資管三 b06705002 黃啟宏

1.

1.

$E(x) = \bar{x}$ (mean of x)

$Var(x) = E((x-\bar{x})^2) = \sigma_x^2$

$E(w) = 0$ (mean of w)

$Var(w) = E((w-0)^2) = \sigma_w^2$

$$s_j^{(l)} = \sum_{i=1}^{d^{(l-1)}} w_{ij}^{(l)} x_i^{(l-1)}$$

$$= w_{1j}^{(l)} x_1 + w_{2j}^{(l)} x_2 + \dots + w_{dj}^{(l)} x_d$$

$$= \sum_{i=1}^{d^{(l-1)}} \underbrace{[w_{i1} x_i, w_{i2} x_i, \dots \dots]}$$

Because $w_{ij}^{(l)}$ are independent to each other and to all $x_i^{(l-1)}$, the covariance between $x_i^{(l-1)}$ and $w_{ij}^{(l)}$ would be 0.

$$cov(x, w) = E((x-\bar{x})(w-0)) = E(x \cdot w) - \bar{x} \cdot 0 = 0$$

From above, we can know that the mean of $s_j^{(l)}$ are $E(x \cdot w)$, which equals to $0 + \bar{x} \cdot 0$, which is 0 and independent to each other ※

2.

2.

$Var(s_j^{(l)})$

$= Var(x \cdot w) \cdot d^{(l-1)}$

$= (E(x^T x w w^T) - E(x \cdot w)^2) \cdot d^{(l-1)}$

$= (Var(x) Var(w) + Var(x) \cdot E(w)^2 + Var(w) E(x)^2) \cdot d^{(l-1)}$

$= (\sigma_x^2 \sigma_w^2 + \sigma_x^2 \cdot 0 + \sigma_w^2 \cdot \bar{x}^2) \cdot d^{(l-1)}$

$= (\sigma_x^2 + \bar{x}^2) \sigma_w^2 d^{(l-1)}$ ※

3.

3.

We know that $x_i^{(l-1)} = \max(s_i^{(l-1)}, 0)$

So, $$E[{x_i^{(l-1)}}^2] = \int_{-\infty}^{\infty} \max(s_i^{(l-1)}, 0)^2 \, p(s_i^{(l-1)}) \, ds_i^{(l-1)}$$

$$= \int_{0}^{\infty} (s_i^{(l-1)})^2 \, p(s_i^{(l-1)}) \, ds_i^{(l-1)}$$

($s_i^{(l-1)}$ are zero-mean, symmetric

$$= \frac{1}{2} \int_{-\infty}^{\infty} (s_i^{(l-1)})^2 \, p(s_i^{(l-1)}) \, ds_i^{(l-1)}$$

$$= \frac{1}{2} E[(s_i^{(l-1)})^2] \quad \#$$

4.

4.

$\mathrm{Var}(s_i^{(l)})$

$$= d^{(l-1)} \, \mathrm{Var}(w_{ij}^{(l)} x_i^{(l-1)})$$

$$= d^{(l-1)} \, \mathrm{Var}(w_{ij}^{(l)}) \, E[(x_i^{(l-1)})^2]$$

(by problem 3)

$$= d^{(l-1)} \, \mathrm{Var}(w_{ij}^{(l)}) \cdot \frac{1}{2} E[(s_i^{(l-1)})^2]$$

(by $s_i^{(l-1)}$ are zero-mean, symmetric random var)

$$= d^{(l-1)} \, \mathrm{Var}(w_{ij}^{(l)}) \, \frac{1}{2} \underbrace{E[(s_i^{(l-1)} - \underbrace{E(s_i^{(l-1)})}_{=0})^2]}_{\mathrm{Var}(s_i^{(l-1)})}$$

$$= d^{(l-1)} \underbrace{\mathrm{Var}(w_{ij}^{(l)})}_{=\sigma_w^2} \frac{1}{2} \mathrm{Var}(s_i^{(l-1)})$$

$$= \frac{d^{(l-1)}}{2} \sigma_w^2 \, \mathrm{Var}(s_i^{(l-1)}) \quad \#$$

5.

5、

$$\mathrm{Var}(s_i^{(l)})$$
$$= d^{(l-1)}\,\mathrm{Var}(w_{ij}^{(l)})\,E[x_i^{(l-1)^2}]$$
$$= \mathrm{Var}(s_i^{(l-1)})$$

$$\boxed{\begin{aligned} & E[x_i^{(l-1)^2}] \\ &= \int_{-\infty}^{\infty} \max(s_i^{(l-1)}, \alpha s_i^{(l-1)})^2\, p(s_i^{(l-1)})\, ds_i^{(l-1)} \\ &= \frac{1+\alpha}{2}\int_{-\infty}^{\infty} (s_i^{(l-1)})^2\, p(s_i^{(l-1)})\, ds_i^{(l-1)} \\ &= \frac{1+\alpha}{2} E[(s_i^{(l-1)})^2] = \frac{1+\alpha}{2}\mathrm{Var}(s_i^{(l-1)}) \end{aligned}}$$

$$d^{(l-1)}\,\mathrm{Var}(w_{ij}^{(l)})\,E[x_i^{(l-1)^2}]$$
$$= \underbrace{d^{(l-1)}\,\sigma_w^2\,\frac{1+\alpha}{2}}_{=1}\,\mathrm{Var}(s_i^{(l-1)}) = \mathrm{Var}(s_i^{(l-1)})$$

$$\frac{d^{(l-1)}\sigma_w^2(1+\alpha)}{2} = 1$$

$$1+\alpha = \frac{2}{d^{(l-1)}\sigma_w^2} \qquad \alpha = \frac{2}{d^{(l-1)}\sigma_w^2} - 1 \ \#$$

6.

6.

臺灣大學

$$V_1 = \beta V_0 + (1-\beta)\Delta_1 = \beta^0(1-\beta)\Delta_1$$
$$V_2 = \beta V_1 + (1-\beta)\Delta_2 = \beta(1-\beta)\Delta_1 + (1-\beta)\Delta_2$$
$$V_3 = \beta V_2 + (1-\beta)\Delta_3 = \beta^2(1-\beta)\Delta_1 + \beta(1-\beta)\Delta_2 + (1-\beta)\Delta_3$$
$$V_4 = \beta V_3 + (1-\beta)\Delta_4 = \beta^3(1-\beta)\Delta_1 + \beta^2(1-\beta)\Delta_2 + \beta(1-\beta)\Delta_3 + (1-\beta)\Delta_4$$

At $t=T$, the formula results in

$$V_T = \sum_{t=1}^{T} \underbrace{\beta^{T-t}(1-\beta)}_{=\alpha_t}\Delta_t$$

$$\alpha_t = \beta^{T-t}(1-\beta) \ \#$$

7.

1.

$\alpha_1 \le \frac{1}{2}$

$\Rightarrow \beta^{T-1}(1-\beta) \le \frac{1}{2} \longrightarrow$ smallest T

$\ln(\beta^{T-1}(1-\beta)) \le \ln\frac{1}{2}$

$\ln\beta^{T-1} + \ln(1-\beta) \le \ln 1 - \ln 2$

$\ln\beta^{T} - \ln\beta + \ln(1-\beta) \le \ln 1 - \ln 2$

$T\ln\beta - \ln\beta + \ln(1-\beta) \le -\ln 2$

$T\ln\beta \le \ln\beta - \ln(1-\beta) - \ln 2$

$T\ln\beta \le \ln\left(\frac{\beta}{2(1-\beta)}\right)$

We know that $0 < \beta < 1$, so $\ln\beta < 0$

$T\ln\beta \le \ln\left(\frac{\beta}{2(1-\beta)}\right)$

$T \ge \frac{\ln\left(\frac{\beta}{2(1-\beta)}\right)}{\ln\beta}$, the smallest T is $\frac{\ln\left(\frac{\beta}{2(1-\beta)}\right)}{\ln\beta}$ #

8.

8.

$$\alpha_t'$$

$$= \frac{\alpha_t}{\sum_{t=1}^{T}\alpha_t}$$

$$= \frac{\beta^{T-t}(1-\beta)}{\sum_{t=1}^{T}\beta^{T-t}(1-\beta)}$$

$$= \frac{\beta^{T-t}\cancel{(1-\beta)}}{\beta^{T-1}\cancel{(1-\beta)}+\beta^{T-2}\cancel{(1-\beta)}+\dots+\beta^{0}\cancel{(1-\beta)}}$$

$$= \frac{\beta^{T-t}}{\beta^{T-1}+\beta^{T-2}+\dots+\beta^{0}}$$

$$= \frac{1}{\beta^{t-1}+\beta^{t-2}+\dots+\beta^{t-T}}$$

$$= \frac{1}{\sum_{k=1}^{T}\beta^{t-k}}$$

9.

臺灣大學

$a_1' \le \frac{1}{2}$

$\Rightarrow \frac{1}{\sum_{k=1}^{T} \beta^{1-k}} \le \frac{1}{2}$ $\longrightarrow$ smallest T

$\frac{1}{\sum_{k=1}^{T} \beta^{1-k}} \le \frac{1}{2}$

$\sum_{k=1}^{T} \beta^{1-k} \ge 2$

$\beta \sum_{k=1}^{T} \beta^{-k} \ge 2$

$\beta \left(\frac{1}{\beta^1} + \frac{1}{\beta^2} + \dots + \frac{1}{\beta^T}\right) \ge 2$

$= \beta \sum_{n=0}^{T-1} \left(\frac{1}{\beta}\right)\left(\frac{1}{\beta}\right)^n \ge 2$

$\beta \left(\frac{1}{\beta}\right)\left(\frac{1-(\frac{1}{\beta})^T}{1-\frac{1}{\beta}}\right) \ge 2$

$\boxed{1 - \frac{1}{\beta} < 0}$ (>1; $0< \; <1$)

$1 - \left(\frac{1}{\beta}\right)^T \le 2\left(1 - \frac{1}{\beta}\right)$

$-\left(\frac{1}{\beta}\right)^T \le 1 - \frac{2}{\beta}$

$-\left(\frac{1}{\beta}\right)^{T-1} \le \beta - 2$

$(-1)(T-1)\ln\left(\frac{1}{\beta}\right) \le \ln(\beta-2)$

$(1-T)(\underbrace{\ln 1}_{0} - \ln\beta) \le \ln(\beta-2)$

$T - 1 \ge \frac{\ln(\beta-2)}{\ln\beta}$

$T \ge \frac{\ln(\beta-2)}{\ln\beta} + 1$

the smallest T is $\frac{\ln(\beta-2)}{\ln\beta} + 1$ #

10.

10.

$T = E_p[\|y - X(w\odot p)\|^2]$

$\frac{d}{dw}(T) = E_p[2(y - X(w\odot p))(-Xp)] = 0$

p is binary vector, whose expectation is 0.5. (between 0 and 1)

$= 2E_p[(X(w\odot p) - y)(Xp)] = 0$

$E_p[(Xw \odot Xp)Xp] - yE_p[Xp] = 0$

$E_p[(0.5w)^T X^T X \cdot \cancel{0.5}] = \cancel{0.5}\, yX$

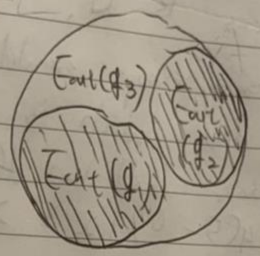
$w = 2yX^{-1}$ #

11.

11.

The best possible $E_{out}(G)$ is 0. $(0.08+0.16+0.32=.56<1)$.

The worst possible $E_{out}(G)$ is the coloring part below.



Because $\underset{\underset{0.32}{\text{|\!|}}}{E_{out}(g_3)} > \underset{\underset{0.08+0.16=0.24}{\text{|\!|}}}{E_{out}(g_1)+E_{out}(g_2)}$

So, the worst possible $E_{out}(G)$ is 0.24

$0 \le E_{out}(G) \le 0.24$ #

12.

12、

If a point is an error, then there should be at least $\frac{K+1}{2}$ wrong classifier.

So, the upper bound of $E_{out}(G)$ would be the sum of each error divided by $\frac{K+1}{2}$, which is $\frac{\sum_{k=1}^{K} e_k}{\frac{K+1}{2}} = \frac{2}{K+1}\sum_{k=1}^{K} e_k$ #

13.

13.

The probability that an example is sampled at least once equals to $N(1-p)$(an example is not sampled at all).

p(not sampled at all)

$= (1-\frac{1}{N})^{N'} = (1-\frac{1}{N})^{pN}$

While N is very large,

$(1-\frac{1}{N})^{pN}$

$= [(1-\frac{1}{N})^{N}]^{p} \Longrightarrow$

$= (e^{-1})^{p}$

$= e^{-p}$

Let $y = \lim_{N \to \infty} (1-\frac{1}{N})^{N}$

$\ln y = \lim_{N \to \infty} N \ln(1-\frac{1}{N})$

$\ln y = \lim_{N \to \infty} \frac{\ln(1-\frac{1}{N})}{N^{-1}}$

(with l'hospital's rule)

$= \lim_{N \to \infty} \frac{(\frac{1}{1-\frac{1}{N}})(0+N^{-2})}{-N^{-2}}$

$= \lim_{N \to \infty} -(\frac{1}{1-\frac{1}{N}})$

$= \lim_{N \to \infty} -(\frac{N}{N-1})$

$= -1$

$y = e^{-1}$

The approximately of the examples are sampled at least once is $N(1-p)$(not sampled at all)

$= N - N \cdot e^{-p}$

$= N - (e^{-p} \cdot N)$ ※

14.

14.

For each dimension, the number of stump should be between $[L, R]$, which is $R-L$.

So, the number of decision stumps should be $2d(R-L)$

However, there exist the situation which is all positive and all negative, so we $+2$. The answer would be $2d(R-L)+2$, which is $2 \cdot 4(5-0)+2 = 42$ ※

15.

15.

$$K_{ds}(x,x') = sign(x_1-\theta_1)\,sign(x'_1-\theta_1) + sign(x_2-\theta_2)\,sign(x'_2-\theta_2) + \dots$$
$$\dots\dots + sign(x_{|G|}-\theta_{|G|})\,sign(x'_{|G|}-\theta_{|G|})$$

if θ_t is between x_t and x'_t, $sign(x_t-\theta_t)\,sign(x'_t-\theta_t) = -1$

otherwise $=1$

There are $|x_t - x'_t|+1$ integers between x_t and x'_t, so there are $\underbrace{2(|x-x'|-1+1)}_{=\,2|x-x'|}$ results that equals -1

We know that $|G| = 2d(R-L)+2$, so the number of $sign(x_t-\theta_t)\,sign(x'_t-\theta_t)=1$ is $2d(R-L)+2-2|x-x'|$.

$$\text{So, } K_{ds}(x,x') = \overbrace{2d(R-L)+2-2|x-x'|}^{+1} - \overbrace{2|x-x'|}^{-1}$$
$$= 2d(R-L)+2-4|x-x'|$$

※

16.

16.

With infinite input vectors, $K_{ds}(x,x')$ should be

$$2d\lfloor R-L\rfloor + 2 - 4\lfloor x-x'\rfloor$$ ※

17.

The lecture I like most is activation in deep learning. Before I studied in this class, the most part that interest me is about neural networks. For me, face recognition is a very awesome thing to me. I feel surprised to predict a value or class with only just lots of data but not logics. Although I find at last that there is lots of derivation behind the models, it is still a cool thing. The part introduced the activation in different layer, which is the point of a neural network model.

18.

The lecture I like least is dual support vector machine. In that part, there is lots of Lagrange dual problem derivation, and I think the part was only mentioned quickly. To be honest, I ’ve never learned about that part before, so I really couldn’t understand it very well.