ML Techniques HW #3

B07902055 謝宗晅

1.

zero-mean：

$$\begin{aligned}
\mathbb{E}[s_j^{(l)}|x_i^{(l-1)}] &= \mathbb{E}\left[\sum_{i=1}^{d^{(l-1)}} w_{ij}^{(l)} x_i^{(l-1)} \middle| x_i^{(l-1)}\right] \\
&= \sum_{i=1}^{d^{(l-1)}} \mathbb{E}\left[w_{ij}^{(l)} \middle| x_i^{(l-1)}\right] \cdot x_i^{(l-1)} \\
&= \sum_{i=1}^{d^{(l-1)}} \mathbb{E}\left[w_{ij}^{(l)}\right] \cdot x_i^{(l-1)} \\
&\quad (\text{since all } w_{ij}^{(l)} \text{ and } x_i^{(l-1)} \text{ are independent}) \\
&= \sum_{i=1}^{d^{(l-1)}} 0 \cdot x_i^{(l-1)} \\
&= 0
\end{aligned}$$

independent to each other conditioned on $\mathbf{x}^{(l-1)}$：

因為所有的 $w_{ij}^{(l)}$ 之間、以及和所有 $x_i^{(l-1)}$ 都是 independent 的，所以有以下的性質：

$$F_{w_j,w_m|x}(w_{ij}^{(l)}, w_{km}^{(l)}) = F_{w|x}(w_{ij}^{(l)}) F_{w|x}(w_{km}^{(l)}), \text{where } i \neq k, j \neq m$$

因此

$$\begin{aligned}
F_{s_j,s_m|x}(s_j^{(l)}, s_m^{(l)}) &= F_{w_j,w_m|x}(w_{ij}^{(l)}, w_{km}^{(l)}) \\
&= F_{w|x}(w_{ij}^{(l)}) F_{w|x}(w_{km}^{(l)}) \\
&= F_{s|x}(s_j^{(l)}) F_{s|x}(s_m^{(l)}) \\
&\implies s_j^{(l)} \text{ and } s_m^{(l)} \text{ are independent conditioned on } \mathbf{x}
\end{aligned}$$

2.

$$
\begin{aligned}
\mathrm{Var}(s_j^{(l)}) &= \mathbb{E}\Big[\big(s_j^{(l)}\big)^2\Big] - \mathbb{E}\Big[s_j^{(l)}\Big]^2 \\
&= \mathbb{E}\Big[\big(s_j^{(l)}\big)^2\Big] - 0 \\
&= \mathbb{E}\Big[\Big(\sum_{i=1}^{d^{(l-1)}} w_{ij}^{(l)}\, x_i^{(l-1)}\Big)^2\Big] \\
&= \mathbb{E}\Big[\sum_{k=1}^{d^{(l-1)}} \Big(w_{kj}^{(l)} x_k^{(l-1)} \Big(\sum_{i=1}^{d^{(l-1)}} w_{ij}^{(l)}\, x_i^{(l-1)}\Big)\Big)\Big] \\
&= \mathbb{E}\Big[\sum_{k=1}^{d^{(l-1)}} \Big(w_{kj}^{(l)} x_k^{(l-1)} \Big(w_{kj}^{(l)} x_k^{(l-1)}\Big) + w_{kj}^{(l)} x_k^{(l-1)} \Big(\sum_{i\neq k} w_{ij}^{(l)}\, x_i^{(l-1)}\Big)\Big)\Big] \\
&= \mathbb{E}\Big[\sum_{k=1}^{d^{(l-1)}} \Big(w_{kj}^{(l)} x_k^{(l-1)}\Big)^2\Big] + \mathbb{E}\Big[\sum_{k=1}^{d^{(l-1)}} \Big(w_{kj}^{(l)} x_k^{(l-1)} \Big(\sum_{i\neq k} w_{ij}^{(l)}\, x_i^{(l-1)}\Big)\Big)\Big] \\
&= \mathbb{E}\Big[\sum_{k=1}^{d^{(l-1)}} \Big(w_{kj}^{(l)} x_k^{(l-1)}\Big)^2\Big] + 0 \\
&= \sum_{k=1}^{d^{(l-1)}} \mathbb{E}\Big[\Big(w_{kj}^{(l)} x_k^{(l-1)}\Big)^2\Big] \\
&= \sum_{k=1}^{d^{(l-1)}} \mathbb{E}\Big[\Big(w_{kj}^{(l)}\Big)^2\Big]\mathbb{E}\Big[\Big(x_k^{(l-1)}\Big)^2\Big] \\
&= \sum_{k=1}^{d^{(l-1)}} \Big(\sigma_w^2 + 0\Big)\Big(\sigma_x^2 + \bar{x}^2\Big) \\
&= d^{(l-1)}\Big(\sigma_w^2\sigma_x^2 + \sigma_w^2\bar{x}^2\Big)
\end{aligned}
$$

3.

因為 $s_i^{(l-1)}$ 是對稱的隨機變數，所以有下列的關係式：

$$
f(s_i^{(l-1)}) = f(-s_i^{(l-1)})
$$

其中 $f(\cdot)$ 是 $s_i^{(l-1)}$ 的機率密度函數。
並且因為上方的式子，所以有以下的關係式：

$$
\begin{aligned}
\mathbb{E}\Big[(s_i^{(l-1)})^2\Big] &= \int_{-\infty}^{\infty} (s_i^{(l-1)})^2 f(s_i^{(l-1)})\, ds_i^{(l-1)} \\
&= \int_{-\infty}^{0} (s_i^{(l-1)})^2 f(s_i^{(l-1)})\, ds_i^{(l-1)} + \int_{0}^{\infty} (s_i^{(l-1)})^2 f(s_i^{(l-1)})\, ds_i^{(l-1)} \\
&= \int_{0}^{\infty} \Big(-(s_i^{(l-1)})'\Big)^2 f\Big(-(s_i^{(l-1)})'\Big)\, d(s_i^{(l-1)})' + \int_{0}^{\infty} (s_i^{(l-1)})^2 f(s_i^{(l-1)})\, ds_i^{(l-1)} \\
&= \int_{0}^{\infty} (s_i^{(l-1)})^2 f(-s_i^{(l-1)})\, ds_i^{(l-1)} + \int_{0}^{\infty} (s_i^{(l-1)})^2 f(s_i^{(l-1)})\, ds_i^{(l-1)} \\
&= \int_{0}^{\infty} (s_i^{(l-1)})^2 f(s_i^{(l-1)})\, ds_i^{(l-1)} + \int_{0}^{\infty} (s_i^{(l-1)})^2 f(s_i^{(l-1)})\, ds_i^{(l-1)} \\
&= 2\int_{0}^{\infty} (s_i^{(l-1)})^2 f(s_i^{(l-1)})\, ds_i^{(l-1)}
\end{aligned}
$$

因此

$$
\begin{aligned}
\mathbb{E}\Big[(x_i^{(l-1)})^2\Big] &= \int_{-\infty}^{\infty} (x_i^{(l-1)})^2 g(x_i^{(l-1)})\, dx_i^{(l-1)} \\
&= \int_{0}^{\infty} (x_i^{(l-1)})^2 g(x_i^{(l-1)})\, dx_i^{(l-1)} \\
&\quad (\text{since } x_i^{(l-1)} \geq 0) \\
&= \int_{0}^{\infty} (s_i^{(l-1)})^2 f(s_i^{(l-1)})\, ds_i^{(l-1)} \\
&\quad (\text{since for } x_i^{(l-1)} \geq 0,\ x_i^{(l-1)} = s_i^{(l-1)} \text{ and } g(x_i^{(l-1)}) = f(s_i^{(l-1)})) \\
&= \frac{1}{2}\mathbb{E}\Big[(s_i^{(l-1)})^2\Big]
\end{aligned}
$$

(以上的 $g(\cdot)$ 是 $x_i^{(l-1)}$ 的機率密度函數)

4.

根據第一題、第二題和第三題得到的結果：

$$
\begin{aligned}
\mathbb{E}[s_j^{(l-1)}] &= 0 \\
\mathrm{Var}(s_j^{(l)}) &= d^{(l-1)}\Big(\sigma_w^2\sigma_x^2 + \sigma_w^2\bar{x}^2\Big) \\
\mathbb{E}\Big[(x_i^{(l-1)})^2\Big] &= \frac{1}{2}\mathbb{E}\Big[(s_i^{(l-1)})^2\Big]
\end{aligned}
$$

就可以得到以下的關係式：

$$
\begin{aligned}
\mathrm{Var}(s_j^{(l)}) &= d^{(l-1)}\left(\sigma_w^2\sigma_x^2 + \sigma_w^2\bar{x}^2\right) \\
&= d^{(l-1)}\left(\sigma_w^2\left(\mathbb{E}\left[(x_i^{(l-1)})^2\right]\right)\right) \\
&= \frac{d^{(l-1)}}{2}\left(\sigma_w^2\left(\mathbb{E}\left[(s_i^{(l-1)})^2\right]\right)\right) \\
&= \frac{d^{(l-1)}}{2}\left(\sigma_w^2\left(\mathbb{E}\left[(s_i^{(l-1)})^2\right] - 0\right)\right) \\
&= \frac{d^{(l-1)}}{2}\left(\sigma_w^2\left(\mathbb{E}\left[(s_i^{(l-1)})^2\right] - \left(\mathbb{E}[s_j^{(l-1)}]\right)^2\right)\right) \\
&= \frac{d^{(l-1)}}{2}\left(\sigma_w^2\left(\mathrm{Var}\left(s_i^{(l-1)}\right)\right)\right)
\end{aligned}
$$

5.

根據第三題的部份結果，我們知道：

$$
\begin{aligned}
\mathbb{E}\left[(s_i^{(l-1)})^2\right] &= 2\int_0^{\infty}(s_i^{(l-1)})^2 f(s_i^{(l-1)})\, ds_i^{(l-1)} \\
&= 2\int_{-\infty}^{0}(s_i^{(l-1)})^2 f(s_i^{(l-1)})\, ds_i^{(l-1)} \\
&\Longrightarrow \int_0^{\infty}(s_i^{(l-1)})^2 f(s_i^{(l-1)})\, ds_i^{(l-1)} \\
&= \int_{-\infty}^{0}(s_i^{(l-1)})^2 f(s_i^{(l-1)})\, ds_i^{(l-1)}
\end{aligned}
$$

所以

$$
\begin{aligned}
\mathbb{E}\left[(x_i^{(l-1)})^2\right] &= \int_{-\infty}^{\infty}(x_i^{(l-1)})^2 g(x_i^{(l-1)})\, dx_i^{(l-1)} \\
&= \int_{-\infty}^{0}(x_i^{(l-1)})^2 g(x_i^{(l-1)})\, dx_i^{(l-1)} + \int_0^{\infty}(x_i^{(l-1)})^2 g(x_i^{(l-1)})\, dx_i^{(l-1)} \\
&= \int_0^{\infty}(a\cdot s_i^{(l-1)})^2 f(s_i^{(l-1)})\, ds_i^{(l-1)} + \int_0^{\infty}(s_i^{(l-1)})^2 f(s_i^{(l-1)})\, ds_i^{(l-1)} \\
&\left(\text{since only one side of } x>0 \text{ and } x<0 \text{ makes } x = a\cdot s\right) \\
&= \frac{a^2+1}{2}\mathbb{E}\left[(s_i^{(l-1)})^2\right]
\end{aligned}
$$

又根據第四題的部份結果，有以下的式子：

$$
\begin{aligned}
\mathrm{Var}(s_j^{(l)}) &= d^{(l-1)}\left(\sigma_w^2\sigma_x^2 + \sigma_w^2\bar{x}^2\right) \\
&= d^{(l-1)}\left(\sigma_w^2\left(\mathbb{E}\left[(x_i^{(l-1)})^2\right]\right)\right) \\
&= d^{(l-1)}\left(\sigma_w^2\left(\frac{a^2+1}{2}\mathbb{E}\left[(s_i^{(l-1)})^2\right]\right)\right) \\
&= d^{(l-1)}\left(\sigma_w^2\left(\frac{a^2+1}{2}\left(\mathbb{E}\left[(s_i^{(l-1)})^2\right] + 0\right)\right)\right) \\
&= d^{(l-1)}\left(\sigma_w^2\left(\frac{a^2+1}{2}\left(\mathbb{E}\left[(s_i^{(l-1)})^2\right] + \mathbb{E}\left[s_i^{(l-1)}\right]^2\right)\right)\right) \\
&= d^{(l-1)}\left(\sigma_w^2\left(\frac{a^2+1}{2}\left(\mathrm{Var}\left(s_i^{(l-1)}\right)\right)\right)\right)
\end{aligned}
$$

所以如果我們令 $\mathbb{E}[w_{ij}^{(l)}] = 0$，$\mathrm{Var}(w_{ij}^{(l)}) = \frac{2}{d^{(l-1)}(a^2+1)}$ 的話，就會變成：

$$
\begin{aligned}
\mathrm{Var}(s_j^{(l)}) &= d^{(l-1)}\left(\sigma_w^2\left(\frac{a^2+1}{2}\left(\mathrm{Var}\left(s_i^{(l-1)}\right)\right)\right)\right) \\
&= d^{(l-1)}\left(\left(\sigma_w^2 + 0\right)\left(\frac{a^2+1}{2}\left(\mathrm{Var}\left(s_i^{(l-1)}\right)\right)\right)\right) \\
&= d^{(l-1)}\left(\left(\sigma_w^2 + \left(\mathbb{E}\left[w_{ij}^{(l)}\right]\right)^2\right)\left(\frac{a^2+1}{2}\left(\mathrm{Var}\left(s_i^{(l-1)}\right)\right)\right)\right) \\
&= d^{(l-1)}\left(\mathrm{Var}(w_{ij}^{(l)})\left(\frac{a^2+1}{2}\left(\mathrm{Var}\left(s_i^{(l-1)}\right)\right)\right)\right) \\
&= \mathrm{Var}\left(s_i^{(l-1)}\right) \\
&\left(\text{since } \mathrm{Var}(w_{ij}^{(l)}) = \frac{2}{d^{(l-1)}(a^2+1)}\right)
\end{aligned}
$$

所以只要將所有 $w_{ij}^{(l)}$ 做以上的初始化，就可以有題目所要求的性質。

6.

觀察一下展開的式子：

$$
\begin{aligned}
\mathbf{v}_1 &= (1-\beta)\boldsymbol{\Delta}_1 \\
\mathbf{v}_2 &= \beta(1-\beta)\boldsymbol{\Delta}_1 + (1-\beta)\boldsymbol{\Delta}_2 \\
\mathbf{v}_3 &= \beta^2(1-\beta)\boldsymbol{\Delta}_1 + \beta(1-\beta)\boldsymbol{\Delta}_2 + (1-\beta)\boldsymbol{\Delta}_3 \\
&\vdots
\end{aligned}
$$

就可以發現到以下的規律：

$$\mathbf{v}_T = \sum_{t=1}^{T} \alpha_t \boldsymbol{\Delta}_t$$
$$\alpha_t = \beta^{T-t}(1-\beta)$$

7.

承第六題，有以下的式子：

$$\begin{aligned}
&\alpha_1 = \beta^{T-1}(1-\beta) \le \frac{1}{2} \\
&\implies \beta^T \le \frac{\beta}{1-\beta} \cdot \frac{1}{2} \\
&\implies T \ge \log_\beta \frac{\beta}{2(1-\beta)} \\
&\text{smallest positive integer } T = \max\left(\left\lceil \log_\beta \frac{\beta}{2(1-\beta)} \right\rceil, 1\right)
\end{aligned}$$

8.

承第六題，α'_t 的關係式如下：

$$\begin{aligned}
\alpha'_t &= \frac{\alpha_t}{\sum_{t=1}^{T} \alpha_t} \\
\sum_{t=1}^{T} \alpha_t &= \beta^{T-1}(1-\beta) + \beta^{T-2}(1-\beta) + \cdots + \beta(1-\beta) + (1-\beta) \\
&= \frac{(1-\beta)(1-\beta^T)}{1-\beta} \\
&= 1 - \beta^T \\
\implies \alpha'_t &= \frac{\beta^{T-t}(1-\beta)}{1-\beta^T} \\
&= \frac{\beta^{T-t}}{1+\beta+\beta^2+\cdots+\beta^{T-1}} \\
&= \frac{\beta^{T-t}}{\sum_{i=0}^{T-1} \beta^i}
\end{aligned}$$

9.

承第八題，有以下的式子：

$$
\begin{aligned}
&\alpha_1' = \frac{\beta^{T-1}}{1+\beta+\beta^2+\cdots+\beta^{T-1}} \le \frac{1}{2} \\
\implies &\frac{\beta^{T-1}}{\frac{\beta^{T-1}(\beta^{-T}-1)}{\beta^{-1}-1}} \le \frac{1}{2} \\
\implies &\frac{\beta^{-1}-1}{\beta^{-T}-1} \le \frac{1}{2} \\
\implies &\beta^{-T}-1 \ge 2(\beta^{-1}-1) \\
\implies &\beta^{-T} \ge 2\beta^{-1}-1 \\
\implies &-T \le \log_\beta\left(2\beta^{-1}-1\right) \\
\implies &T \ge -\log_\beta\left(2\beta^{-1}-1\right)
\end{aligned}
$$

$$
\text{therefore, smallest positive integer } T = \max\left(\left\lceil -\log_\beta\left(2\beta^{-1}-1\right)\right\rceil, 1\right)
$$

10.

將題目中的式子展開：

$$
\begin{aligned}
&\min_{\mathbf{w}} \mathbb{E}_{\mathbf{p}}\Big((\mathbf{w}\odot\mathbf{p})^T X^T X(\mathbf{w}\odot\mathbf{p}) - 2\Big(X(\mathbf{w}\odot\mathbf{p})\Big)^T\mathbf{y} + \mathbf{y}^T\mathbf{y}\Big) \\
&= \min_{\mathbf{w}} \mathbb{E}_{\mathbf{p}}\Big(\sum_{i=1}^{N}\Big(\sum_{j=1}^{d} x_{i,j}w_j p_j\Big)^2 - 2\sum_{i=1}^{N}\sum_{j=1}^{d} x_{i,j}w_j p_j y_i\Big) + \mathbf{y}^T\mathbf{y} \\
&= \min_{\mathbf{w}} \mathbb{E}_{\mathbf{p}}\Big(\sum_{i=1}^{N}\Big(\sum_{j=1}^{d}\sum_{k=1}^{d} x_{i,j}x_{i,k}w_j w_k p_j p_k\Big)\Big) - \sum_{i=1}^{N}\sum_{j=1}^{d} x_{i,j}w_j y_i + \mathbf{y}^T\mathbf{y} \\
&= \min_{\mathbf{w}} \sum_{i=1}^{N}\Big(\sum_{j=1}^{d}\sum_{k=1}^{d} x_{i,j}x_{i,k}w_j w_k \mathbb{E}_{\mathbf{p}}\Big(p_j p_k\Big)\Big) - \sum_{i=1}^{N}\sum_{j=1}^{d} x_{i,j}w_j y_i + \mathbf{y}^T\mathbf{y} \\
&= \min_{\mathbf{w}} \sum_{i=1}^{N}\Big(\sum_{j=1}^{d}\sum_{k=1}^{d} x_{i,j}x_{i,k}w_j w_k \mathbb{E}_{\mathbf{p}}\Big(p_j p_k\Big)\Big) - \mathbf{w}^T X^T\mathbf{y} + \mathbf{y}^T\mathbf{y} \\
&= \min_{\mathbf{w}} \sum_{i=1}^{N}\Big(\sum_{j\neq k} x_{i,j}x_{i,k}w_j w_k \cdot \frac{1}{4} + \sum_{j=k} x_{i,j}x_{i,k}w_j w_k \cdot \frac{1}{2}\Big) - \mathbf{w}^T X^T\mathbf{y} + \mathbf{y}^T\mathbf{y} \\
&\quad (\text{since } \mathbb{E}[p_j] = \frac{1}{2}, \mathbb{E}[p_j p_k] = \frac{1}{4} (\text{where } j \neq k),\ \ \mathbb{E}[p_j^2] = \frac{1}{2}) \\
&= \min_{\mathbf{w}} \sum_{i=1}^{N}\Big(\frac{1}{4}\Big(\sum_{j=1}^{d} x_{i,j}w_j\Big)^2 + \frac{1}{4}\sum_{j=1}^{d}(x_{i,j}w_j)^2\Big) - \mathbf{w}^T X^T\mathbf{y} + \mathbf{y}^T\mathbf{y} \\
&= \min_{\mathbf{w}} \frac{1}{4}(\mathbf{w}^T X^T X\mathbf{w}) - \mathbf{w}^T X^T\mathbf{y} + \mathbf{y}^T\mathbf{y} + \frac{1}{4}\sum_{i=1}^{N}\sum_{j=1}^{d}(x_{i,j}w_j)^2
\end{aligned}
$$

將以上的式子對 w_i 做偏微分：

$$
\begin{aligned}
&\frac{\partial}{\partial w_i}\Big(\frac{1}{4}(\mathbf{w}^T X^T X\mathbf{w}) - \mathbf{w}^T X^T\mathbf{y} + \mathbf{y}^T\mathbf{y} + \frac{1}{4}\sum_{i=1}^{N}\sum_{j=1}^{d}(x_{i,j}w_j)^2\Big) \\
&= \frac{1}{2}(X^T X\mathbf{w})_i - (X^T\mathbf{y})_i + \frac{1}{2}\sum_{k=1}^{N} x_{k,i}^2 w_i
\end{aligned}
$$

令以上的式子等於 0，就可以得到 optimal $\mathbf{w}$：

$$\text{Let } Z = \frac{1}{2}(X^TX) + \frac{1}{2}\begin{bmatrix} \sum_{k=1}^N x_{k,1}^2 & 0 & \cdots & 0 \\ 0 & \sum_{k=1}^N x_{k,2}^2 & 0 & \vdots \\ \vdots & 0 & \ddots & 0 \\ 0 & \cdots & 0 & \sum_{k=1}^N x_{k,d}^2 \end{bmatrix}$$

$$\text{then the equation in previus section is equivalent to :}$$

$$\mathbf{0} = Z\mathbf{w} - X^T\mathbf{y}$$

$$\text{therefore the optimal } \mathbf{w} = Z^\dagger X^T\mathbf{y}$$

11.

這個 uniform blending classifier 要做錯的話，就必須有 2 個以上的 g_k 是做錯的，所以這個 G 的 E_{out} 最大的可能就是將 g_1 和 g_2 做錯的地方和 g_3 重疊，並且 g_1 和 g_2 做錯的地方不重疊，那麼就會有最大的 $E_{\text{out}} = 0.08 + 0.16 = 0.24$；而可能的最小 E_{out} 就是所有的 g_k 做錯的地方都不重疊，那麼就會有最小的 $E_{\text{out}} = 0$ (因為 $0.08 + 0.16 + 0.32 = 0.56 < 1$，所以有不重疊的可能)。

$$E_{\text{out}} \in [0, 0.24]$$

12.

承第十一題的想法，如果要最大化一個 G 的 E_{out}，就要有一半以上的 g_k 都是做錯的，因此只要有 $\frac{K+1}{2}$ 個 g_k 是做錯的話，G 就會做錯，因此最理想的情況就是，每一個 g_k 所做錯的區域都重疊了 $\frac{K+1}{2}$ 次，也就是說每一個 g_k 的每單位 E_{out} 都為 G 的 E_{out} 做了 $\frac{2}{K+1}$ 單位的貢獻 (因為要做出 1 單位 G 的 E_{out}，就必須有 $\frac{K+1}{2}$ 個 g_k 的 1 單位 E_{out})，所以將式子列出來就會變成：

$$E_{\text{out}}(G) \leq \sum_{k=1}^K e_k \cdot \frac{2}{K+1} = \frac{2}{K+1}\sum_{k=1}^K e_k$$

上述的情況是最佳的情況，因此是 $E_{\text{out}}(G)$ 的上界。

13.

至少被 sample 到一次的所有點的可能就是 " 所有點的數量 " 減掉 " 沒有被 sample 到的所有點的數量 "，而 " 一個點沒有被 sample 到的機率 " 就是：

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

而當 N 非常大的時候，$\left(1 - \frac{1}{N}\right)^{pN} \approx e^{-p}$ 因此所有沒有被 sample 到的點的數量就是：

$$N \cdot \left(1 - \frac{1}{N}\right)^{pN} \approx N \cdot e^{-p}$$

所以所有被 sample 到一次以上的點的數量就是：

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

14.

這題可以依照 θ 的範圍分為兩個部份來討論：
第一個部份是 $\theta \leq L$ 或是 $\theta > R$，也就是使得所有 $g_{s,i,\theta}(\mathbf{x}) = 1$ 或是所有 $g_{s,i,\theta}(\mathbf{x}) = -1$ 的，雖然 s 的不同也會影響到 output，但是只有「全都是 $+1$」和「全都是 -1」的兩種情況。
第二個部份是 $L < \theta \leq R$ 的情況，因為 $x_i \in \{0, \cdots, 5\}$，並且 $\text{sign}(0) = +1$，所以考慮單一一個維度，可以知道當 $n < \theta \leq n+1$ 時，都會使得

$$x_i \in \{n, n-1, \cdots\},\ s \cdot \text{sign}(x_i - \theta) = -s$$

以及

$$x_i \in \{n+1, n+2, \cdots\},\ s \cdot \text{sign}(x_i - \theta) = s$$

所以考慮單一維度的話，θ 總共可以做出 5 個不一樣的 decision stumps，並且因為 s 有兩種可能，維度 $d = 4$，因此所有可能的 decision stumps 的數量就是兩個部份的數量加總：

$$2 + 5 \times 2 \times 4 = 42$$

15.

先計算 $|\mathcal{G}|$，承第 14 題，可以觀察到在 (d, L, R) 的條件下，所有可能的 decision stumps 的數量是：

$$|\mathcal{G}| = 2 + 2d(R - L)$$

然後看看 $\phi_{ds}(\mathbf{x})$ 的第 t 維相乘的結果：

$$\begin{aligned} g_t(\mathbf{x}) g_t(\mathbf{x}') &= s_t \cdot \text{sign}(x_{i(t)} - \theta_t) \cdot s_t \cdot \text{sign}(x'_{i(t)} - \theta_t) \\ &= s_t^2 \text{sign}(x_{i(t)} - \theta_t) \cdot \text{sign}(x'_{i(t)} - \theta_t) \\ &= \text{sign}(x_{i(t)} - \theta_t) \cdot \text{sign}(x'_{i(t)} - \theta_t) \\ &\quad (\text{since every } s_t^2 = 1) \end{aligned}$$

其中 $x_{i(t)}$ 代表的是 g_t 所使用的 x_i，可以觀察到當

$$\min(x_{i(t)}, x'_{i(t)}) < \theta_t \leq \max(x_{i(t)}, x'_{i(t)})$$

則 $g_t(\mathbf{x}) g_t(\mathbf{x}') = -1$，如果再加上 $s_t \in \{-1, +1\}$ 進考慮的話，對於 $\mathbf{x}$ 和 $\mathbf{x}'$ 的第 i 維，就會使得以下這麼多種 decision stumps 的值是 -1 (因為所有 x_i 都是整數)：

$$2|x_i - x'_i|$$

再考慮每一個維度，則有以下這麼多種的 decision stumps 的值是 -1：

$$\sum_{k=1}^{d} 2|x_k - x'_k|$$

因此就有以下這麼多種的 decision stumps 的值是 $+1$：

$$|\mathcal{G}| - \sum_{k=1}^{d} 2|x_k - x'_k| = 2 + 2d(R-L) - \sum_{k=1}^{d} 2|x_k - x'_k|$$

所以把所有的值加總起來：

$$\begin{aligned} K_{ds}(\mathbf{x}, \mathbf{x}') &= +1\Big(2 + 2d(R-L) - \sum_{k=1}^{d} 2|x_k - x'_k|\Big) - 1\Big(\sum_{k=1}^{d} 2|x_k - x'_k|\Big) \\ &= 2 + 2d(R-L) - 4\sum_{k=1}^{d} |x_k - x'_k| \end{aligned}$$

16.

參考自 http://work.caltech.edu/~htlin/publication/doc/infkernel.pdf

因為這一題的 $|\mathcal{G}|$ 是無限大的，所以這題要使用積分來計算，因為對於每一個 θ_t，只要考慮在 $[L, R]$ 範圍之內的所有情況就好了，因為 decision stumps 的數量是無限大的，所以第 15 題先考慮邊際情況的方法會行不通：

$$
\begin{aligned}
K_{ds}(\mathbf{x}, \mathbf{x}') &= \sum_{s=+1,-1} \sum_{i=1}^{d} \int_{L}^{R} s^2 \cdot \text{sign}(x_i - \theta) \cdot \text{sign}(x_i' - \theta) d\theta \\
&= 2 \sum_{i=1}^{d} \int_{L}^{R} \text{sign}(x_i - \theta) \cdot \text{sign}(x_i' - \theta) d\theta \\
&= 2 \sum_{i=1}^{d} \left((R - L) - \int_{\min(x_i, x_i')}^{\max(x_i, x_i')} 1 \, d\theta - \int_{\min(x_i, x_i')}^{\max(x_i, x_i')} 1 \, d\theta \right) \\
&\quad (\text{the first and second components are for } +1 \text{ , the third one is for } -1) \\
&= 2 \sum_{i=1}^{d} \left((R - L) - 2|x_i - x_i'| \right) \\
&= 2 \sum_{i=1}^{d} (R - L) - 4 \sum_{i=1}^{d} |x_i - x_i'| \\
&= 2d(R - L) - 4 \sum_{i=1}^{d} |x_i - x_i'|
\end{aligned}
$$

17.

我最喜歡的課程是 4/10 後半段的 neural network，因為在還沒有聽過那堂課之前，對於 neural network 的想像是非常困難而且深奧的 (雖然實際上應該是非常困難且深奧的沒有錯)，但是在那一堂課之後，就覺得自己距離知道 neural network 在做什麼又更近了一點，好像終於學到了一點點就在生活周遭的機器學習的知識，因此是我這學期的課最喜歡的部份。

18.

我最不喜歡的課程是 5/1 的 CNN 課程，絕大部分的原因是因為教授在直播時的器材問題，會使得聲音一直有小小的爆音，聽起來蠻不舒服的，但是在課程的後半段就解決了這個問題，因此那堂課還是很不錯的，只是有美中不足的地方。