Machine Learning Hw4

r07922100 楊力權

1.

46個weight，一層hidden layer一定至少有兩個hidden node(一個neuron+一個bias)，假設最基本的network是$9-1-1$共有2個hidden node，在第一層hidden layer加入一個node會增加11個weight，若加入新的hidden layer會需要兩個hidden node且只增加2個weight，而在非第一層hidden layer加入一個node會多3個weight。因此要建立最少weight的network必是一直新增hidden layer提升深度而不提升廣度。

36個hidden node能建立18個hidden layer，因此共有18*2+10*1=46個weight。

2.

由遞迴程式計算最多weight的層數與架構，得結構為10->22->14->1，共10*21+22*13+14*1=510個weight。

3.

$$\nabla_w \text{err}_n(w) = \nabla_w \|x_n - ww^T x_n\|^2 = \nabla_w (x_n^T x_n - 2x_n^T ww^T x_n + x_n^T ww^T ww^T x_n)$$
$$= \nabla_w (x_n^T x_n - 2(w^T x_n)^2 + (w^T x_n)^2 w^T w) = -4w^T x_n x_n + 2w^T x_n x_n w^T w + 2(w^T x_n)^2 w$$

4.

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

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

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

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

ϵ_n mean 0 $\Rightarrow E(\epsilon_n) = 0 \Rightarrow E(2x_n^T ww^T ww^T \epsilon_n - 2x_n^T ww^T \epsilon_n) = 0$

$\Rightarrow \Omega(w) = E(\frac{1}{N}\sum_{n=1}^{N}(\epsilon_n^T w)w^T w(w^T\epsilon_n)) = E(\frac{1}{N}\sum_{n=1}^{N}\epsilon_n^T w w^T \epsilon_n)w^T w = E(\frac{1}{N}\sum_{n=1}^{N} w^T \epsilon_n \epsilon_n^T w)w^T w$

$= w^T \frac{1}{N}\sum_{n=1}^{N} E(\epsilon_n \epsilon_n^T) w w^T w$ 因為ϵ_n mean 0, variance 1 $\Rightarrow E(\epsilon_n \epsilon_n^T) = I$

$\Rightarrow \Omega(w) = (w^T w)^2$

5.

$$\sum_{t=1}^{d}((\sum_{i=1}^{d}\sum_{j=1}^{\tilde{d}}\tanh(x_i u_{ij})u_{tj}) - x_t)^2$$

6.

non-tied auto encoder : $E = (d(e(x)) - x)^2 = (W^{(2)}e(x) - x)^2 = (W^{(2)}(W^{(1)}x) - x)^2$

$$\Rightarrow \frac{\partial E}{\partial W_{ij}^{(1)}} = \frac{\partial E}{\partial(d(e(x)) - x)}\frac{\partial(W^{(2)}(W^{(1)}(x)) - x)}{\partial W_{ij}^{(1)}} = \frac{\partial E}{\partial(d(e(x)) - x)}(W_{j\to all}^{(2)}x_i)$$

$$\Rightarrow \frac{\partial E}{\partial W_{ji}^{(2)}} = \frac{\partial E}{\partial(d(e(x)) - x)}\frac{\partial(W^{(2)}(W^{(1)}(x)) - x)}{\partial W_{ji}^{(2)}} = \frac{\partial E}{\partial(d(e(x)) - x)}(W_{all\to j}^{(1)}x)$$

tied auto encoder : $E = (d(e(x)) - x)^2 = (U^T e(x) - x)^2 = (U^T(Ux) - x)^2$

此時$W^{(1)} = W^{(2)^T} = U$

$$\Rightarrow \frac{\partial E}{\partial U_{ij}} = \frac{\partial E}{\partial(d(e(x)) - x)}\frac{\partial(U^T(U(x)) - x)}{\partial U_{ij}} = \frac{\partial E}{\partial(d(e(x)) - x)}(U_{all\to j}x + U_{j\to all}^T x_i)$$

$= \frac{\partial E}{\partial W_{ij}^{(1)}} + \frac{\partial E}{\partial W_{ji}^{(2)}}$，得證。

(activation function只是微分的其中一步驟，因此不論tied non-tied梯度值相同)

7.

兩點x_+, x_-做1Nearest Neighbor表示以兩點正中間的平面$w^T x + b = 0$為分類標準。

平面的垂直向量$\vec{n} = x_+ - x_-$，且過點$\frac{x_+ + x_-}{2}$因此得平面

$$(x_+ - x_-)^T(x - \frac{x_+ + x_-}{2}) = 0$$

$$\Rightarrow (x_+ - x_-)^T x - \frac{(x_+ - x_-)^T(x_+ + x_-)}{2} = (x_+ - x_-)^T x - \frac{x_+^T x_+ - x_-^T x_-}{2} = 0$$

$$\Rightarrow g_{LIN}(x) = \text{sign}(w^T x + b) = \text{sign}((x_+ - x_-)^T x - \frac{\|x_+\|^2 - \|x_-\|^2}{2})$$

8.

一sample x若要歸類為正則

$\beta_+ e^{-\|x-\mu_+\|^2} + \beta\_e^{-\|x-\mu_-\|^2} > 0 \Rightarrow \beta_+ e^{-\|x-\mu_+\|^2} > -\beta\_e^{-\|x-\mu_-\|^2}$ (因$\beta_- < 0$)兩邊取ln

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

$$\Rightarrow \|x\|^2 + \|\mu_+\|^2 - 2\mu_+^T x - \|x\|^2 - \|\mu_-\|^2 + 2\mu_-^T x < \ln -\frac{\beta_+}{\beta_-}$$

$$\Rightarrow 2(\mu_- - \mu_+)^T x + \|\mu_+\|^2 - \|\mu_-\|^2 < \ln -\frac{\beta_+}{\beta_-} \Rightarrow 2(\mu_+ - \mu_-)^T x + (\|\mu_-\|^2 - \|\mu_+\|^2) + \ln(-\frac{\beta_+}{\beta_-}) > 0$$

已知 x 帶入後為正必須滿足上線性式，因此可得

$$g_{LIN}(x) = \text{sign}(w^T x + b) = \text{sign}((2\mu_+ - 2\mu_-)^T x + (\|\mu_-\|^2 - \|\mu_+\|^2) + \ln(-\frac{\beta_+}{\beta_-}))$$

9.

經過步驟2.1也就是固定V最佳化W，已知objective function為

$\min_W \sum_{m=1}^{M} \sum_{(x_n, r_{nm}) \in D_m} (r_{nm} - w_m^T v_n)^2$，解第m部電影的optimal w_m：

$$\frac{\partial}{\partial w_m} \sum_{(x_n, r_{nm}) \in D_m} (r_{nm} - w_m^T v_n)^2 = \sum_{(x_n, r_{nm}) \in D_m} (-2 r_{nm} v_n + 2 w_m^T v_n v_n) = 0$$

因為V是為1的constant matrix所以$v_n = 1$

$$\Rightarrow \sum_{(x_n, r_{nm}) \in D_m} (-2 r_{nm} + 2 w_m^T) = -2 \sum_{(x_n, r_{nm}) \in D_m} r_{nm} + 2 \sum_{(x_n, r_{nm}) \in D_m} w_m^T = 0$$

$$\Rightarrow \sum_{(x_n, r_{nm}) \in D_m} w_m^T = \sum_{(x_n, r_{nm}) \in D_m} r_{nm} \Rightarrow w_m^T = \frac{\sum_{(x_n, r_{nm}) \in D_m} r_{nm}}{|(x_n, r_{nm}) \in D_m|}$$

可得w_m^T 是由所有給m-th movie評分的分數的平均。

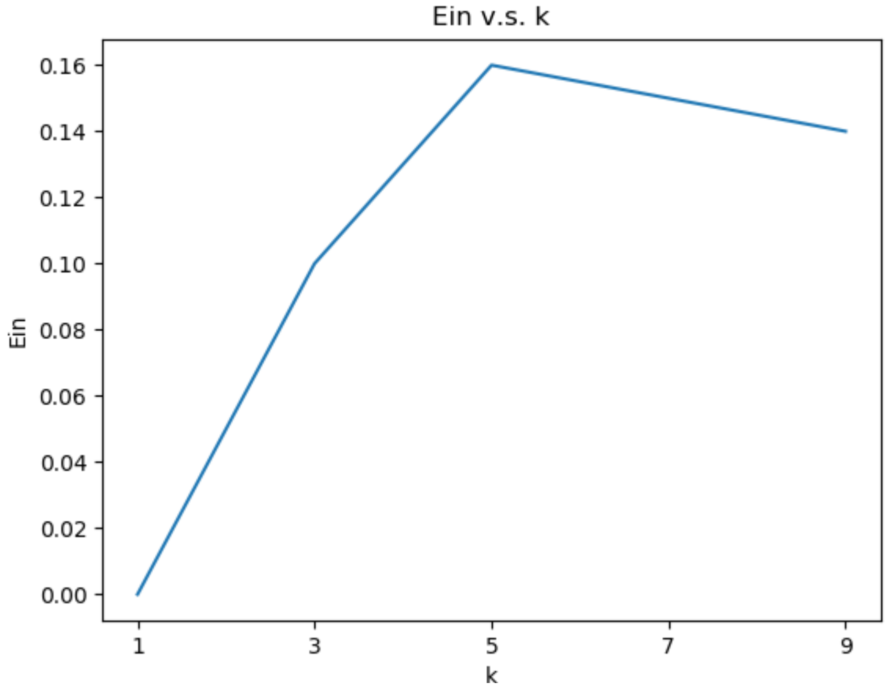
10.

對新使用者而言每一部電影的分數為

$$r_{(N+1)m} = v_{N+1}^T w_m = \frac{1}{N} \sum_{n=1}^{N} v_n^T w_m = \frac{1}{N} \sum_{n=1}^{N} r_{nm}$$

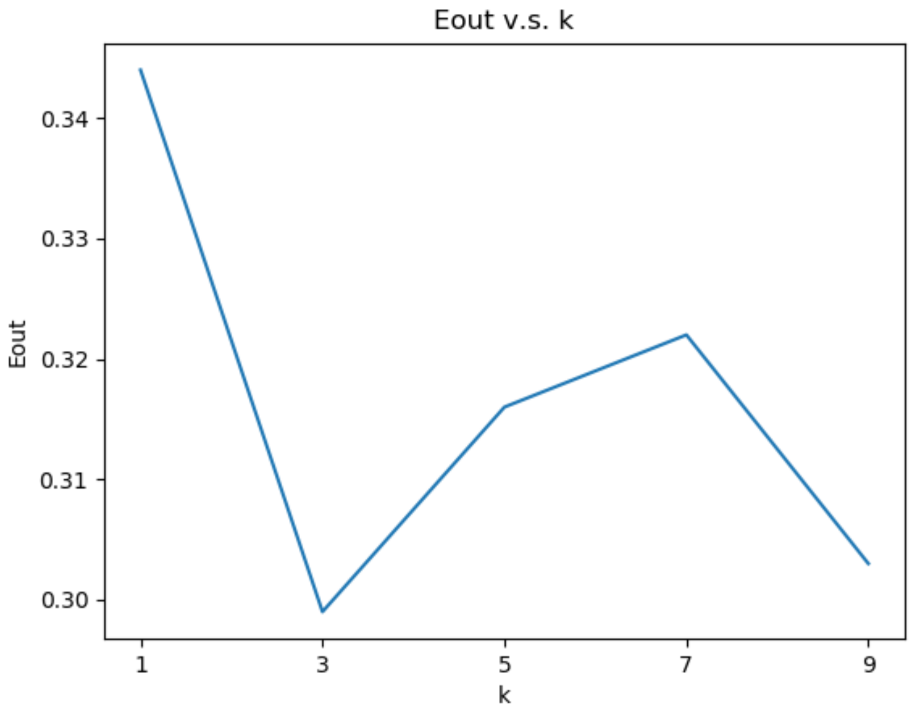
是所有舊使用者評分的平均，因此推薦電影是最高平均分數。

11.



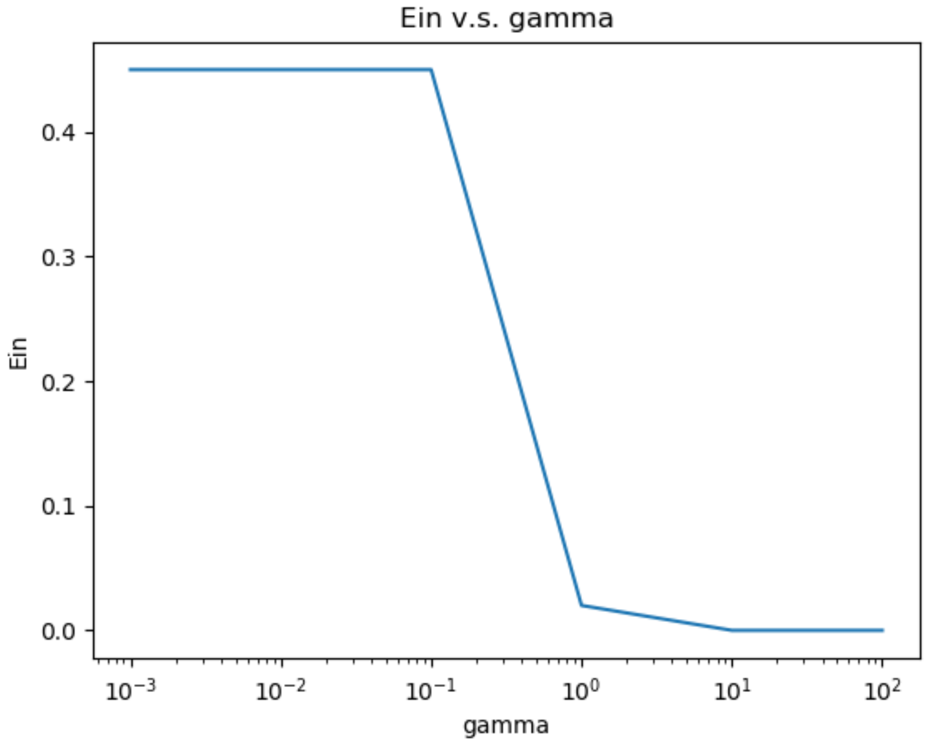
k=1 必是自己因此Ein=0，隨著鄰居點增加Ein變大，到了k=5趨於穩定。

12.



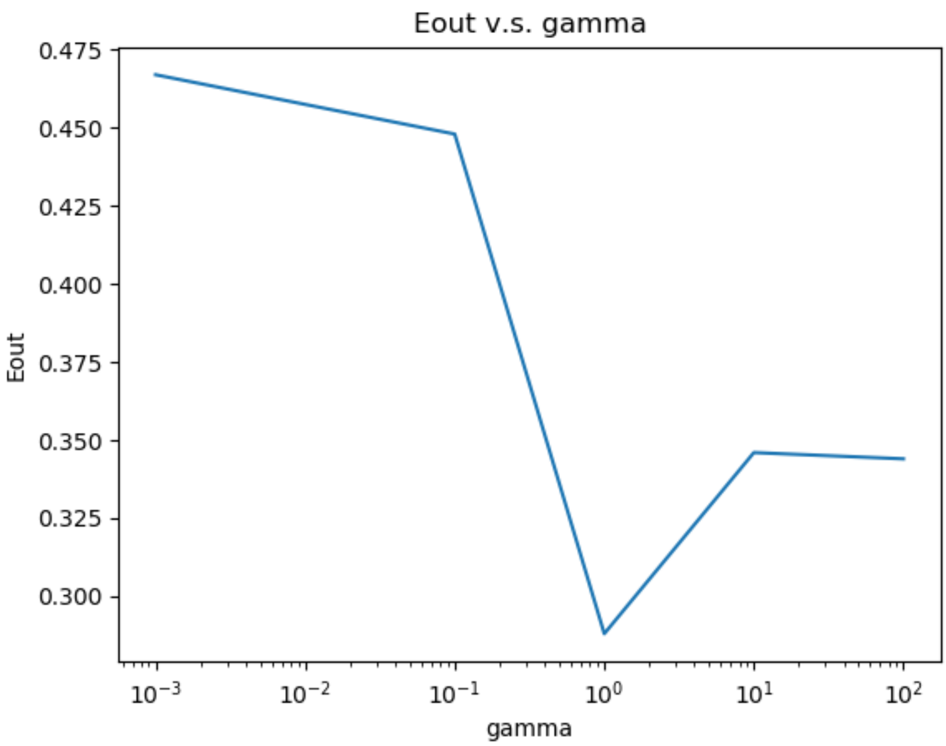
k=1顯然過於overfit因而Eout非常高，k=3的效果是最好的，而k=9趨於穩定。

13.



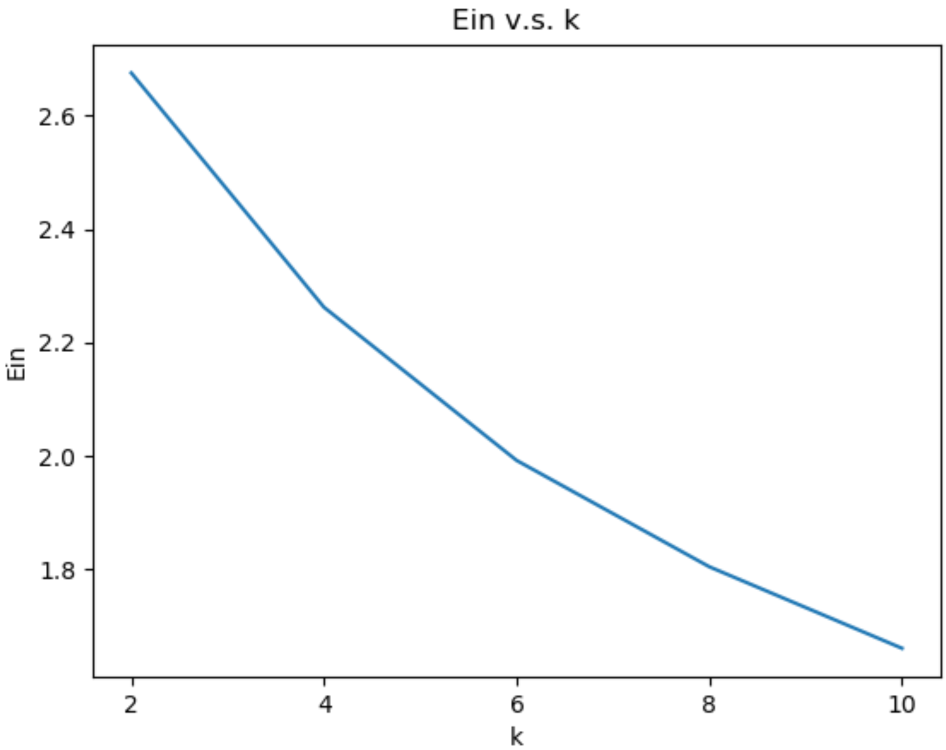
gamma愈大Ein愈小。

14.



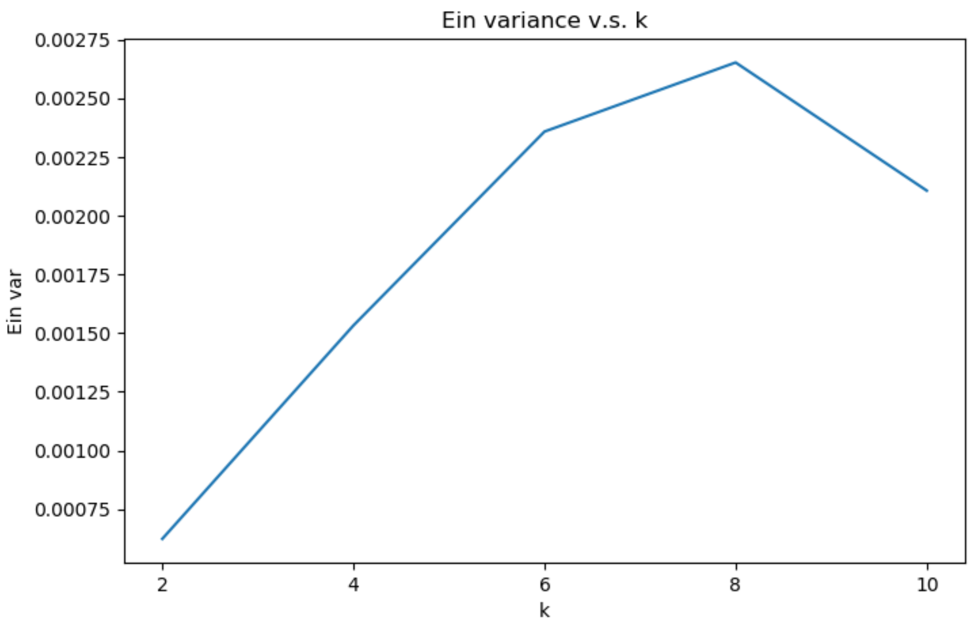
gamma愈大Eout會變小，大概在1的時候Eout最低之後overfit。

15.



k愈大Ein愈小，因為Ein是與距離有關，分成愈多群，最後converge的距離一定較小。

16.



k愈大variance愈大，k愈大隨機的初始點愈多，因此error較不穩定。

18.

因為第1層至第2層轉換函數已固定，因此只考慮第0層到第1層時，從d->3可看成三個d維perceptron，而d維perceptron之vc dimension為d+1，因此最多的dichotomy數量=bounding function $B(N, d+2) \leq N^{(d+1)} + 1$。

三個d維perceptron $|dichotomy| \leq B(N, d+2)^3 \leq (N^{(d+1)} + 1)^3 < N^{3(d+1)+1} + 1$

套用17.不等式$for \triangle \geq 2, if\ N \geq 3 \triangle \log_2 \triangle, N^{\triangle} + 1 < 2^N$

$\triangle = 3(d+1) + 1 \geq 2$，因此若資料量$N \geq 3(3(d+1)+1)\log_2(3(d+1)+1)$，則

$|dichotomy| \leq B(N, d+2)^3 \leq (N^{(d+1)} + 1)^3 < N^{3(d+1)+1} + 1 < 2^N$

也就是無法shatter $3(3(d+1)+1)\log_2(3(d+1)+1)$筆資料，因此

$d_{vc} < 3(3(d+1)+1)\log_2(3(d+1)+1)$