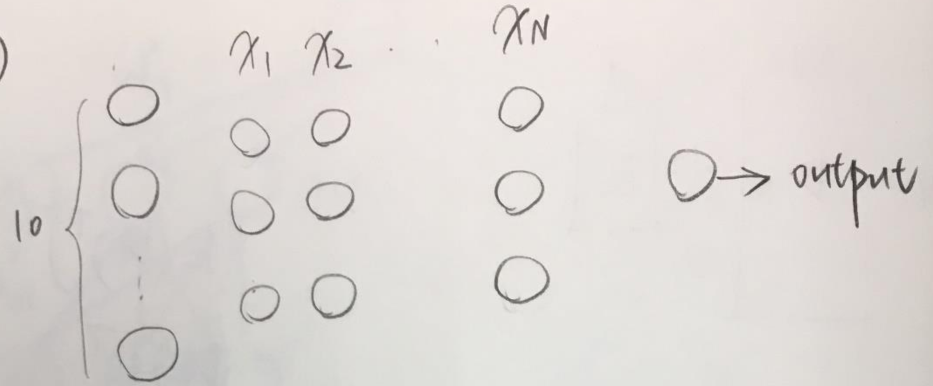
B04901069 電機四 林志皓

Problem 1

①



Let the ith layer has x_i neurons.

$\sum_{n=1}^{N} x_n = 36$. $x_n \ge 2 \ \forall n = 1 \cdots N$

Total weights number

$= 10(x_1-1) + x_1(x_2-1) + \cdots x_N \cdot 1$

$= 10x_1 + x_1x_2 + \cdots x_N \cdot 1 - (10 + x_1 + \cdots x_{N-1})$

$= 10x_1 + x_1x_2 + \cdots x_N \cdot 1 - (10 + 36 - x_N)$

$= (10x_1 + x_1x_2 + \cdots x_N \cdot 2) - 46$.

For each neuron to any layer ⇒ at least has 4 weights (product of neuron number of preceding & back layers).

⇒ $x_1 = x_2 = \cdots = x_N = 2$ leads to minimum weights

⇒ $(10 \cdot 2 + 4 \cdot 18) - 46 = 46$ (minimum)

#

Problem 2

②

According to question ①.

Total weights number $= (10x_1 + x_1x_2 + \cdots + x_N \cdot 2) - 46$.

and $\sum_{n=1}^{N} x_n = 36$. $x_i \geq 2 \ \forall i = 1, 2 \cdots N$

Assume 1 hidden layer: $10 \cdot 36 + 36 \cdot 2 - 46 = 386$.

Assume 2 hidden layer:

let $x_1 = x$, $x_2 = 36 - x$

$10x + x(36-x) + (36-x) \cdot 2 - 46 = -x^2 + 44x + 26 = -(x-22)^2 + 510$.

when $x = 22$ has max 510

Assume ≥ 3 hidden layer. $\Rightarrow$ < 510 weights.

$\Rightarrow$ maximum possible number of weights $= 510$.

#

Problem 3

③

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

$$= (x_n - ww^T x_n)^T (x_n - ww^T x_n)$$

$$= (x_n^T - x_n^T w w^T)(x_n - ww^T x_n)$$

$$= x_n^T x_n - 2 x_n^T w w^T x_n + x_n^T w w^T w w^T x_n$$

$$= x_n^T x_n - 2(w^T x_n)^2 + (w^T x_n)^2 (w^T w)$$

$(\because w^T x_n = x_n^T w = k$, k is a constant$)$.

$$\therefore \nabla_w err_n(w) = \frac{\partial x_n^T x_n}{\partial w} - 4(w^T x_n)\frac{\partial w^T x_n}{\partial w}$$

$$+ 2(w^T x_n)\frac{\partial w^T x_n}{\partial w}(w^T w) + (w^T x_n)^2 \frac{\partial w^T w}{\partial w}$$

$$= 0 - 4(w^T x_n) x_n + 2(w^T x_n)(w^T w) x_n + 2(w^T x_n)^2 w$$

#

Problem 4

④ $\epsilon_n \sim \text{uniform}(0, 1)$.

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

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

$$= \frac{1}{N}\sum_{n=1}^{N} (k_n - ww^T\epsilon_n)^T (k_n - ww^T\epsilon_n)$$

$$= \frac{1}{N}\sum_{n=1}^{N} (k_n^T - \epsilon_n^T ww^T)(k_n - ww^T\epsilon_n)$$

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

$$= \frac{1}{N}\sum_{n=1}^{N} \left[k_n^T k_n - 2(w^T k_n)(w^T\epsilon_n) + (w^T\epsilon_n)^2 \cdot (w^T w) \right]$$

$$= \frac{1}{N}\sum_{n=1}^{N} k_n^T k_n - 2\underbrace{\mathbb{E}[(w^T k_n)(w^T\epsilon_n)]}_{=0} + \mathbb{E}[(w^T\epsilon_n)^2] \cdot (w^T w)$$

$$= \frac{1}{N}\sum_{n=1}^{N} k_n^T k_n - 0 + \mathbb{E}[w^T\epsilon_n\epsilon_n^T w] \cdot (w^T w)$$

and. $\epsilon_{n,i} \sim \text{uniform}(0,1)$. $\therefore$

$$\mathbb{E}[\epsilon_{n,i} \cdot \epsilon_{n,k}] = \begin{cases} 0 & \text{if } i \neq k. \ (\because \mathbb{E}[\epsilon_{n,i}] \cdot \mathbb{E}[\epsilon_{n,k}] = 0) \\ 1 & \text{if } i = k. \ (\because \mathbb{E}[x^2] = (\mathbb{E}[x])^2 + Var(x) = 1 \end{cases}$$

$$\therefore E_{in}(w) = \frac{1}{N}\sum_{n=1}^{N} k_n^T k_n + \mathbb{E}[w^T I w] \cdot (w^T w)$$

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

$$\therefore \Omega(w) = (w^T w)^2 \quad \#$$

Problem 5

⑤ loss of basic autoencoder: $\sum_{i=1}^{d}(g_i(x)-x_i)^2$.

let $h = \begin{bmatrix} h_1 \\ h_2 \\ \vdots \\ h_d \end{bmatrix}$ represent the vector of hidden layer.

$\because u_{ij} = w_{ij}^{(1)} = w_{ji}^{(2)}$

$$g_i(x) = \sum_{n=1}^{\tilde{d}} w_{ni}^{(2)} h_n = \sum_{n=1}^{\tilde{d}} u_{in} h_n$$

$$= \sum_{n=1}^{\tilde{d}} u_{in} \tanh\left(\sum_{m=1}^{d} w_{mn}^{(1)} x_m\right)$$

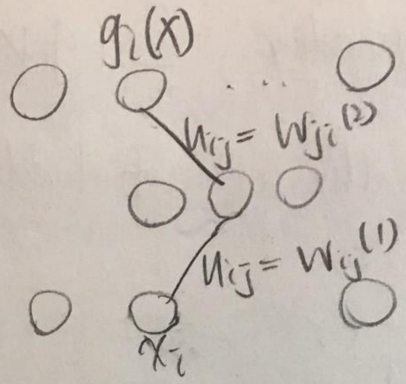
$$= \sum_{n=1}^{\tilde{d}} u_{in} \tanh\left(\sum_{m=1}^{d} u_{mn} x_m\right)$$

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

$$= \sum_{i=1}^{d}\left(\left(\sum_{n=1}^{\tilde{d}} u_{in} \tanh\left(\sum_{m=1}^{d} u_{mn} x_m\right)\right) - x_i\right)^2$$

Problem 6

(b)



let output of ith hidden neuron $= h_i$

$$E = \sum_{n=1}^{d} \left(g_n(x) - x_n \right)^2$$

$$\frac{\partial E}{\partial g_n(x)} = 2\left(g_n(x) - x_n \right)$$

let $\delta_i = \dfrac{\partial(\text{output of ith hidden})}{\partial(\text{input of ith hidden})}$

if $n \neq i$, $\dfrac{\partial g_n(x)}{\partial u_{ij}} = \dfrac{\partial(u_{nj} h_j)}{\partial u_{ij}} = u_{nj} \delta_j x_i$

if $n = i$

$$\frac{\partial g_n(x)}{\partial u_{ij}} = \frac{\partial(u_{ij} h_j)}{\partial(u_{ij})} = h_j + u_{ij} \frac{\partial h_j}{\partial u_{ij}}$$

$$= h_j + u_{ij} \delta_j x_i$$

$$\therefore \frac{\partial E}{\partial u_{ij}} = \sum_{n=1}^{d} 2\left(g_n(x) - x_n \right) \cdot u_{nj} \delta_j x_i + 2\left(g_i(x) - x_i \right) h_j$$

$$\frac{\partial g_n(x)}{\partial W_{ij}^{(1)}} = \frac{\partial (W_{jn}^{(2)} h_j)}{\partial W_{ij}^{(1)}} = W_{jn}^{(2)} \delta_j x_i$$

$$= U_{nj} \delta_j x_i$$

$$\frac{\partial g_n(x)}{\partial W_{ji}^{(2)}} = \frac{\partial (W_{ji}^{(2)} h_j)}{\partial W_{ji}^{(2)}} = \begin{cases} h_j, & \text{if } n=i \\ 0, & \text{if } n \neq i \end{cases}$$

$$\therefore \frac{\partial E}{\partial W_{ij}^{(1)}} + \frac{\partial E}{\partial W_{ji}^{(2)}}$$

$$= \sum_{n=1}^{d} 2(g_n(x) - x_n) U_{nj} \delta_j x_i$$

$$+ 2(g_i(x) - x_i) h_j$$

$$= \frac{\partial E}{\partial U_{ij}} \Rightarrow \text{proved !}$$

Problem 7

(7) hypothesis $g_{LIN}(x) = \text{sign}(w^T x + b)$

$w = x_+ - x_-$ and hyperplane $w^T x + b = 0$

passes through mid-point of x_+ & x_- $\rightarrow$ $\frac{x_+ + x_-}{2}$

$\therefore (x_+ - x_-)^T \left(\frac{x_+ + x_-}{2}\right) + b = 0$. $\frac{1}{2}\left(\|x_+\|^2 - \|x_-\|^2\right) + b = 0$

$b = -\frac{1}{2}\left(\|x_+\|^2 - \|x_-\|^2\right)$

$\therefore g_{LIN}(x) = \text{sign}\left((x_+ - x_-)^T x + \frac{1}{2}\left(\|x_+\|^2 - \|x_-\|^2\right)\right)$

Problem 8

(8)

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

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

$$\|x-\mu_-\|^2 - \|x-\mu_+\|^2 = (x-\mu_-)^T(x-\mu_-) - (x-\mu_+)^T(x-\mu_+)$$

$$= (x^Tx - 2\mu_-^T x + \mu_-^T\mu_-) - (x^Tx - 2\mu_+^T x + \mu_+^T\mu_+)$$

$$= 2(\mu_+ - \mu_-)^T x + (\mu_-^T\mu_- - \mu_+^T\mu_+)$$

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

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

$$= \text{sign}(w^T x + b).$$

Where $w = 2(\mu_+ - \mu_-)$. $b = \mu_-^T\mu_- - \mu_+^T\mu_+ - \ln\left(-\frac{\beta_-}{\beta_+}\right)$

Problem 9

⑨ $v_n = 1 \quad \forall n = 1, 2, \dots, N$ after initialization.

w_m is optimized via least square error over $\{v_n, r_{nm}\}$

let error $= E = \sum_{n=1}^{N} (v_n w_m - r_{nm})^2 = \sum_{n=1}^{N} (w_m - r_{nm})^2$

$$\frac{\partial E}{\partial w_m} = \sum_{n=1}^{N} 2 (w_m - r_{nm})$$

$\frac{\partial E}{\partial w_m} = 0$ when $\sum_{n=1}^{N} 2 (w_m - r_{nm}) = 0$

$$\Rightarrow w_m = \frac{1}{N} \sum_{n=1}^{N} r_{nm}$$

$\Rightarrow$ average rating of the m-th movie

Problem 10

(10)

$$v_{N+1}^T w_m = \left(\frac{1}{N}\sum_{n=1}^{N} v_n\right)^T w_m$$

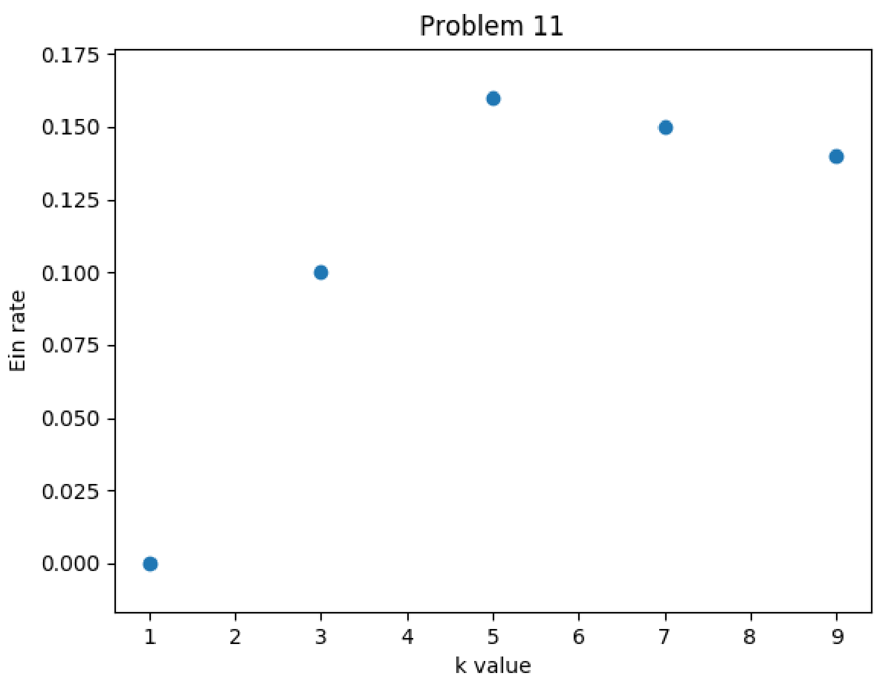
$$= \frac{1}{N}\sum_{n=1}^{N} v_n^T w_m = \frac{1}{N}\sum_{n=1}^{N} r_{nm}$$

$\equiv$ The average rating of the m th movie

$\therefore$ maximum predicted score $v_{N+1}^T w_m$

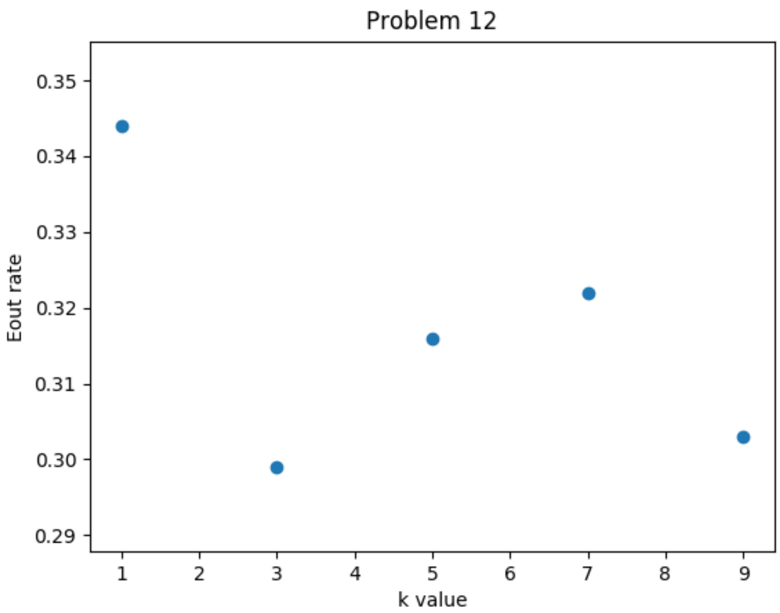
$\Rightarrow$ the movie with largest average rating

Problem 11



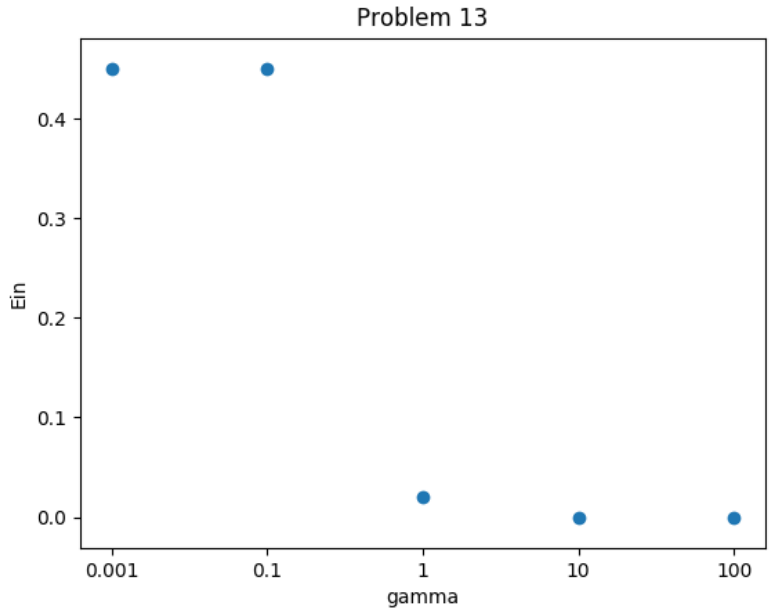
When k = 1, the prediction is obtained by the data itself. so the Ein is 0 as a result, and we can see that while k is increasing, the E in rate is increasing, too.

Problem 12



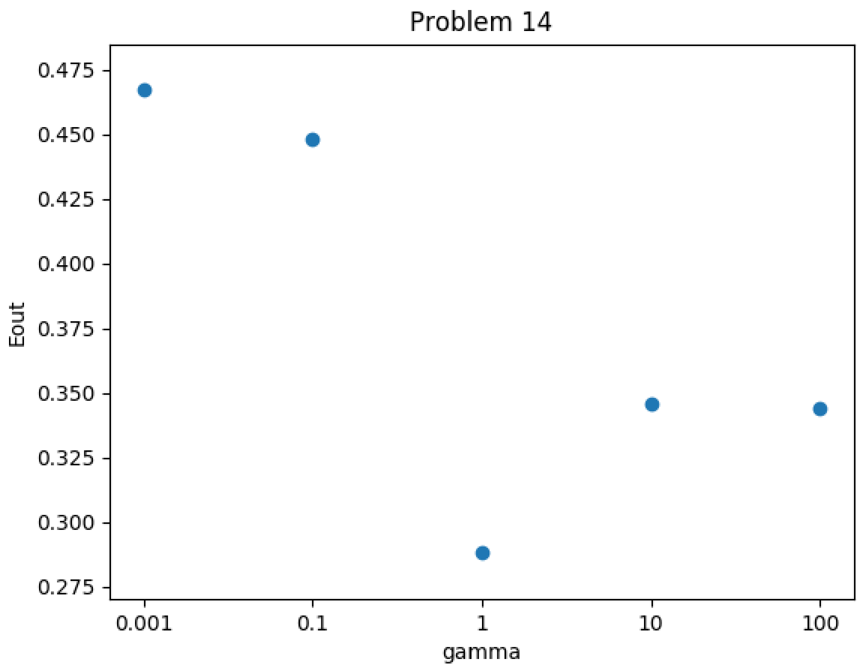
As we can see in this picture, when k= 1, the prediction is obtained by the nearest data in the training data, and it's not sufficient; so when k gets bigger, the E out rate are smaller because the hypothesis consider more data around the testing data.

Problem 13



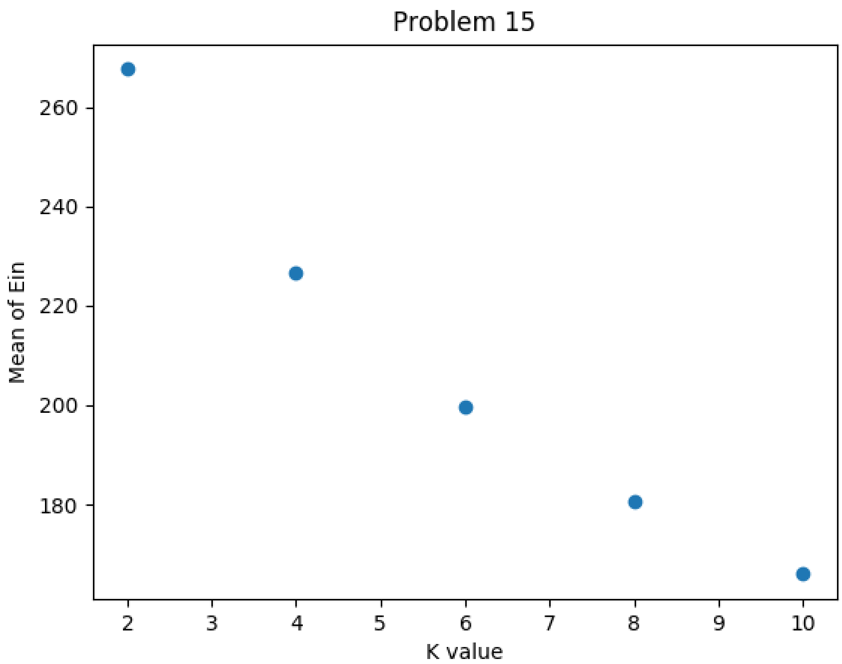
When gamma is small, the data far from the target can influence it largely, so the Error is high as a result, and when gamma gets bigger, the hypothesis tends to obtain prediction from the nearest data, so the error is low.

Problem 14



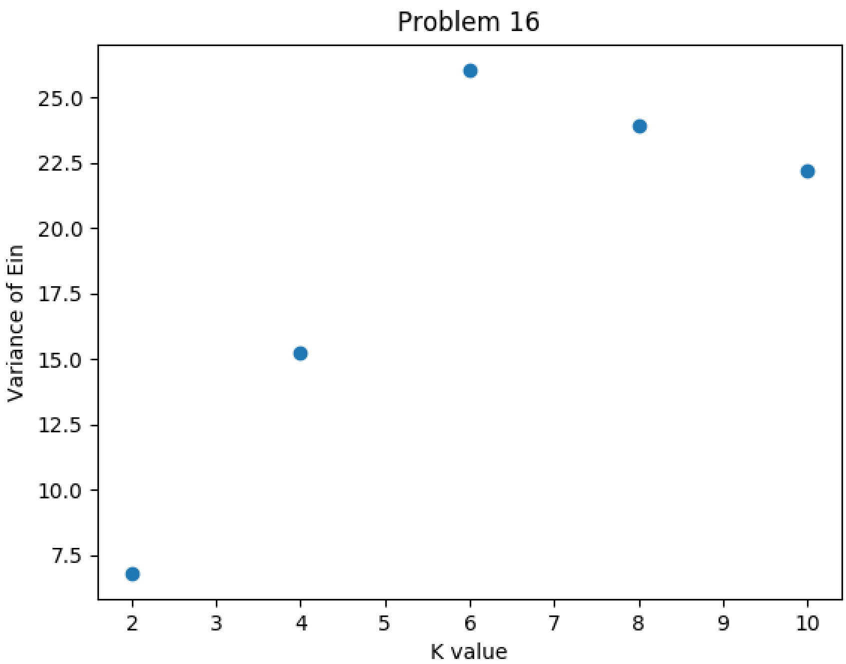
As we can see, when we chose a proper size for gamma (in this case = 1), can obtain the best performance.

Problem 15



While k value gets bigger, the Ein is decreasing, it's obivious becase the average distance to nearest center is decreased.

Problem 16



The variance is increasing while k gets bigger, it's probably because the more centers can be learned, the more possilbe solution for convergence, which leads to larger variance.

Problem 17

17

consider extreme case: $N = 3\Delta \log_2 \Delta$

$$\frac{N^\Delta+1}{2^N} = \frac{(3\Delta\log_2\Delta)^\Delta+1}{2^{3\Delta\log_2\Delta}} = \frac{(3\Delta\log_2\Delta)^\Delta+1}{\Delta^{3\Delta}}$$

$$= \left(\frac{3\Delta\log_2\Delta}{\Delta^3}\right)^\Delta + \frac{1}{\Delta^{3\Delta}} = \left(\frac{3\log_2\Delta}{\Delta^2}\right)^\Delta + \frac{1}{\Delta^{3\Delta}}$$

let $f(\Delta) = \left(\frac{3\log_2\Delta}{\Delta^2}\right)^\Delta + \frac{1}{\Delta^{3\Delta}} = \left[\left(\frac{3}{\Delta}\right)\cdot\left(\frac{\log_2\Delta}{\Delta}\right)\right]^\Delta + \frac{1}{\Delta^{3\Delta}}$

when $\Delta=2$. $f(2) = \frac{9}{16}+\frac{1}{64} = \frac{37}{64} < 1$.

when $\Delta>2$. $f(\Delta)<f(2)<1$. ($\because \frac{3}{\Delta}, \frac{\log_2\Delta}{\Delta}, \frac{1}{\Delta^{3\Delta}}$ are decreasing)

$\therefore \frac{N^\Delta+1}{2^N}<1$ for all $\Delta\ge2$ when $N=3\Delta\log_2\Delta$.

$\therefore N^\Delta+1<2^N$ for all $\Delta\ge2$ when $N=3\Delta\log_2\Delta$.

if $\Delta'<\Delta$. $N^{\Delta'}+1 < N^\Delta+1 < 2^N$, $N = 3\Delta\log_2\Delta > 3\Delta'\log_2\Delta'$

$\therefore$ for $\Delta\ge2$. if $N\ge 3\Delta\log_2\Delta$, $N^\Delta+1<2^\Delta$