1.
設有 k 層 hidden layer，每層有 d_k 個 units，即要求以下式子：

$$\min \prod_{i=0}^{i=k-1} d_i(d_{i+1}-1)\,,\text{其中}d_0=10, d_1+d_2+\cdots+d_k=36, d_i\geq 2$$

易知，當$d_1=d_2=\cdots=d_k=2$ 時(即每層均只有x_0和一個 unit)，會有最小值 $10+2*18=46$ 個 weights。
(註：因若將其中一個增加為 k，所需 weights 數量亦會增加；且通常在 neuron 數相同時，NN 深度和 weights 數量呈遞減趨勢)

2.
所求式子即為將上式之 min 改成 max。而 NN 若要有最多的 weight，hidden layer 的層數不是一層就是兩層(視 input units 和 hidden units 數量關係而定)。
一層 hidden layer：需要 $10*35+36*1=386$ 個 weights。
兩層 hidden layer：設兩層分別有$x, 36-x$ 個 units，$num=10(x-1)+x(35-x)+(36-x)=-x^2+44x+26$，故在$x=22$ 時有極值 $10*21+22*13+14=510$(個 weights)
由以上可知，最大值為 510 個 weights。

3.
題目表示：

$$\mathrm{err_n}(w)=\|\mathrm{x_n}-\mathrm{ww^T x_n}\|^2$$

故由 chain rule 可得其梯度為：

$$\nabla_{\mathrm{w}} err_n(w)=2(\mathrm{x_n}-\mathrm{ww^T x_n})(2\mathrm{w^T x_n})=4(\mathrm{x_n}-\mathrm{ww^T x_n})\mathrm{w^T x_n}$$

4.
題目給定以下：

$$\mathrm{E_{in}}(w)=\frac{1}{N}\sum_{n=1}^{N}\|\mathrm{x_n}-\mathrm{ww^T}(\mathrm{x_n}+\epsilon_n)\|^2$$

稍微展開可得：

$$\mathrm{E_{in}}(w)=\frac{1}{N}\sum_{n=1}^{N}\|\mathrm{x_n}-\mathrm{ww^T x_n}-\mathrm{ww^T}\epsilon_n\|^2$$

$$=\frac{1}{N}\sum_{n=1}^{N}\|\mathrm{x_n}-\mathrm{ww^T x_n}\|^2-2(\mathrm{x_n}-\mathrm{ww^T x_n})(\mathrm{ww^T}\epsilon_n)+(\mathrm{ww^T}\epsilon_n)^2$$

故題目要求之$\Omega(\mathrm{w})$為：

$$\Omega(\mathrm{w}) = \frac{1}{N}\sum_{n=1}^{N} -2(\mathrm{x_n} - \mathrm{ww^T x_n})(\mathrm{ww^T}\epsilon_n) + (\mathrm{ww^T}\epsilon_n)^2$$

又因為ϵ_n為 normal distribution 且平均為 0、Var( )為 1，故有以下性質：

$$\mathrm{E}[\epsilon_\mathrm{n}] = 0, \mathrm{E}[\epsilon_\mathrm{n}\epsilon_n^T] = \mathrm{I_n}$$

故取期望值後，$\mathrm{E}[\Omega(\mathrm{w})]$為：

$$\mathrm{E}[\Omega(\mathrm{w})] = \mathrm{E}[(\mathrm{ww^T}\epsilon_n)^T(\mathrm{ww^T}\epsilon_n)] = E[\epsilon_n^T \mathrm{ww^T ww^T}\epsilon_n]$$

因$\epsilon_n^T \mathrm{ww^T ww^T}\epsilon_n$為一個值，故可等價為$\mathrm{trace}(\epsilon_n^T \mathrm{ww^T ww^T}\epsilon_n)$

$$\mathrm{E}[\Omega(\mathrm{w})] = E\left[\mathrm{trace}\Big((\epsilon_n^T \mathrm{ww^T})\mathrm{ww^T}\epsilon_n\Big)\right] = E\left[\mathrm{trace}\Big((\mathrm{ww^T}\epsilon_n)\epsilon_n^T \mathrm{ww^T}\Big)\right]$$

$$= E\left[\mathrm{trace}\Big((\mathrm{ww^T})\mathrm{ww^T}\Big)\right] = E\left[\mathrm{trace}\Big(\mathrm{ww^T}(\mathrm{ww^T})\Big)\right]$$

$$= E[\mathrm{trace}(\mathrm{w^T}(\mathrm{ww^T})w)] = E[\mathrm{w^T}(\mathrm{ww^T})w] = \mathrm{w^T}(\mathrm{ww^T})w$$

故可知：

$$\Omega(\mathrm{w}) = (w^T w)^2$$

5.

題目定$u_{ij} = w_{ij}^{(1)} = w_{ji}^{(2)}$

定義 hidden layer units 的 input為 y：

$$\mathrm{y_j} = \sum_{i=1}^{d} u_{ij} x_i$$

定義 output layer units為 h(x)：

$$h(x_i) = \sum_{j=1}^{d} u_{ij} tanh\,(y_j)$$

故 error function E 可表示成：

$$\mathrm{E} = \sum_{i=1}^{d} (h(x_i) - x_i)^2 = \sum_{i=1}^{d}\left(\sum_{j=1}^{d} u_{ij}\, tanh(y) - x_i\right)^2$$

$$= \sum_{i=1}^{d}\left(\sum_{j=1}^{d} u_{ij}\, tanh\left(\sum_{i=1}^{d} u_{ij} x_i\right) - x_i\right)^2$$

6.

使用符號延續上題。

首先，將上題導出之式子中，u_{ij}還原成還未令$w_{ij}^{(1)} = w_{ji}^{(2)}$之時：

$$E = \sum_{i=1}^{d} (h(x_i) - x_i)^2$$

$$h(x_i) = \sum_{j=1}^{d} w_{ji}^{(2)} tanh\,(y_j)$$

$$\mathrm{y_j} = \sum_{i=1}^{d} w_{ij}^{(1)} x_i$$

(1) 對$w_{ij}^{(1)}$偏微：

$$\frac{\partial \mathrm{E}}{\partial w_{ij}^{(1)}} = 2\sum_{i=1}^{d}\left((h(x) - x_i)\frac{\partial h(x_i)}{\partial w_{ij}^{(1)}}\right)$$

$$= 2\sum_{i=1}^{d}\left((h(x) - x_i)\sum_{k=1}^{d}\left(w_{ki}^{(2)}\tanh'(y_k)\frac{\partial y_k}{\partial w_{ij}^{(1)}}\right)\right)$$

$$= 2\sum_{i=1}^{d}\left((h(x) - x_i) w_{ji}^{(2)}\, tanh'(y_j)\, x_i\right)$$

(2) 對$w_{ji}^{(2)}$偏微：

$$\frac{\partial \mathrm{E}}{\partial w_{ji}^{(2)}} = 2\sum_{i=1}^{d}\left((h(x) - x_i)\frac{\partial h(x_i)}{\partial w_{ji}^{(2)}}\right) = 2\sum_{i=1}^{d}\left((h(x) - x_i)\tanh\,(y_j)\right)$$

(3) 令$u_{ij} = w_{ij}^{(1)} = w_{ji}^{(2)}$時，對$u_{ij}$偏微：：

一樣先列出 E 和 y：

$$h(x_i) = \sum_{j=1}^{d} u_{ij} tanh\,(y_j)$$

$$\mathrm{y_j} = \sum_{i=1}^{d} u_{ij} x_i$$

對u_{ij}偏微：

$$\frac{\partial \mathrm{E}}{\partial u_{ij}} = 2\sum_{i=1}^{d}(h(x) - x_i)\frac{\partial h(x_i)}{\partial u_{ij}}$$

$$= 2\sum_{i=1}^{d}(h(x) - x_i)\left(tan\,h(y_j) + \sum_{k=1}^{d}\left(u_{ik} tanh'(y_k)\frac{\partial y_k}{\partial u_{ij}}\right)\right)$$

$$= 2\sum_{i=1}^{d}(h(x) - x_i)\left(tanh(y_j) + u_{ij} tanh'(y_j) x_i\right)$$

將(1), (2)的$w_{ij}^{(1)}$和$w_{ji}^{(2)}$換回u_{ij}：

$$\frac{\partial \mathrm{E}}{\partial w_{ij}^{(1)}} + \frac{\partial \mathrm{E}}{\partial w_{ji}^{(2)}} = 2\sum_{i=1}^{d}\left((h(x) - x_i)\left(w_{ji}^{(2)}\, tanh'(y_j)\, x_i + tanh(y_j)\right)\right)$$

$$= 2\sum_{i=1}^{d}\Big((h(x)-x_i)\big(u_{ij}\,tanh'(y_j)\,x_i + tanh(y_j)\big)\Big) = \frac{\partial \mathrm{E}}{\partial u_{ij}}$$

得證。

7.

中點為：$\frac{\mathrm{x}_+ + \mathrm{x}_-}{2}$，法向量為：$\mathrm{x}_+ - \mathrm{x}_-$

對於每個要預測的點 $\mathrm{x_p}$，預測方式為：

$$\mathrm{sign}((\mathrm{x}_+ - \mathrm{x}_-)^{\mathrm{T}}(\mathrm{x_p} - \frac{\mathrm{x}_+ + \mathrm{x}_-}{2}))$$

展開可得：

$$\mathrm{sign}\left((\mathrm{x}_+ - \mathrm{x}_-)^{\mathrm{T}}\mathrm{x_p} - \frac{(\mathrm{x}_+ - \mathrm{x}_-)^{\mathrm{T}}(\mathrm{x}_+ + \mathrm{x}_-)}{2}\right)$$

$$= \mathrm{sign}\left((\mathrm{x}_+ - \mathrm{x}_-)^{\mathrm{T}}\mathrm{x_p} - \frac{\mathrm{x}_+^{\mathrm{T}}\mathrm{x}_+ - \mathrm{x}_-^{\mathrm{T}}\mathrm{x}_-}{2}\right)$$

8.

當$\mathrm{g_{RBFNET}}(\mathrm{x}) = 0$ 時：

$$\beta_+ e^{-\|\mathrm{x}-\mu_+\|^2} + \beta_- e^{-\|\mathrm{x}-\mu_-\|^2} = 0$$

$$-\frac{\beta_+}{\beta_-} = e^{\|\mathrm{x}-\mu_+\|^2 - \|\mathrm{x}-\mu_-\|^2}$$

$$\|\mathrm{x}-\mu_+\|^2 - \|\mathrm{x}-\mu_-\|^2 = \ln(\beta_+) - \ln(-\beta_-)$$

$$-2\mathrm{x}^{\mathrm{T}}(\mu_+ - \mu_-) + (\mu_+^T\mu_+ - \mu_-^T\mu_-) = \ln(\beta_+) - \ln(-\beta_-)$$

全部移至等號左邊並變號，即可推得判斷式$\mathrm{g_{LIN}}(\mathrm{x})$：

$$\mathrm{g_{LIN}}(\mathrm{x}) = sign(2(\mu_+ - \mu_-)^{\mathrm{T}}\mathrm{x} - \ (\mu_+^T\mu_+ - \mu_-^T\mu_-) + \ln(\beta_+) - \ln(-\beta_-))$$

9.

因$\tilde{\mathrm{d}} = 1$，故設$\mathrm{V} = [v_1, v_2, \dots, v_N]^{\mathrm{T}}, \mathrm{W} = [w_1, w_2, \dots, w_M]^{\mathrm{T}}$

由 slide p.7 可推得，我們要最小化之 loss function 為：

$$\mathrm{E_{in}}(\{w\},\{v\}) = \sum (r_{nm} - w_m v_n)^2$$

步驟 2.1 為 optimize w，對w_m偏微可得：

$$\sum_{i=1}^{N}(r_{im} - w_m v_i)v_i = 0$$

$$w_m = \frac{v_1 r_{1m} + v_2 r_{2m} + \cdots + v_N r_{Nm}}{v_1^2 + v_2^2 + \cdots + v_N^2}$$

又因 V 初始化為全為 1 的矩陣，故帶入可得$\mathrm{w_m}$為：

$$w_m = \frac{r_{1m} + r_{2m} + \ldots + r_{Vm}}{N}$$

即為每人對該電影的平均評價。

10.

Predict score $\mathrm{v}_{\mathrm{N+1}}^{\mathrm{T}} w_m$可表達成：

$$\mathrm{v}_{\mathrm{N+1}}^{\mathrm{T}} w_m = \frac{1}{N}(\mathrm{v}_1 + \mathrm{v}_2 + \cdots + \mathrm{v}_{\mathrm{N}})^T w_m = \frac{1}{\mathrm{N}}(v_1^T w_m + v_2^T w_m + \cdots + v_N^T w_m)$$

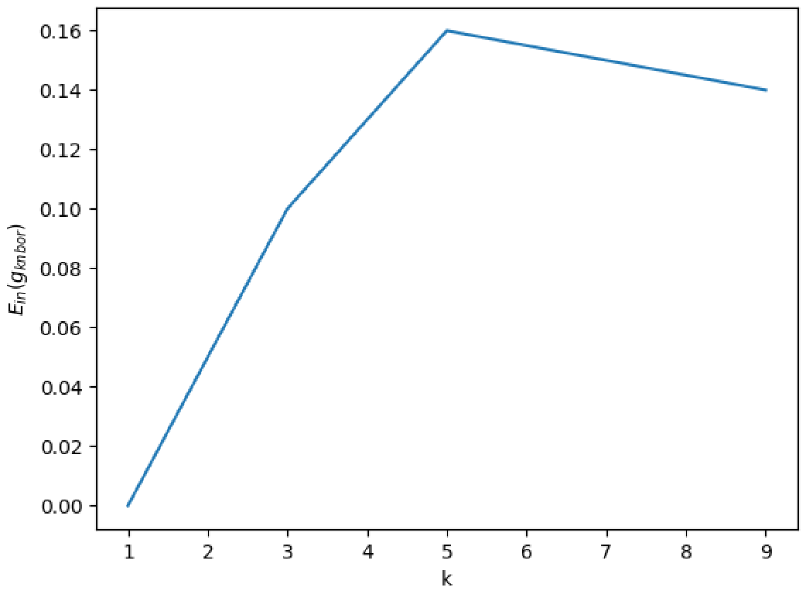
又因題目說已得到 perfect matrix factorization，即$v_i^T w_j = r_{ij}$

故以上式子可轉成：

$$\mathrm{v}_{\mathrm{N+1}}^{\mathrm{T}} w_m = \frac{1}{\mathrm{N}}(r_{1m} + r_{2m} + \cdots + r_{Nm})$$

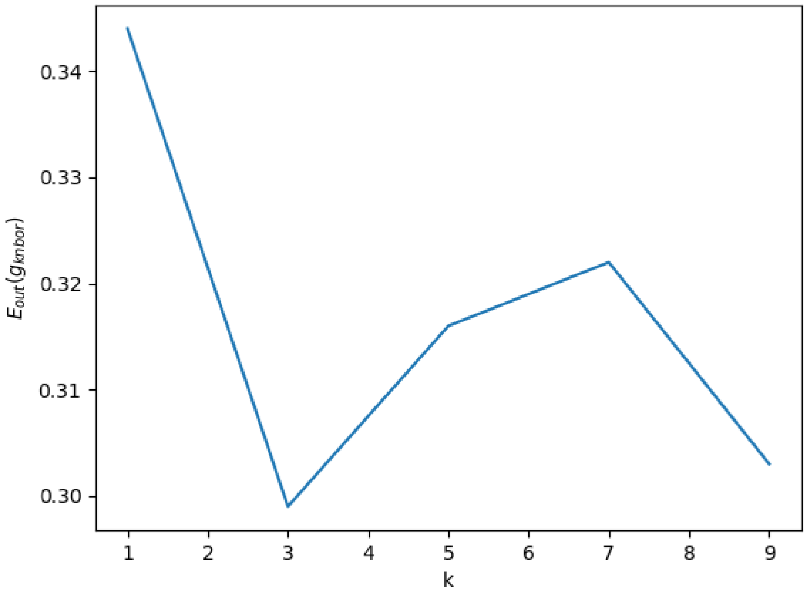
可知對 m 電影的 predict score 為所有 N 個人對其評分的平均。故推薦最大值的 $\mathrm{v}_{\mathrm{N+1}}^{\mathrm{T}} w_m$，即表示推薦電影平均評分之最大值。

11.



可以看出隨著 k 增加，$\mathrm{E_{in}}$會先上升再下降。而因為在計算 k 個最近的鄰居時，我使用的實行方式會將其本身的點考慮進去，故在$\mathrm{k} = 1$ 時，$\mathrm{E_{in}} = 0$。

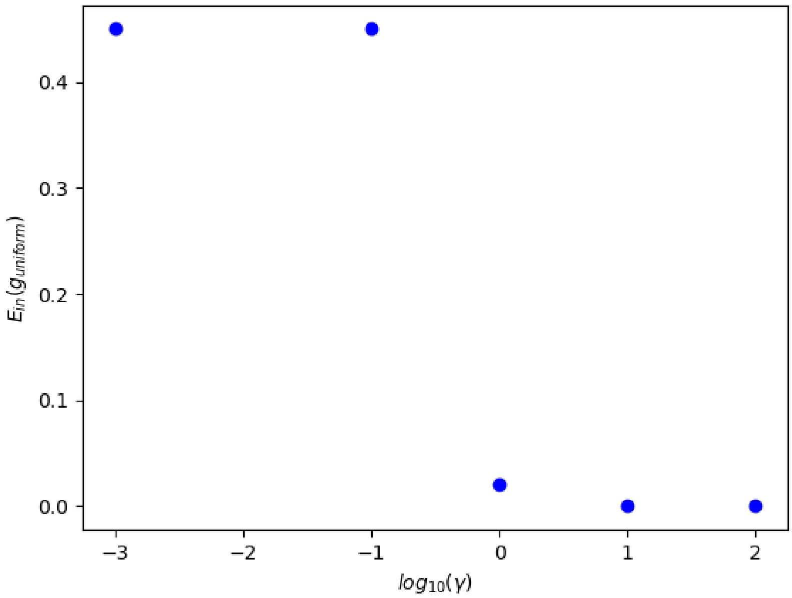
12.



在 $k \in [1, 9]$時，可以看出除了 $k = 1$ 時 E_{out}特別高之外，其餘的E_{out}走勢相當的不穩定。而E_{out}最低落在$k = 3$ 時，其值為 0.3 附近。0.3 可以算是個相當大的誤差，故可以看出使用 K-nearest 這種算法的表現並不會太好。

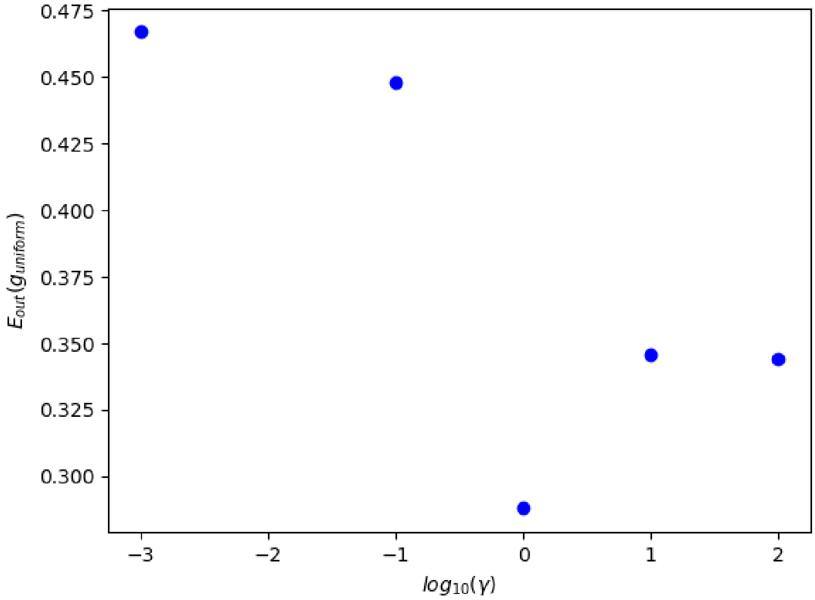
13.

註：13, 14 題因為使用 γ 當 x 軸做圖會使有些點重合，故特別修正成以$\log(\gamma)$做為 x 軸，使圖較易讀一點。



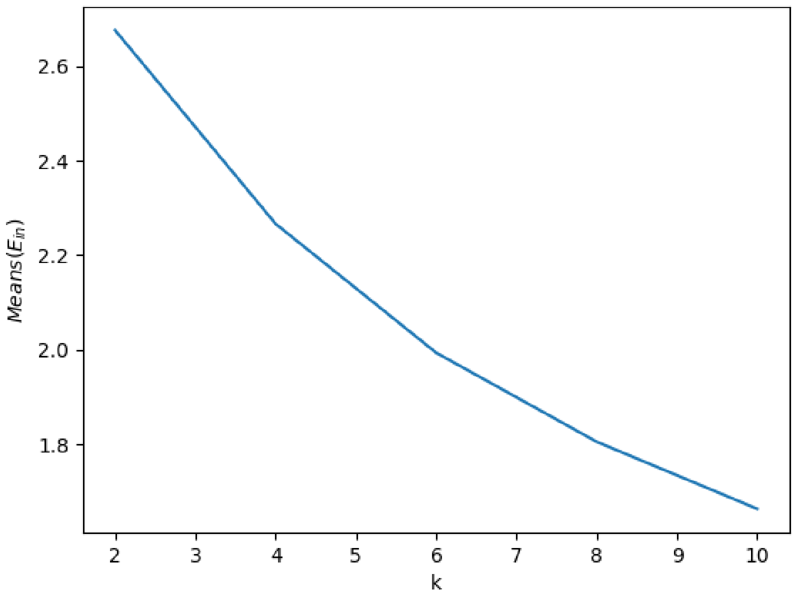
可以看出，隨著 γ 的增加，E_{in}呈現遞減的狀況。而且在 $\gamma = [0.1, 1]$ 這個區間中，E_{in} 變化幅度相當大。而和上題一樣，因為會把本身的點也納入考慮，故在圖中可看到，可以期望在 $\gamma \geq 10$ 時，$E_{in} = 0$。

14.



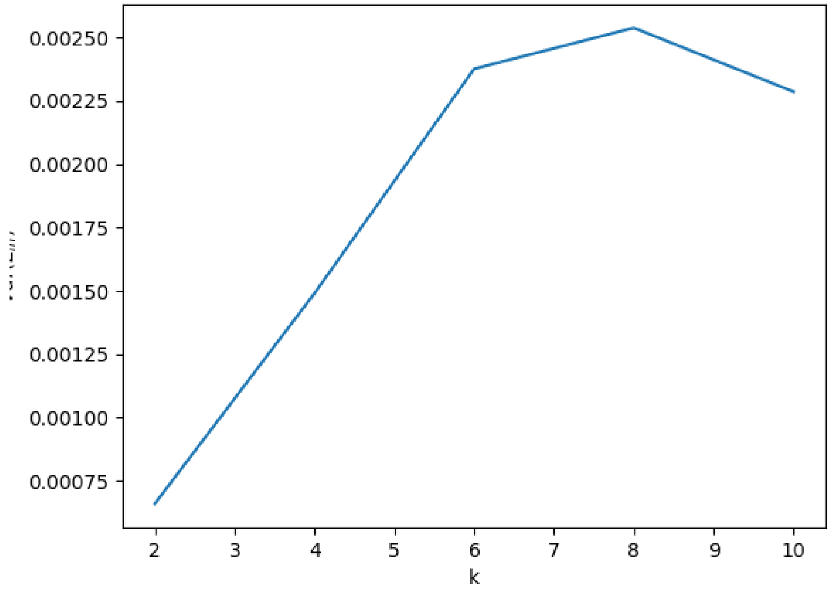
隨著 γ 的增加，E_{in}的變化並不太穩定。但大致走向應為先下降，之後回升，最後($\gamma \geq 10$)趨近一個定值。

15.



可以看出隨著 k 越大，$E[E_{in}]$會有遞減之趨勢。這相當得符合理論，本來分越多個 cluster，題目定義之E_{in}的期望值就會越低，在$k =$ data 個數時，$E_{in} = 0$。

16.



隨著 k 越大，$\mathrm{Var}(\mathrm{E_{in}})$會有先遞增再遞減之趨勢，不過其數值其實都很小。而我推測，若一開始選的點不重複，在$\mathrm{k} = \mathrm{data}$ 個數時，$\mathrm{Var}(\mathrm{E_{in}}) = 0$。

17.

為求方便，將 Δ 用 k 表示。且原要證明的$2^{\mathrm{N}} > N^k + 1$經由移項可得：

$$\frac{\ln(2^{\mathrm{N}} - 1)}{\ln N} > k$$

將不等式左側對 N 進行微分，可得：

$$\frac{\frac{2^N \ln 2}{2^N - 1} \ln N - \frac{\ln(2^N - 1)}{N}}{(\ln N)^2} > \frac{2^N \ln 2}{(2^N - 1)\ln N} - \frac{\ln 2^N}{N(\ln N)^2}$$

$$= \frac{\ln 2}{\ln N} + \frac{\ln 2}{(2^N - 1)\ln N} - \frac{\ln 2}{(\ln N)^2} > \frac{\ln 2}{\ln N} - \frac{\ln 2}{(\ln N)^2} = \frac{\ln 2}{\ln N}(1 - \frac{1}{\ln N})$$

又因為$\mathrm{k} \geq 2$。故有 $\mathrm{N} \geq 3\mathrm{k} * \log_2 k \geq 3 * 2 * 1 = 6$，並可得知 $\ln N > 1$。故：

$$\frac{\ln 2}{\ln N}\left(1 - \frac{1}{\ln N}\right) > 0$$

由 $\frac{\ln(2^{\mathrm{N}}-1)}{\ln N}$ 微分恆大於 0 可知，其為遞增函數。

接著只要確認當在$\mathrm{N} = 3\mathrm{k} * \log_2 k$時，$\frac{\ln(2^{\mathrm{N}}-1)}{\ln N} > k$ 成立，即證畢。

當$\mathrm{N} = 3\mathrm{k} * \log_2 k$時，不等式變成：

$$\ln(\mathrm{k}^{3\mathrm{k}} - 1) > k(\ln 3\mathrm{k} * \log_2 k)$$
$$\mathrm{k}^{3\mathrm{k}} > 3^k k^k (\log_2 k)^k + 1$$

欲證明此關係，先證以下在$\mathrm{k} \geq 2$ 時成立：

$$\mathrm{k}^2 \geq (3\log_2 k + 1)$$

在$\mathrm{k} \geq 2$時，易知$2\mathrm{k} > \frac{3}{k \ln 2}$，即左式遞增較右式快；且在$\mathrm{k} = 2$時，$\mathrm{k}^2 = (3\log_2 k + 1) = 4$。由以上得證。

從以上得出之小結可推得：

$$\mathrm{k}^{3\mathrm{k}} = k^k k^{2k} \geq k^k (3\log_2 k + 1)^k = (3k\log_2 k + 1)^k > (3k\log_2 k)^k + 1$$

以上證得在$\mathrm{N} = 3\mathrm{k} * \log_2 k$時，題目所求之不等式成立。再搭配最一開始證之遞增性，即證畢。

18.

使用 ML Foundation 中的符號以及下列關係式來解題：

$$m_H(N) \leq B(N,k) = \sum_{i=0}^{k-1} \binom{N}{i} \leq N^{k-1}$$

而因 NN 架構可以看成 model 的 blending，其等於有 3 個 model，每個最多有$\mathrm{N}^{\mathrm{d}+1} + 1$種 Hypothesis (因 break point 為 d+2，後面的"+1"在$\mathrm{N} = 1$時的特例)，再加上$\mathrm{d}_0^{(1)}$的 constant Hypothesis。故可知題目架構可以產生的 Hypothesis 數量為：$\left(\mathrm{N}^{\mathrm{d}+1} + 1\right)^3 + 1$。

將$m_H(N) = \left(\mathrm{N}^{\mathrm{d}+1} + 1\right)^3 + 1$化成$\mathrm{N}^{\Delta} + 1$形式以便進行證明。易知在$\mathrm{N} > \Delta$時：

$$m_H(N) = \left(\mathrm{N}^{\mathrm{d}+1} + 1\right)^3 + 1 < N^{3(d+1)+1} + 1$$

完成以上處理，接著證明$m_H(N) < 2^N$時的 N 值條件。

上題已知，當 $\Delta \geq 2, \mathrm{N} \geq 3\Delta \log_2 \Delta$ 時，$\mathrm{N}^{\Delta} + 1 < 2^N$。而前面已推導出$m_H(N)$有以下關係：

$$m_H(N) = \left(\mathrm{N}^{\mathrm{d}+1} + 1\right)^3 + 1 < N^{3(d+1)+1} + 1$$

而在$\mathrm{N} \geq 3(3(\mathrm{d}+1)+1) \log_2(3(\mathrm{d}+1)+1)$ 時，結合上題之式子可知：

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

即表示$m_H(N)$無法 shatter N 個點。

故可知其 VC dimension 必比$3(3(\mathrm{d}+1)+1) \log_2(3(\mathrm{d}+1)+1)$小。