underline{y(t)y(t)}x^T(t)x(t)$

$= y(t)w^T(t)x(t) + \underline{y^2(t)}\,x^T(t)x(t)$

∵ either $1 \cdot 1 = 1$
or $-1 \cdot -1 = 1$

$= \underbrace{y(t)w^T(t)x(t)}_{\text{scalar}} + \underbrace{x^T(t)x(t)}_{\text{scalar}}$ ⌊ $\|x(t)\|_2^2 \geq 0$

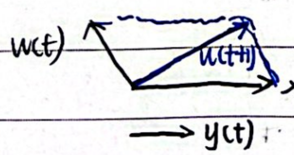
∴ $y(t)w^T(t+1)x(t) > y(t)w^T(t)x(t)$ & ∵ $x_0 = 1$ ∴ $> 0$ ✓

(c) ∵ only if when we disclassified the sample, we may update $w(t)$, that is $w(t+1) = w(t) + y(t) \cdot x(t)$

∴ ① $w(t) \cdot x \geq 0$, correctly classified

② $w(t) \cdot x < 0$, misclassified $\begin{cases} y(t) > 0, = 1 & \text{case } 1 \\ y(t) < 0, = -1 & \text{case } 2 \end{cases}$
L also denote $x(t)$

⇒ For case 1, $y(t) = 1$, $w(t+1) = w(t) + x(t)$

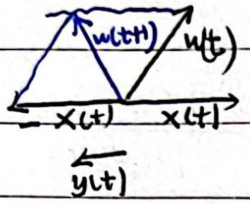



∴ $w(t) \cdot x(t) < 0$, but $\underline{w(t+1) \cdot x(t) > 0}$, that is next time, we will classify correctly. And In extreme case, even if $w(t+1)$ is still point to the inverse direction, but after round by round updating. finally it will point to the right direction step by step.

No.

Date. / /

$\Rightarrow$ For case 2. $w(t)\cdot x(t) < 0$, $y(t) = -1$, $w(t+1) = W(t)+(-x(t))$




$\therefore w(t)\cdot x(t) < 0$. but $w(t+1)\cdot x(t) > 0$.

$\therefore$ At first, we disclassified. but, every time we will make $w(t)$ towards the correct direction little by little too.

$\therefore$ Finally, we get it.

No.

Date.

HW3 For question in (a). we don't need to update $D_t(i)$ $\because$ for every $i$ $D_t(i) = \frac{1}{m} = \frac{1}{10}$

2.(a) ① $h^{(1)}(x_i)$: $\varepsilon_1 = \sum_{i=1}^{n} D_t(i) I(y_i \neq h_1(x_i))$

$= \frac{4}{10} = \frac{2}{5} = 0.4$

② $h^{(2)}(x_i)$: $\varepsilon_2 = \sum_{i=1}^{n} D_t(i) I(y_i \neq h_2(x_i))$

$= \frac{4}{10} = \frac{2}{5} = 0.4$

③ $h^{(3)}(x_i)$: $\varepsilon_3 = \sum_{i=1}^{n} D_t(i) I(y_i \neq h_3(x_i))$

$= \frac{5}{10} = \frac{1}{2} = 0.5$

④ $h^{(4)}(x_i)$: $\varepsilon_4 = \sum_{i=1}^{n} D_t(i) I(y_i \neq h_4(x_i))$

$= \frac{6}{10} = \frac{3}{5} = 0.4$

⑤ $h^{(5)}(x_i)$: $\varepsilon_5 = \sum_{i=1}^{n} D_t(i) I(y_i \neq h_5(x_i))$

$= \frac{6}{10} = \frac{3}{5} = 0.4$

⑥ $h^{(6)}(x_i)$: $\varepsilon_6 = \sum_{i=1}^{n} D_t(i) I(y_i \neq h_6(x_i))$

$= \frac{5}{10} = \frac{1}{2} = 0.5$

$\therefore$ for $h^{(1)}(x_i)$ and $h^{(2)}(x_i)$, they have min $\varepsilon = 0.4$
$h^{(4)}(x_i)$ and $h^{(5)}(x_i)$

(b) (1) choose $h^{(1)}$ as the first round classifier

$\alpha_1 = \frac{1}{2} \ln\left(\frac{1-\varepsilon_1}{\varepsilon_1}\right) = \frac{1}{2} \cdot \ln\left(\frac{1-0.4}{0.4}\right) \approx 0.2027$

(2) ① $\because D_{t+1}(i) = \frac{D_t(i)}{Z_t} e^{(-\alpha_t y_i h_t(x_i))}$

and $\because \sum_i D_{t+1}(i) = 1$

$\therefore Z_t = \sum_i D_t(i) e^{(-\alpha_t y_i h_t(x_i))}$

$= \sum_{i: y_i = h_t(x_i)} D_t(i) e^{-\alpha_t} + \sum_{i: y_i \neq h_t(x_i)} D_t(i) e^{\alpha_t}$

$= e^{-\alpha_t}\left(\sum_{i: y_i = h_t(x_i)} D_t(i)\right) + e^{\alpha_t}\left(\sum_{i: y_i \neq h_t(x_i)} D_t(i)\right)$

$= e^{-\alpha_t} \cdot (1-\varepsilon_t) + e^{\alpha_t} \varepsilon_t$

substitute $\alpha_t = 2\sqrt{\varepsilon_t(1-\varepsilon_t)}$

② Because $\alpha_t$ represents the weigh of this classifier in the final vote

$\therefore$ And $\prod_t Z_t$ represents all samples loss in t rounds in total

$\therefore \downarrow Z_t$, loss $\downarrow$

$\therefore \Rightarrow \min Z_t$ $\quad \nabla_{\alpha_t} Z_t = -e^{-\alpha_t}(1-\varepsilon_t) + e^{\alpha_t} \varepsilon_t = 0$

$e^{\alpha_t} \varepsilon_t - e^{-\alpha_t}(1-\varepsilon_t)$

No.

Date. / /

$$e^{2\alpha_t}\cdot\varepsilon_t - (1-\varepsilon_t) = 0$$

$$e^{2\alpha_t} = \frac{1-\varepsilon_t}{\varepsilon_t}$$

$$\alpha_t = \frac{1}{2}\ln\frac{1-\varepsilon_t}{\varepsilon_t}$$

∴ we get $\alpha_t = \frac{1}{2}\ln\frac{1-\varepsilon_t}{\varepsilon_t}$

(3) $D_1(i) = \frac{1}{m} = \frac{1}{10}$ for every i, and in the first round

$x_1, x_7, x_8, x_9$ will be misclassified

| | | | | |
|---|---|---|---|---|
| predict | -1 | 1 | 1 | 1 |
| true | 1 | -1 | -1 | -1 |

$\Rightarrow Z_1 = 2\sqrt{\varepsilon_1(1-\varepsilon_1)} = 2\sqrt{\frac{6}{25}} = \frac{2\sqrt{6}}{5}$, $\alpha_1 = \frac{1}{2}\ln\frac{1-0.4}{0.4} = \frac{1}{2}\ln\frac{3}{2}$

$\therefore D_2(x_1) = \frac{D_1(x_1)}{Z_1}e^{\alpha_1} = \frac{1}{10}\cdot\frac{5}{2\sqrt{6}}\cdot\frac{\sqrt{6}}{2} = \frac{1}{8}$

$D_2(x_7) = \frac{D_1(x_7)}{Z_1}e^{\alpha_1} = \frac{1}{10}\cdot\frac{5}{2\sqrt{6}}\cdot\frac{\sqrt{6}}{2} = \frac{1}{8}$

$D_2(x_8) = \frac{D_1(x_8)}{Z_1}e^{\alpha_1} = \frac{1}{10}\cdot\frac{5}{2\sqrt{6}}\cdot\frac{\sqrt{6}}{2} = \frac{1}{8}$

$D_2(x_9) = \frac{D_1(x_9)}{Z_1}e^{\alpha_1} = \frac{1}{10}\cdot\frac{5}{2\sqrt{6}}\cdot\frac{\sqrt{6}}{2} = \frac{1}{8}$

∴ these points new value in $D_2(\ )$ is $\frac{1}{8}$, for the remaining correct points is $\frac{1}{12}$

(4) Because the weigh of $x_1, x_7, x_8, x_9$ ↑, and the remaining decrease

∴ $h_2$ $x_1 x_2 x_3 x_9$ new $\varepsilon_t$ in second round $\varepsilon_2 = \frac{1}{8}+\frac{1}{12}+\frac{1}{12}+\frac{1}{8} = \frac{20}{48}$

misclassified point $h_3$ $x_1 x_2 x_3 x_4 x_5$ $\varepsilon_3 = \frac{1}{8}+\frac{1}{12}+\frac{1}{12}+\frac{1}{12}+\frac{1}{12} = \frac{22}{48}$

$h_4$ $x_3 x_4 x_5 x_7$, $\varepsilon_4 = \frac{1}{12}+\frac{1}{12}+\frac{1}{12}+\frac{1}{8} = \frac{18}{48}$

$h_5$ $x_5 x_6 x_7 x_8$ $\varepsilon_5 = \frac{1}{12}+\frac{1}{12}+\frac{1}{8}+\frac{1}{8} = \frac{20}{48}$

$h_6$ $x_5 x_6 x_7 x_8 x_9$ $\varepsilon_6 = \frac{1}{12}+\frac{1}{8}+\frac{1}{8}+\frac{1}{8}+\frac{1}{12} = \frac{26}{48}$

∴ it is $h_4$ minimize $\varepsilon_2$ and it is equal to $\frac{18}{48} = \frac{3}{8} = 0.375$

(5) $H(x) = \text{sign}(h_1(x)\alpha_1 + \alpha_2 h_2(x))$, $\alpha_1 = \frac{1}{2}\ln\frac{3}{2}$, $\alpha_2 = \frac{1}{2}\ln\frac{5}{3}$

$\therefore H(x_1) = S[-\alpha_1+\alpha_2] = +1$ $H(x_6) = S[\alpha_1-\alpha_2] = -1$

$H(x_2) = S[\alpha_1+\alpha_2] = +1$ $H(x_7) = S[+\alpha_1+\alpha_2] = +1$ $\therefore \varepsilon = \frac{|1|+|1|}{10} = 0.4$

$H(x_3) = S[\alpha_1-\alpha_2] = -1$ $H(x_8) = S[+\alpha_1-\alpha_2] = -1$

$H(x_4) = S[\alpha_1-\alpha_2] = -1$ $H(x_9) = S[+\alpha_1-\alpha_2] = -1$ $\therefore \underset{0.4}{\varepsilon} = \underset{0.4}{\varepsilon_1}$ ✓

$H(x_5) = S[\alpha_1-\alpha_2] = -1$ $H(x_{10}) = S[\alpha_1-\alpha_2] = -1$

∴ the same