1. 因為每一層 layer 都是 fully connected，其中每一層 layer 都有一個常數節點不會受 weight 影響，第一層數量為 10，最後一層為 1，對於層數 L 的 neural network，總 weight 數量為：

$$\sum_{l=0}^{L-2} d^{(l)} * (d^{(l+1)} - 1) + d^{L-1}$$

Constraint 為$\sum_{l=1}^{L-1} d^{(l)} = 36$

這個不等式不是很容易求，只好使用窮舉法找出最小與最大 weight 數量，其程式碼如下：

```python
import numpy as np

maxw  = -1
minw = 10000
minstruc = []
maxstruc = []

def layer_structure( number, unitnum, hidden = []):
    global maxw
    global minw
    global minstruc
    global maxstruc
    if np.sum(hidden) != None:
        if np.sum(hidden) == unitnum:
            wnum = 10*(hidden[0]-1) + hidden[len(hidden) - 1]
            for i in range(len(hidden) - 1):
                wnum += hidden[i]*(hidden[i+1]-1)
            if maxw < wnum:
                maxw = wnum
                maxstruc = hidden
                #print('max:', maxw, maxstruc)
            if minw > wnum:
                minw = wnum
                minstruc = hidden
                #print('min:', minw, minstruc)
            return
        elif np.sum(hidden) > unitnum:
            return
    for i in number:
        layer_structure(number, unitnum,  hidden + [i])

maxw = -1
minw = 10000
number = np.arange(2, 36)
layer_structure(number, 36)

print('final min:', minw, minstruc)
print('final max:', maxw, maxstruc)
```

最後得出的最小 weight 數為 46，layer 排列為

{10, 2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,1}

2. 最後得出的最大 weight 數為 510，layer 排列為
   {10, 22, 14, 1}

3.

$$\nabla_w err_n(w) = \nabla_w\left(\|x_n - ww^Tx_n\|^2\right)$$
$$= \nabla_w\left(x_n^Tx_n - 2ww^Tx_n \cdot x_n + (ww^Tx_n)^2\right)$$
$$= -4w^Tx_n \cdot x_n + \nabla_w\left(x_n^Tww^T \cdot ww^Tx_n\right) \quad (x_n^Tw = w^Tx_n \text{ 都是一常數})$$
$$= -4w^Tx_n \cdot x_n + \nabla_w\left((x_n^Tw)^2(w^Tw)\right)$$
$$= -4w^Tx_n \cdot x_n + 2x_n^Tw \cdot (w^Tw) \cdot x_n + 2w \cdot (x_n^Tw)^2 \;\#$$

4.

$$E_{in}(w) = \frac{1}{N}\sum_{n=1}^{N}\|x_n - ww^T(x_n + \varepsilon_n)\|^2$$
$$= \frac{1}{N}\sum_{n=1}^{N}\left(x_n^Tx_n - 2(x_n+\varepsilon_n)^Tww^Tx_n + (x_n+\varepsilon_n)^Tww^Tww^T(x_n+\varepsilon_n)\right)$$
$$= \frac{1}{N}\left(\sum_{n=1}^{N}\left(x_n^Tx_n - 2x_n^Tww^Tx_n + x_n^Tww^Tww^Tx_n\right) + \sum_{n=1}^{N}\left(-2\varepsilon_n^Tww^Tx_n + \varepsilon_n^Tww^Tww^T\varepsilon_n + 2\varepsilon_n^Tww^Tww^Tx_n\right)\right)$$
$$= \frac{1}{N}\sum_{n=1}^{N}\|x_n - ww^Tx_n\|^2 + \sum_{n=1}^{N}\left(\left(-2ww^Tx_n + 2ww^Tww^Tx_n\right)\cdot\varepsilon_n + \varepsilon_n^Tww^Tww^T\varepsilon_n\right)$$

已知 $E(\varepsilon_n) = 0$, $E(\varepsilon_n\varepsilon_n^T) = I_n$

$$E(E_{in}(w)) = \frac{1}{N}\sum_{n=1}^{N}\|x_n - ww^Tx_n\|^2 + \cancel{\sum_{n=1}^{N}\left(-2ww^Tx_n + 2ww^Tww^Tx_n\right)E(\varepsilon_n)}^{0} + E\left(\sum_{n=1}^{N}\varepsilon_n^Tww^Tww^T\varepsilon_n\right)$$
$$= \frac{1}{N}\sum_{n=1}^{N}\|x_n - ww^Tx_n\|^2 + \frac{1}{N}\sum_{n=1}^{N}E\left(tr(\varepsilon_n^Tww^Tww^T\varepsilon_n)\right)$$
$$= \frac{1}{N}\sum_{n=1}^{N}\|x_n - ww^Tx_n\|^2 + \frac{1}{N}\sum_{n=1}^{N}tr\left(E(ww^T\varepsilon_n \cdot \varepsilon_n^Tww^T)\right)$$
$$= \frac{1}{N}\sum_{n=1}^{N}\|x_n - ww^Tx_n\|^2 + \frac{1}{N}\sum_{n=1}^{N}tr\left(ww^T\cancel{E(\varepsilon_n\varepsilon_n^T)}\,ww^T\right)$$
$$= \frac{1}{N}\sum_{n=1}^{N}\|x_n - ww^Tx_n\|^2 + \cancel{\frac{1}{N}}\cancel{N}(w^Tw)^2 \Rightarrow \Omega(w) = (w^Tw)^2 \;\#$$

5.

error function $E = \sum_{i=1}^{d} (g_i(x) - x_i)^2$

let $U = \begin{bmatrix} u_{11} & \cdots & u_{1d} \\ \vdots & & \vdots \\ u_{\hat{d}1} & \cdots & u_{\hat{d}d} \end{bmatrix} \Rightarrow g_i(x) = \sum_{j=1}^{\hat{d}} u_{ij} \tanh\left(\sum_{k=1}^{d} u_{kj} x_k\right)$

$$E = \sum_{i=1}^{d} \left( \sum_{j=1}^{\hat{d}} u_{ij} \tanh\left(\sum_{k=1}^{d} u_{kj} x_k\right) - x_i \right)^2$$

6.

for W: $E = \sum_{i=1}^{d} \left( \sum_{j=1}^{\hat{d}} w_{ji}^{(2)} \tanh\left(\sum_{k=1}^{d} w_{kj}^{(1)} x_k\right) - x_i \right)^2$

$$\frac{\partial E}{\partial w_{IJ}^{(1)}} = \sum_{i=1}^{d} \underbrace{2\left( w_{Ji}^{(2)} \tanh\left(\sum_{k=1}^{d} w_{kJ}^{(1)} x_k\right) - x_i \right)}_{=C} \cdot \partial\left( \sum_{j=1}^{d} w_{ji}^{(2)} \tanh\left(\sum_{k=1}^{d} w_{kj}^{(1)} x_k\right) - x_i \right) \Big/ \partial w_{IJ}^{(1)}$$

$$= \sum_{i=1}^{d} 2\left( w_{Ji}^{(2)} \tanh\left(\sum_{k=1}^{d} w_{kJ}^{(1)} x_k\right) - x_i \right) w_{Ji} \tanh'\left(\sum_{k=1}^{d} w_{kJ}^{(1)} x_k\right) \cdot \partial \sum_{k=1}^{d} w_{kJ}^{(1)} x_k \Big/ \partial w_{IJ}^{(1)}$$

$$= \sum_{i=1}^{d} 2\left( w_{Ji}^{(2)} \tanh\left(\sum_{k=1}^{d} w_{kJ}^{(1)} x_k\right) - x_i \right) w_{Ji} \tanh'\left(\sum_{k=1}^{d} w_{kJ}^{(1)} x_k\right) \cdot x_I$$

$$\frac{\partial E}{\partial w_{JI}^{(2)}} = 2\left( w_{JI}^{(2)} \tanh\left(\sum_{k=1}^{d} w_{kJ}^{(1)} x_k\right) - x_I \right) \partial\left( w_{JI}^{(2)} \tanh\left(\sum_{k=1}^{d} w_{kJ}^{(1)} x_k\right) - x_I \right) \Big/ \partial w_{JI}^{(2)}$$

$$= 2\left( w_{JI}^{(2)} \tanh\left(\sum_{k=1}^{d} w_{kJ}^{(1)} x_k\right) - x_I \right) \cdot \tanh\left(\sum_{k=1}^{d} w_{kJ}^{(1)} x_k\right)$$

$$\frac{\partial E}{\partial u_{IJ}} = \partial\left( \sum_{j=1}^{\hat{d}} u_{Ij} \tanh\left(\sum_{k=1}^{d} u_{kj} x_k\right) - x_I \right)^2 \Big/ \partial u_{IJ} + \partial\left( \sum_{i=1, i\neq I}^{d} \left( \sum_{j=1}^{\hat{d}} u_{ij} \tanh\left(\sum_{k=1}^{d} u_{kj} x_k\right) - x_i \right)^2 \right) \Big/ \partial u_{IJ}$$

$$= 2\left( u_{IJ} \tanh\left(\sum_{k=1}^{d} u_{kJ} x_k\right) - x_I \right) \cdot \left( \tanh\left(\sum_{k=1}^{d} u_{kJ} x_k\right) + u_{IJ} \cdot \tanh'\left(\sum_{k=1}^{d} u_{kJ} x_k\right) \cdot x_I \right)$$

$$+ \sum_{i=1, i\neq I}^{d} 2\left( u_{iJ} \tanh\left(\sum_{k=1}^{d} u_{kJ} x_k\right) - x_i \right) \cdot u_{iJ} \tanh'\left(\sum_{k=1}^{d} u_{kJ} x_k\right) \cdot x_I$$

$$= \sum_{i=1}^{d} 2\left( u_{iJ} \tanh\left(\sum_{k=1}^{d} u_{kJ} x_k\right) - x_i \right) \cdot u_{iJ} \tanh'\left(\sum_{k=1}^{d} u_{kJ} x_k\right) \cdot x_I + 2\left( u_{IJ} \tanh\left(\sum_{k=1}^{d} u_{kJ} x_k\right) - x_I \right) \tanh\left(\sum_{k=1}^{d} u_{kJ} x_k\right)$$

take $w_{ij}^{(1)}, w_{ji}^{(2)}$ as $u_{ij}$

$$\frac{\partial E}{\partial w_{IJ}^{(1)}} + \frac{\partial E}{\partial w_{JI}^{(2)}} = \sum_{i=1}^{d} 2\left( u_{iJ} \tanh\left(\sum_{k=1}^{d} u_{kJ} x_k\right) - x_i \right) \cdot u_{iJ} \tanh'\left(\sum_{k=1}^{d} u_{kJ} x_k\right) x_I + 2\left( u_{IJ} \tanh\left(\sum_{k=1}^{d} u_{kJ} x_k\right) - x_I \right) \tanh\left(\sum_{k=1}^{d} u_{kJ} x_k\right)$$

$$= \frac{\partial E}{\partial u_{IJ}} \quad \#$$

7.

兩個类的 1 Nearest Neighbor 即為求任意点離哪一個类比較近,
因此 hypothesis 為 2 类的中垂面.
對於平面上的点 $x$

$$(x^+ - x^-, 2)\cdot\left(x - \frac{x^+ + x^-}{2}, 0\right) = 0$$

$$\Rightarrow (x^+ - x^-)\cdot x - \frac{\|x^+\|^2 - \|x^-\|^2}{2} = 0$$

$$g_{\text{1NN}}(x) = \text{sign}(w^T x + b) = \text{sign}\left((x^+ - x^-)x - \frac{\|x^+\|^2 - \|x^-\|^2}{2}\right)$$

8.

已知 $\text{sign}(x) = \text{sign}(\alpha x)$ when $\alpha > 0$，且 $\exp(x) > 0$

$$g_{\text{RBFNET}}(x) = \text{sign}\left(\beta_+ \exp(-\|x-\mu_+\|^2) + \beta_- \exp(-\|x-\mu_-\|^2)\right)$$

$$= \text{sign}\left(\exp(+\|x-\mu_-\|^2)\cdot\left(\beta_+ \exp(-\|x-\mu_+\|^2) + \beta_- \exp(-\|x-\mu_-\|^2)\right)\right)$$

$$= \text{sign}\left(\beta_+ \exp(\|x-\mu_-\|^2 - \|x-\mu_+\|^2) + \beta_-\right)$$

$$= \text{sign}\left(\exp(\|x-\mu_-\|^2 - \|x-\mu_+\|^2) + \frac{\beta_-}{\beta_+}\right) \quad (\because \beta_+ > 0)$$

若 $\exp(\|x-\mu_-\|^2 - \|x-\mu_+\|^2) + \frac{\beta_-}{\beta_+} > 0$：

$$\exp(\|x-\mu_-\|^2 - \|x-\mu_+\|^2) > -\frac{\beta_-}{\beta_+}$$

$$\Longleftrightarrow \|x-\mu_-\|^2 - \|x-\mu_+\|^2 > \ln\left(-\frac{\beta_-}{\beta_+}\right)$$

$$\Longleftrightarrow \|x-\mu_-\|^2 - \|x-\mu_+\|^2 - \ln\left(-\frac{\beta_-}{\beta_+}\right) > 0$$

同理，若 $\exp(\|x-\mu_-\|^2 - \|x-\mu_+\|^2) + \frac{\beta_-}{\beta_+} \le 0 \Longleftrightarrow \|x-\mu_-\|^2 - \|x-\mu_+\|^2 - \ln\left(-\frac{\beta_-}{\beta_+}\right) \le 0$

$$\Rightarrow \text{sign}\left(\exp(\|x-\mu_-\|^2 - \|x-\mu_+\|^2) + \frac{\beta_-}{\beta_+}\right) = \text{sign}\left(\|x-\mu_-\|^2 - \|x-\mu_+\|^2 - \ln\left(-\frac{\beta_-}{\beta_+}\right)\right)$$

$$= \text{sign}\left((\cancel{x^T x} - 2\mu_-^T x + \mu_-^T\mu_-) - (\cancel{x^T x} - 2\mu_+^T x + \mu_+^T\mu_+) - \ln\left(-\frac{\beta_-}{\beta_+}\right)\right)$$

$$= \text{sign}\left(2(\mu_+ - \mu_-)^T x + \left(\mu_-^T\mu_- - \mu_+^T\mu_+ - \ln\left(-\frac{\beta_-}{\beta_+}\right)\right)\right) \#$$

9.

現在要 fix V 求 $E_m(W_m) = \sum_{(x_n, r_{nm}) \in D_m} (r_{nm} - W_m^T V_n)^2$ 的最小值.

$\nabla_{W_m} E_m(W_m) = \sum_{(x_n, r_{nm}) \in D_m} -2(r_{nm} - W_m^T V_n) V_n$

已知 $V_n = 1$，令上式為 0：

$\nabla_{W_m} E_m(W_m) = \sum_{(x_n, r_{nm}) \in D_m} -2(r_{nm} - W_m) = 0 \Rightarrow \sum_{(x_n, r_{nm}) \in D_m} r_{nm} = \sum_{(x_n, r_{nm}) \in D_m} W_m$

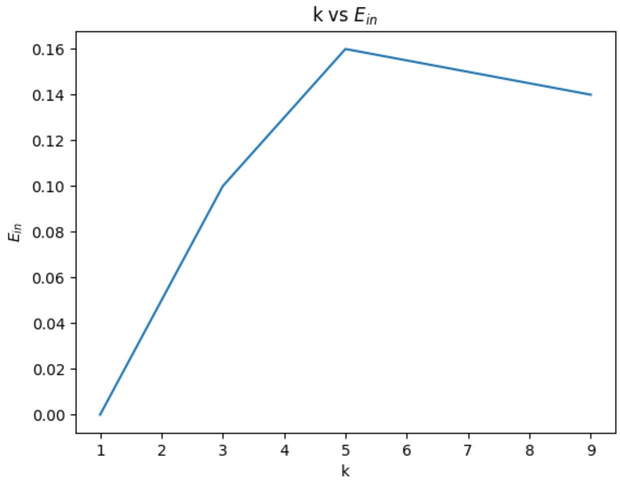
$W_m = \dfrac{\sum_{(x_n, r_{nm}) \in D_m} r_{nm}}{\sum_{(x_n, r_{nm}) \in D_m} 1}$ = m-th 電影的平均分數. #

10.

$V_{N+1}^T W_m = \frac{1}{N} \sum_{n=1}^{N} V_n^T \cdot W_m = \frac{1}{N} \sum_{n=1}^{N} r_{n,m}$

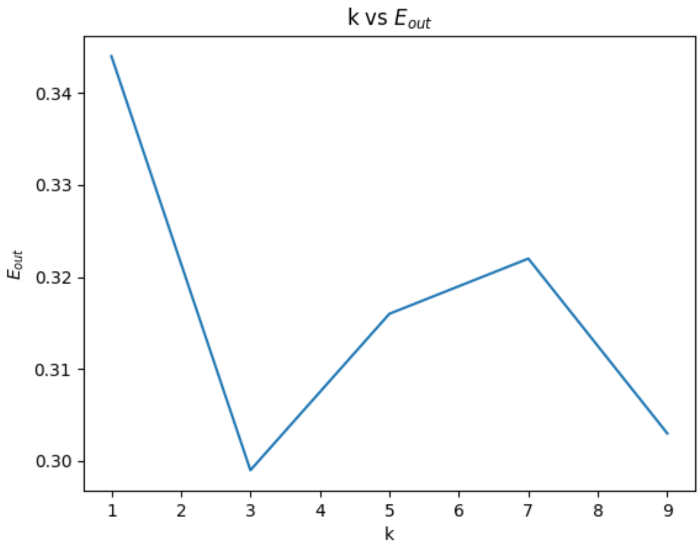
$\frac{1}{N} \sum_{n=1}^{N} r_{n,m}$ 最大的電影即為平均分數最高的電影 #

11.



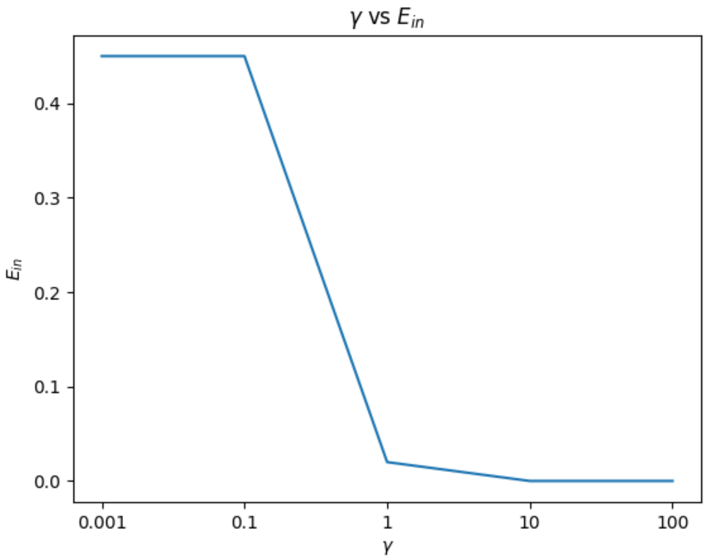

對於任意 sample 點 x，當 k = 1 時距離最近的點就是自己，所以 $E_{in}$ = 0。若 k != 1 則也會受周遭點的影響，因為不能保證 hw4_train.dat 數據 1 與-1 與附近分佈具相關性，因此 $E_{in}$ 跟 k 並沒有明顯的正負相關性。

12.



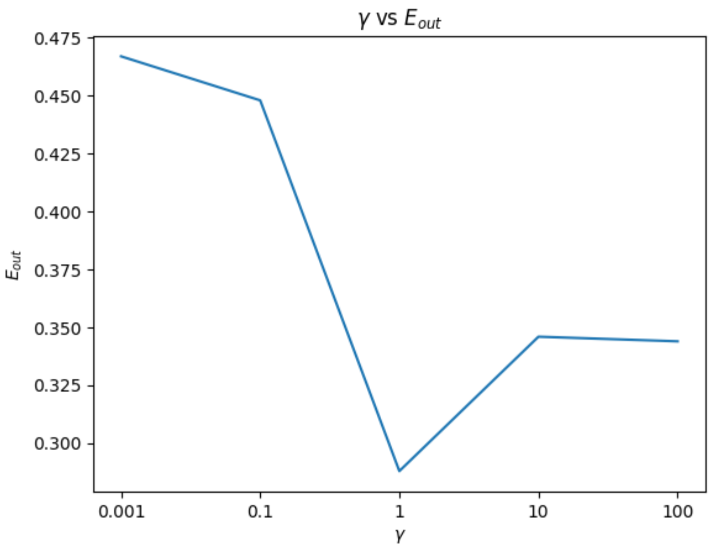

跟 $E_{in}$ 不同，當 k = 1 時 $E_{out}$ 反而最大，因為最近的點不一定會跟自己的 y 值一樣，可以看出 k Nearest Neighbor 的表現並不好，而且做 predict 也很吃運算資源。

13.



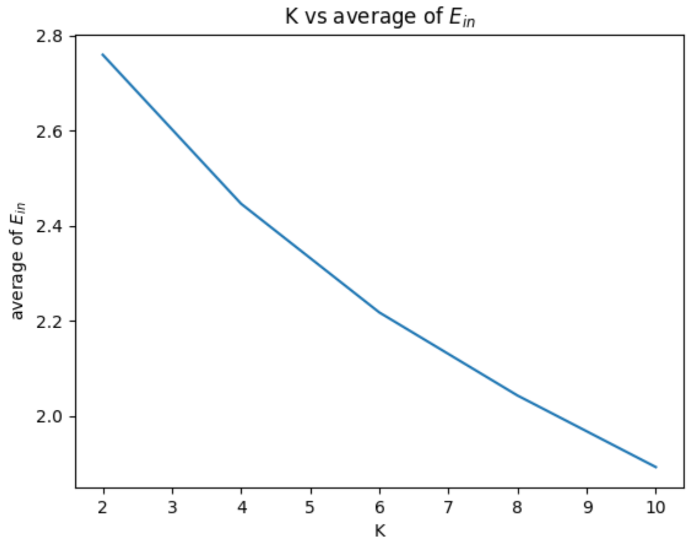

$g_{uniform}$ = sign($\sum_{m=1}^{N} y_m \exp(-\gamma\|x - x_m\|^2)$)
可以看出當$\gamma$越大 $E_{in}$ 越小。因為$\gamma$越大代表與 x 越接近的點所擁有的權重越大，而最接近 x 的 sample 即為 x 自己，所以當$\gamma$夠大時 x 會 dominant 整個決策，此時 $E_{in}$ = 0。

14.



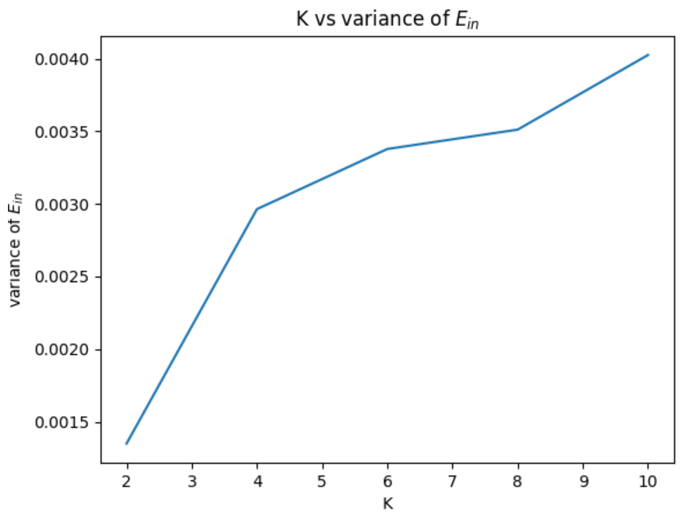

$\gamma$跟 $E_{out}$ 的相關性相對來說就沒有上題那麼明顯了，只知道$\gamma$越大 $g_{uniform}$ 的值也會跟 x 相近的點越接近。理論上來說會呈負相關，但沒有一定的嚴格遞減性。

15.




可以看出 K 越大平均 $E_{in}$ 越小，因為 K 越大能夠分越多群，$\mu_m$跟 $x_m$ 之間的平均距離當然也會越小。

16.




K 越大 $E_{in}$ 的 variance 就越大。因為當 K 很小時，分的群體多樣性也低，因此會收斂到差不多的位置，$E_{in}$ 也會差不多。當 K 變大時，群體多樣性增加，最後收斂的$\mu_m$位置受到起始 random 位置的影響也會變大，因此 $E_{in}$ 也較容易差很多。

17.

for $\Delta \geq 2$, $N \geq 3\Delta\log_2\Delta \Rightarrow N^\Delta + 1 < 2^N$

先證 $N^\Delta < 2^N$:

當 $N = 3\Delta\log_2\Delta$:

$N^\Delta < 2^N \Leftrightarrow N^\Delta < (2^{\log_2\Delta})^{3\Delta}$

$\Leftrightarrow N^\Delta < \Delta^{3\Delta} \Leftrightarrow N < \Delta^3$

$\Leftrightarrow 3\cancel{\Delta}\log_2\Delta < \Delta^{\cancel{3}2}$

當 $\Delta = 2$ $3 \cdot 1 < 2^2$ 成立

取 $f(\Delta) = \Delta^2 - 3\log_2\Delta$

$f'(\Delta) = 2\Delta - \frac{3}{\Delta\ln(2)} > 0$ for $\Delta > 2$

$\Rightarrow f(\Delta)$ 是嚴格遞增 for $\Delta > 2 \Rightarrow \Delta^2 - 3\log_2\Delta > 0 \Leftarrow 3\log_2\Delta < \Delta^2 \Leftrightarrow N^\Delta < 2^N$ for $\Delta \geq 2$ —①

$N^\Delta < 2^N \Leftrightarrow \Delta\ln N < N\ln 2$

取 $f(N) = \Delta\ln N - N\ln 2$

$\because N^\Delta < 2^N$ for $N = 3\Delta\log_2\Delta$

$\therefore f(3\Delta\log_2\Delta) < 0$ for $N = 3\Delta\log_2\Delta$

$f'(N) = \frac{\Delta}{N} - \ln 2 = 0 \Rightarrow N = \frac{\Delta}{\ln 2}$

$\Rightarrow$ 當 $N > \frac{\Delta}{\ln 2}$ $f'(N) = \frac{\Delta}{N} - \ln 2 < \frac{\Delta}{\frac{\Delta}{\ln 2}} - \ln 2 = 0$

$\Rightarrow f(N)$ 是嚴格遞減 for $N > \frac{\Delta}{\ln 2}$

又 $3\Delta\log_2\Delta = \frac{3\Delta\ln\Delta}{\ln 2} > \frac{\Delta}{\ln 2}$

$\Rightarrow f(N) \leq f(3\Delta\log_2\Delta) < 0$ for $N \geq 3\Delta\log_2\Delta$

$\Rightarrow N^\Delta < 2^N$

根據下題, $\Delta = 3(d+1)+1$, $N$ 為 sample 數, 兩者皆為整數, 因此 $N^\Delta + 1 < 2^N$ 也成立 #

18.

we know that bounding function $B(N,k)$:

maximum possible $m_H(N)$ when break point $= k$

且对 N 组数據，d 维 input，d 维 perceptrons 能表示组合数 $\le B(N, d+1)$

而在題中第 0 层到第 1 层任一節点即為一 d 维 perceptrons，

3 節点能表示组合数 $\le B(N, d+1)^3$

已知 $B(N,k) \le \sum_{i=0}^{k-1} C^N_i \le N^{k-1}+1$ (ML foundation hw 2)

$\Rightarrow$ 3 節点能表示组合数 $\le B(N, d+1)^3 \le (N^d+1)^3$

要證 $H_{3A}$ 的 VC dimension $\le 3(3(d+1)+1)\log_2(3(d+1)+1)$：

let $\Delta = 3(d+1)+1$ for $d \ge 0 \Rightarrow \Delta \ge 4 \ge 2$

let $N \ge 3\Delta\log_2\Delta$，則根據上題：

$N^{3(d+1)+1}+1 < 2^N$

$$3\text{節点能表示的组合数} \le B(N, d+1)^3 \le (N^d+1)^3$$
$$\le (N^{d+1}+1)^3 = N^{3(d+1)} + 3N^{2(d+1)} + 3N^{d+1} + 1$$

對於 $N \ge 3\Delta\log_2\Delta \ge 3\cdot 4\log_2 4 = 24$ ($\because \Delta \ge 4$)

$$\Rightarrow (N^{d+1}+1)^3 = N^{3(d+1)} + 3N^{2(d+1)} + 3N^{d+1} + 1 < N^{3(d+1)} + N^{3(d+1)} + N^{3(d+1)} + 1$$
$$= 3N^{3(d+1)} + 1 < N^{3(d+1)+1} + 1 < 2^N$$

$\Rightarrow$ $H_{3A}$ 無法 shatter $3\Delta\log_2\Delta$ 個点

VC dimension $< 3\Delta\log_2\Delta = 3(3(d+1)+1)\log_2(3(d+1)+1)$ #