Machine Learning

Answer Sheet for Homework 8

Da-Min HUANG

R04942045

Graduate Institute of Communication Engineering, National Taiwan University

January 20, 2016

Problem 1

1. Forward:

$$(A+1)\times B+(B+1)\times 1=(A+2)\,B+1 \tag{1}$$

2. Backward:

$\delta_1^{(L)}=-2\left(y_n-s_1^{(L)}\right)x_i^{(L-1)}$ counts and

$$\frac{\partial e_n}{\partial w_{ij}^{(\ell)}}=\delta_j^{(\ell)}x_i^{(\ell-1)}\text{ for }0\leq i\leq d^{(\ell-1)}\text{ and }1\leq j\leq d^{(\ell)} \tag{2}$$

with

$$\delta_j^{(\ell)}=\sum_k\left(\delta_k^{(\ell+1)}\right)\left(w_{jk}^{(\ell+1)}\right)\left(\tanh'\left(s_j^{\ell}\right)\right) \tag{3}$$

So one backward counts

$$\underbrace{(B+1)\times 1}_{\text{output layer}}+\underbrace{B\times(A+1)}_{\text{hidden layer}}+\underbrace{B}_{\text{hidden layer }\delta_j^{(\ell)}}=(A+3)\,B+1 \tag{4}$$

Hence, total number of operations required in a single iteration of backpropagation is

$$((A+2)\,B+1)+((A+3)\,B+1)=(2A+5)\,B+2 \tag{5}$$

□

Problem 2

Suppose we have k hidden layers, which means $L = k + 1$, with $d^{(1)}$, $d^{(2)}$,..., $d^{(k)}$ units ($x_0^{(\ell)}$ is not counted here) in each layer. The number of total weights is

$$\sum_{i=0}^{k-1} \left(d^{(i)} + 1\right) d^{(i+1)} + \left(d^{(k)} + 1\right) \times 1 = \sum_{i=0}^{k-1} d^{(i)} d^{(i+1)} + \sum_{j=1}^{k} d^{(j)} + \left(d^{(k)} + 1\right) := N_w \tag{6}$$

with

$$\sum_{j=1}^{k} \left(d^{(j)} + 1\right) = \left(\sum_{j=1}^{k} d^{(j)}\right) + k = 36 \text{ and } d^{(0)} = 9 \tag{7}$$

So we have

$$N_w = (37 - k) + 9d^{(1)} + \left(\sum_{i=1}^{k-1} d^{(i)} d^{(i+1)}\right) + d^{(k)} \tag{8}$$

Since $d^{(\ell)} \geq 1$ for $0 \leq \ell \leq k + 1$, so we have $1 \leq k \leq 18$.

Claim: $k = 18$ minimizes N_w.

Proof of Claim:

If $k = 18$, we have 2 units in each hidden layer (one is $x_0^{(\ell)}$, not counted in $d^{(\ell)}$), so

$$N_w|_{k=18} = (37 - 18) + 9 \times 1 + \left(\sum_{i=1}^{17} 1 \times 1\right) + 1 = 46 \tag{9}$$

If $k = 18 - m$, $m \in \mathbb{N}$ and $1 \leq m \leq 17$, we have

$$\begin{aligned} N_w|_{k=18-m} &= (19 + m) + 9d'^{(1)} + \left(\sum_{i=1}^{17-m} d'^{(i)} d'^{(i+1)}\right) + d'^{(18-m)} && (10) \\ &\geq (19 + m) + 9 + (17 - m) + 1 && (11) \\ &= 19 + 9 + 17 + 1 = N_w|_{k=18} && (12) \end{aligned}$$

where $d'^{(\ell)}$ is the new number of each hidden layer if $k = 18 - m$ and (11) holds due to $d'^{(i)} d'^{(i+1)} \geq 1$ and $d'^{(i)} \geq 1$, $\forall i$ (by definition).

Hence, we have $N_w \geq 46$.

$\square$

Problem 3

Following the setting of Problem 2.

Claim: $k=2$ with 21 units (not included $x_0^{(1)}$) in $d^{(1)}$ and 13 units (not included $x_0^{(2)}$) in $d^{(2)}$ maxmizes N_w.

Proof of Claim:

If $k=2$ with 21 units (not included $x_0^{(1)}$) in $d^{(1)}$ and 13 units (not included $x_0^{(2)}$) in $d^{(2)}$, we have

$$N_w|_{k=2} = (37-2) + 9\times 21 + (21\times 13) + 13 = 510 \tag{13}$$

Consider following cases,

1. If $k=2$ with $34-m$ units (not included $x_0^{(1)}$) in $d^{(1)}$ and m units (not included $x_0^{(2)}$) in $d^{(2)}$, where $m \in \mathbb{N}$ and $1 \le m \le 33$, we have

   $$N_w|_{k=2} = (37-2) + 9\times(34-m) + ((34-m)\times m) + m = -(m-13)^2 + 510 \tag{14}$$

   Hence, $m=13$ maximize $N_w|_{k=2}$.

2. If $k=1$.

   $$N_w|_{k=1} = (37-1) + 9\times 35 + 35 = 386 < 510 \tag{15}$$

3. If $k=3$ with $33-n_1-n_2$ units (not included $x_0^{(1)}$) in $d^{(1)}$, n_1 units (not included $x_0^{(2)}$) in $d^{(2)}$ and n_2 units (not included $x_0^{(3)}$) in $d^{(3)}$, where $n_1, n_2 \in \mathbb{N}$ and $1 \le n_1, n_2 \le 31$, we have

   $$N_w|_{k=3} = (37-3) + 9\times(33-n_1-n_2) + ((33-n_1-n_2)\times n_1 + n_1\times n_2) + n_2 \tag{16}$$

   $$= -(n_1-12)^2 - 8n_2 + 475 \le -8n_2 + 475 \le 467 < 510 \tag{17}$$

   We can see that this results in 9-20-12-1-1 neuron network, which is equal to grab one neuron (and one constant neuron) from 9-21-13-1 neuron network to create the third hidden layer.

Claim: The structure of neuron network of s hidden layers and maximum number of weights is constructed by adding single neuron hidden layer to neuron network of $s-1$ hidden layers with maximum number of weights if $18 \ge s \ge 3$ and $s \in \mathbb{N}$.

Proof of Claim:

We will prove this claim by induction.

1. For $s=3$, it is proven above.

2. Suppose this holds for $s=\ell$.

3. Let $\max\left(N_w|_{k=\ell}\right) = N_w^{(\ell)}$, if neuron network of $\ell+1$ hidden layers is constructed by adding single neuron hidden layer to neuron network, we have

$$\begin{aligned} N_w|_{k=\ell+1} = &(37-(\ell+1)) + 9\left(d^{(1)}-1\right) + \left(\left(d^{(1)}-1\right)\left(d^{(2)}-1\right) + \left(d^{(2)}-1\right)d^{(3)}\right) \\ &+ \left(\sum_{i=3}^{\ell-1} d^{(i)}d^{(i+1)}\right) + d^{(\ell)}d^{(\ell+1)} + d^{(\ell+1)} \end{aligned} \tag{18}$$

where $d^{(i)}$ is the number of neurons of i-th hidden layer in ℓ hidden layers neuron network. So the number of first and second hidden layer in new neuron network is $d^{(1)}-1$ and $d^{(2)}-1$. Then we have

$$\begin{aligned} N_w|_{k=\ell+1} &= -1-9-d^{(1)}-d^{(2)}+1-d^{(3)}-d^{(\ell)}+d^{(\ell)}d^{(\ell+1)}+d^{(\ell+1)}+N_w^{(\ell)} && (19) \\ &= -9-d^{(1)}-d^{(2)}-d^{(3)}-d^{(\ell)}+d^{(\ell)}d^{(\ell+1)}+d^{(\ell+1)}+N_w^{(\ell)} && (20) \\ &= -9-d^{(1)}-d^{(2)}+N_w^{(\ell)} && (21) \end{aligned}$$

(29) holds due to $d^{(3)} = d^{(\ell)} = d^{(\ell+1)} = 1$. Since $d^{(i)} = 1$ for $i \geq 3$, we need to take at least two neurons from first and second layer when adding one more layer. If let $d^{(\ell+1)} = 2$, one is from $d^{(1)}$ the other two is from $d^{(2)}$, we have

$$\begin{aligned} N_w|_{k=\ell+1} = &(37-(\ell+1)) + 9\left(d^{(1)}-1\right) + \left(\left(d^{(1)}-1\right)\left(d^{(2)}-2\right) + \left(d^{(2)}-2\right)d^{(3)}\right) \\ &+ \left(\sum_{i=3}^{\ell-1} d^{(i)}d^{(i+1)}\right) + d^{(\ell)}d^{(\ell+1)} + d^{(\ell+1)} && (22) \\ = &-1-9-2d^{(1)}-d^{(2)}+2-2d^{(3)}-d^{(\ell)}+d^{(\ell)}d^{(\ell+1)}+d^{(\ell+1)}+N_w^{(\ell)} && (23) \\ = &-8-2d^{(1)}-d^{(2)}-2d^{(3)}-d^{(\ell)}+d^{(\ell)}d^{(\ell+1)}+d^{(\ell+1)}+N_w^{(\ell)} && (24) \\ = &-7-2d^{(1)}-d^{(2)}+N_w^{(\ell)} < -9-d^{(1)}-d^{(2)}+N_w^{(\ell)} && (25) \end{aligned}$$

For $17 \geq \ell \geq 3$, we have $d^{(1)} > 2$ for sure. If we take two from $d^{(1)}$, then we have

$$\begin{aligned} N_w|_{k=\ell+1} = &(37-(\ell+1)) + 9\left(d^{(1)}-2\right) + \left(\left(d^{(1)}-2\right)\left(d^{(2)}-1\right) + \left(d^{(2)}-1\right)d^{(3)}\right) \\ &+ \left(\sum_{i=3}^{\ell-1} d^{(i)}d^{(i+1)}\right) + d^{(\ell)}d^{(\ell+1)} + d^{(\ell+1)} && (26) \\ = &-1-18-d^{(1)}-2d^{(2)}+2-d^{(3)}-d^{(\ell)}+d^{(\ell)}d^{(\ell+1)}+d^{(\ell+1)}+N_w^{(\ell)} && (27) \\ = &-17-d^{(1)}-2d^{(2)}-d^{(3)}-d^{(\ell)}+d^{(\ell)}d^{(\ell+1)}+d^{(\ell+1)}+N_w^{(\ell)} && (28) \\ = &-15-d^{(1)}-2d^{(2)}+N_w^{(\ell)} < -9-d^{(1)}-d^{(2)}+N_w^{(\ell)} && (29) \end{aligned}$$

Similarly, take more neurons from the first two layers will make $N_w|_{k=\ell+1}$. Hence, the way to get $N_w^{(\ell+1)}$ is to put only one neuron to the last hidden layer.

From the conclusion above, we see that for any $s \geq 3$ hidden layers, if s is bigger, then the maximum number is smaller. Hence $N_w \leq 510$.

$\square$

Problem 4

$$
\begin{aligned}
\nabla_{\mathbf{w}} \mathrm{err}_n(\mathbf{w}) &= \frac{\partial}{\partial \mathbf{w}} \left\| \mathbf{x}_n - \mathbf{w}\mathbf{w}^T \mathbf{x}_n \right\|^2 = \frac{\partial}{\partial \mathbf{w}} \left(\mathbf{x}_n - \mathbf{w}\mathbf{w}^T \mathbf{x}_n \right)^T \left(\mathbf{x}_n - \mathbf{w}\mathbf{w}^T \mathbf{x}_n \right) && (30) \\
&= \frac{\partial}{\partial \mathbf{w}} \left(\mathbf{x}_n^T - \mathbf{x}_n^T \left(\mathbf{w}\mathbf{w}^T \right)^T \right) \left(\mathbf{x}_n - \mathbf{w}\mathbf{w}^T \mathbf{x}_n \right) && (31) \\
&= \left(-2\mathbf{x}_n^T \mathbf{w} \right) \left(\mathbf{x}_n - \mathbf{w}\mathbf{w}^T \mathbf{x}_n \right) + \left(\mathbf{x}_n^T - \mathbf{x}_n^T \left(\mathbf{w}\mathbf{w}^T \right) \right) \left(-2\mathbf{w}\mathbf{x}_n \right) && (32) \\
&= -2 \left(\mathbf{x}_n^T \mathbf{w} \right) \mathbf{x}_n + 2 \left(\mathbf{x}_n^T \mathbf{w} \right) \mathbf{w} \left(\mathbf{x}_n^T \mathbf{w} \right)^T - 2 \left(\mathbf{x}_n^T \mathbf{w} \right) \mathbf{x}_n + 2\mathbf{x}_n^T \left(\mathbf{w}^T \mathbf{w} \right) \mathbf{w}\mathbf{x}_n && (33) \\
&= -4 \left(\mathbf{x}_n^T \mathbf{w} \right) \mathbf{x}_n + 2 \left(\mathbf{x}_n^T \mathbf{w} \right)^2 \mathbf{w} + 2\mathbf{x}_n^T \left(\mathbf{w}^T \mathbf{w} \right) \mathbf{w}\mathbf{x}_n && (34)
\end{aligned}
$$

where we have used

$$
\begin{aligned}
\left(\mathbf{x}_n^T \mathbf{w} \right) \mathbf{w} \left(\mathbf{x}_n^T \mathbf{w} \right)^T &= \left(\begin{pmatrix} x_1 & x_2 & \cdots & x_n \end{pmatrix} \begin{pmatrix} w_1 \\ w_2 \\ \vdots \\ w_n \end{pmatrix} \right) \begin{pmatrix} w_1 \\ w_2 \\ \vdots \\ w_n \end{pmatrix} \left(\mathbf{x}_n^T \mathbf{w} \right)^T && (35) \\
&= c \begin{pmatrix} w_1 \\ w_2 \\ \vdots \\ w_n \end{pmatrix} c = \begin{pmatrix} c^2 w_1 \\ c^2 w_2 \\ \vdots \\ c^2 w_n \end{pmatrix} = \left(\mathbf{x}_n^T \mathbf{w} \right)^2 \mathbf{w} && (36)
\end{aligned}
$$

and

$$
\left(\mathbf{w}\mathbf{w}^T \right) \mathbf{w}\mathbf{x}_n = \mathbf{w} \left(\mathbf{w}^T \mathbf{w} \right) \mathbf{x}_n = \mathbf{w} \left(\sum_{i=1}^{n} w_i^2 \right) \mathbf{x}_n = \left(\sum_{i=1}^{n} w_i^2 \right) \mathbf{w}\mathbf{x}_n = \left(\mathbf{w}^T \mathbf{w} \right) \mathbf{w}\mathbf{x}_n \quad (37)
$$

and

$$
\left(\mathbf{w}\mathbf{w}^T \right)^T = \left(w_i w_j \right)_{ij}^T = \left(w_j w_i \right)_{ij}^T = \left(w_j w_i \right)_{ji} = \mathbf{w}\mathbf{w}^T \quad (38)
$$

with $c := \sum_{i=1}^{n} x_i w_i$.

$\square$

Problem 5

$$
\begin{aligned}
E_{\text{in}}(\mathbf{w}) &= \frac{1}{N}\sum_{n=1}^{N}\left(\mathbf{x}_n - \mathbf{w}\mathbf{w}^T(\mathbf{x}_n+\boldsymbol{\epsilon}_n)\right)^T\left(\mathbf{x}_n - \mathbf{w}\mathbf{w}^T(\mathbf{x}_n+\boldsymbol{\epsilon}_n)\right) &(39)\\
&= \frac{1}{N}\sum_{n=1}^{N}\left(\mathbf{x}_n^T - (\mathbf{x}_n+\boldsymbol{\epsilon}_n)^T\mathbf{w}^T\mathbf{w}\right)\left(\mathbf{x}_n - \mathbf{w}\mathbf{w}^T(\mathbf{x}_n+\boldsymbol{\epsilon}_n)\right) &(40)\\
&= \frac{1}{N}\sum_{n=1}^{N}\left\|\mathbf{x}_n - \mathbf{w}\mathbf{w}^T\mathbf{x}_n\right\|^2 - \boldsymbol{\epsilon}_n^T\mathbf{w}^T\mathbf{w}\left(\mathbf{x}_n - \mathbf{w}\mathbf{w}^T\mathbf{x}_n\right)\\
&\quad - \mathbf{w}\mathbf{w}^T\boldsymbol{\epsilon}_n\left(\mathbf{x}_n^T - \mathbf{x}_n^T\mathbf{w}^T\mathbf{w}\right) + (\boldsymbol{\epsilon}_n)^2\left(\mathbf{w}^T\mathbf{w}\right)^2 &(41)
\end{aligned}
$$

Since $\boldsymbol{\epsilon}_n$ is generated from a zero-mean, unit variance Gaussian distribution, so $\mathcal{E}(\boldsymbol{\epsilon}) = 0$ and $\mathcal{E}\left(\|\boldsymbol{\epsilon}\|^2\right) = 1$. Hence

$$
\mathcal{E}\left(E_{\text{in}}(\mathbf{w})\right) = \frac{1}{N}\sum_{n=1}^{N}\left\|\mathbf{x}_n - \mathbf{w}\mathbf{w}^T\mathbf{x}_n\right\|^2 + \left(\mathbf{w}^T\mathbf{w}\right)^2 \tag{42}
$$

So $\Omega(\mathbf{w}) = \left(\mathbf{w}^T\mathbf{w}\right)^2$.

□

Problem 6

<u>Claim</u>: $\mathbf{w} = 2(\mathbf{x}_+ - \mathbf{x}_-)$, $b = -\|\mathbf{x}_+\|^2 + \|\mathbf{x}_-\|^2$.
<u>Proof of Claim</u>:
Consider the following cases.

1. If $\mathbf{x} = \mathbf{x}_+$, $\mathbf{w}^T\mathbf{x} + b > 0$.

$$
\begin{aligned}
\mathbf{w}^T\mathbf{x} + b &= 2\left(\mathbf{x}_+^T - \mathbf{x}_-^T\right)\mathbf{x}_+ + \left(-\|\mathbf{x}_+\|^2 + \|\mathbf{x}_-\|^2\right) &(43)\\
&= \|\mathbf{x}_+\|^2 - 2\mathbf{x}_-^T\mathbf{x}_+ + \|\mathbf{x}_-\|^2 = \|\mathbf{x}_+\|^2 - \mathbf{x}_-^T\mathbf{x}_+ - \mathbf{x}_+^T\mathbf{x}_- + \|\mathbf{x}_-\|^2 &(44)\\
&= \|\mathbf{x}_+ - \mathbf{x}_-\|^2 > 0 &(45)
\end{aligned}
$$

2. If $\mathbf{x} = \mathbf{x}_-$, $\mathbf{w}^T\mathbf{x} + b < 0$.

$$
\begin{aligned}
\mathbf{w}^T\mathbf{x} + b &= 2\left(\mathbf{x}_+^T - \mathbf{x}_-^T\right)\mathbf{x}_- + \left(-\|\mathbf{x}_+\|^2 + \|\mathbf{x}_-\|^2\right) &(46)\\
&= -\|\mathbf{x}_+\|^2 + 2\mathbf{x}_-^T\mathbf{x}_+ - \|\mathbf{x}_-\|^2 = -\|\mathbf{x}_+\|^2 + \mathbf{x}_-^T\mathbf{x}_+ + \mathbf{x}_+^T\mathbf{x}_- - \|\mathbf{x}_-\|^2 &(47)\\
&= -\|\mathbf{x}_+ - \mathbf{x}_-\|^2 < 0 &(48)
\end{aligned}
$$

3. If $\mathbf{x} = (\mathbf{x}_+ + \mathbf{x}_-)/2$, $\mathbf{w}^T\mathbf{x} + b = 0$.

$$\begin{aligned} \mathbf{w}^T\mathbf{x} + b &= 2\left(\mathbf{x}_+^T - \mathbf{x}_-^T\right)(\mathbf{x}_+ + \mathbf{x}_-)/2 + \left(-\|\mathbf{x}_+\|^2 + \|\mathbf{x}_-\|^2\right) && (49)\\ &= \left(\|\mathbf{x}_+\|^2 - \|\mathbf{x}_-\|^2\right) + \left(-\|\mathbf{x}_+\|^2 + \|\mathbf{x}_-\|^2\right) = 0 && (50) \end{aligned}$$

4. If $\mathbf{x} = (\mathbf{x}_+ + \mathbf{x}_-)/2 + \mathbf{x}'$, $\mathbf{w}^T\mathbf{x} + b > 0$ if $\|\mathbf{x}_+ - \mathbf{x}'\|^2 < \|\mathbf{x}_- - \mathbf{x}'\|^2$; $\mathbf{w}^T\mathbf{x} + b < 0$ if $\|\mathbf{x}_+ - \mathbf{x}'\|^2 > \|\mathbf{x}_- - \mathbf{x}'\|^2$.

$$\begin{aligned} \mathbf{w}^T\mathbf{x} + b &= 2\left(\mathbf{x}_+^T - \mathbf{x}_-^T\right)\left((\mathbf{x}_+ + \mathbf{x}_-)/2 + \mathbf{x}'\right) + \left(-\|\mathbf{x}_+\|^2 + \|\mathbf{x}_-\|^2\right) && (51)\\ &= 2\left(\mathbf{x}_+^T - \mathbf{x}_-^T\right)\mathbf{x}' && (52) \end{aligned}$$

   If $\|\mathbf{x}_+ - \mathbf{x}'\|^2 < \|\mathbf{x}_- - \mathbf{x}'\|^2$, then $\mathbf{x}_+^T\mathbf{x}' > 0$ and $\mathbf{x}_-^T\mathbf{x}' < 0 \Rightarrow 2\left(\mathbf{x}_+^T - \mathbf{x}_-^T\right)\mathbf{x}' > 0$; if $\|\mathbf{x}_+ - \mathbf{x}'\|^2 > \|\mathbf{x}_- - \mathbf{x}'\|^2$, then $\mathbf{x}_+^T\mathbf{x}' < 0$ and $\mathbf{x}_-^T\mathbf{x}' > 0 \Rightarrow 2\left(\mathbf{x}_+^T - \mathbf{x}_-^T\right)\mathbf{x}' < 0$.

Hence, we have proved the claim.

□

Problem 7

If g_{RBFNET} outputs $+1$, which means

$$\beta_+ \exp\left(-\|\mathbf{x} - \boldsymbol{\mu}_+\|^2\right) + \beta_- \exp\left(-\|\mathbf{x} - \boldsymbol{\mu}_-\|^2\right) > 0 \tag{53}$$

Since $\beta_+ > 0 > \beta_-$, we have

$$\left|\frac{\beta_+}{\beta_-}\right| \frac{\exp\left(-\|\mathbf{x} - \boldsymbol{\mu}_+\|^2\right)}{\exp\left(-\|\mathbf{x} - \boldsymbol{\mu}_-\|^2\right)} > 1 \tag{54}$$

$$\left|\frac{\beta_+}{\beta_-}\right| \exp\left(-\|\mathbf{x} - \boldsymbol{\mu}_+\|^2 + \|\mathbf{x} - \boldsymbol{\mu}_-\|^2\right) > 1 \tag{55}$$

$$\exp\left(-\|\mathbf{x} - \boldsymbol{\mu}_+\|^2 + \|\mathbf{x} - \boldsymbol{\mu}_-\|^2\right) > \left|\frac{\beta_-}{\beta_+}\right| \tag{56}$$

$$2\left(\boldsymbol{\mu}_+ - \boldsymbol{\mu}_-\right)^T\mathbf{x} > \ln\left|\frac{\beta_-}{\beta_+}\right| + \|\boldsymbol{\mu}_+\|^2 - \|\boldsymbol{\mu}_-\|^2 \tag{57}$$

$$2\left(\boldsymbol{\mu}_+ - \boldsymbol{\mu}_-\right)^T\mathbf{x} + \left(\ln\left|\frac{\beta_+}{\beta_-}\right| - \|\boldsymbol{\mu}_+\|^2 + \|\boldsymbol{\mu}_-\|^2\right) > 0 \tag{58}$$

Similarly, if g_{RBFNET} outputs -1, we have

$$2\left(\boldsymbol{\mu}_+ - \boldsymbol{\mu}_-\right)^T\mathbf{x} + \left(\ln\left|\frac{\beta_+}{\beta_-}\right| - \|\boldsymbol{\mu}_+\|^2 + \|\boldsymbol{\mu}_-\|^2\right) < 0 \tag{59}$$

Hence

$$2\left(\boldsymbol{\mu}_+ - \boldsymbol{\mu}_-\right),\ b = \ln\left|\frac{\beta_+}{\beta_-}\right| - \|\boldsymbol{\mu}_+\|^2 + \|\boldsymbol{\mu}_-\|^2 \tag{60}$$

□

Problem 8

$$\text{optimal } \beta_n = \left(\left(\mathrm{Z}^T\mathrm{Z}\right)^{-1}\mathrm{Z}^T\mathbf{y}\right)_n \tag{61}$$

where Z is

$$\mathrm{Z} = \begin{pmatrix} [\![\mathbf{x}_1 \neq \mathbf{x}_1]\!] & [\![\mathbf{x}_1 \neq \mathbf{x}_2]\!] & \cdots & [\![\mathbf{x}_1 \neq \mathbf{x}_N]\!] \\ [\![\mathbf{x}_2 \neq \mathbf{x}_1]\!] & [\![\mathbf{x}_2 \neq \mathbf{x}_2]\!] & \cdots & [\![\mathbf{x}_2 \neq \mathbf{x}_N]\!] \\ \vdots & \vdots & \ddots & \vdots \\ [\![\mathbf{x}_N \neq \mathbf{x}_1]\!] & [\![\mathbf{x}_N \neq \mathbf{x}_2]\!] & \cdots & [\![\mathbf{x}_N \neq \mathbf{x}_N]\!] \end{pmatrix} = \begin{pmatrix} 0 & 1 & \cdots & 1 \\ 1 & 0 & \cdots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \cdots & 0 \end{pmatrix} \tag{62}$$

so

$$\left(\mathrm{Z}^T\mathrm{Z}\right)^{-1}\mathrm{Z}^T = \frac{1}{N-1}\begin{pmatrix} -(N-2) & 1 & \cdots & 1 \\ 1 & -(N-2) & \cdots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \cdots & -(N-2) \end{pmatrix} \tag{63}$$

Hence we have

$$\beta_n = \frac{1}{N-1}\left(\sum_{i \neq n} y_i - (N-2)\, y_n\right) \tag{64}$$

□

Problem 9

V is initialized as

$$\mathrm{V}_{\tilde{d}\times N} = \begin{pmatrix} 1 & 1 & \cdots & 1 \end{pmatrix} \tag{65}$$

so

$$\min_{w_m} \frac{1}{N}\sum_{n}^{N}\left(r_{nm} - w_m v_n\right)^2 = \min_{w_m} \frac{1}{N}\sum_{n}^{N}\left(r_{nm} - w_m\right)^2 \tag{66}$$

and we have

$$\frac{\partial}{\partial w_m}\frac{1}{N}\sum_{n}^{N}\left(r_{nm} - w_m\right)^2 = -\frac{2}{N}\sum_{n}^{N}\left(r_{nm} - w_m\right) = -2\left(\left(\frac{1}{N}\sum_{n}^{N} r_{nm}\right) - w_m\right) = 0 \tag{67}$$

Hence,

$$w_m = \frac{1}{N}\sum_{n}^{N} r_{nm} = \text{average rating of the } m\text{-th movie} \tag{68}$$

□

Problem 10

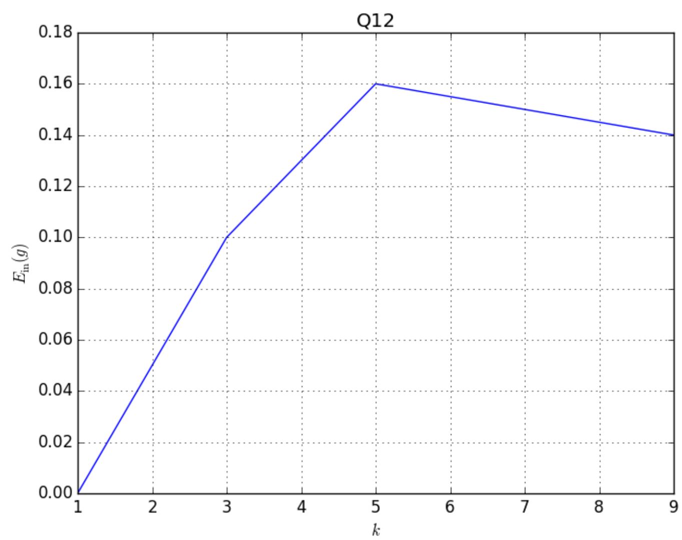
$$\mathbf{v}_{N+1}^T \mathbf{w}_m = \frac{1}{N} \left(\sum_{n=1}^{N} \mathbf{v}_n^T \right) \mathbf{w}_m = \frac{1}{N} \sum_{n=1}^{N} \mathbf{v}_n^T \mathbf{w}_m = \frac{1}{N} \sum_{n=1}^{N} r_{nm} \tag{69}$$

Hence, we have

$$\max_m \mathbf{v}_{N+1}^T \mathbf{w}_m = \max_m \frac{1}{N} \sum_{n=1}^{N} r_{nm} = \text{the movie with largest average rating} \tag{70}$$

□

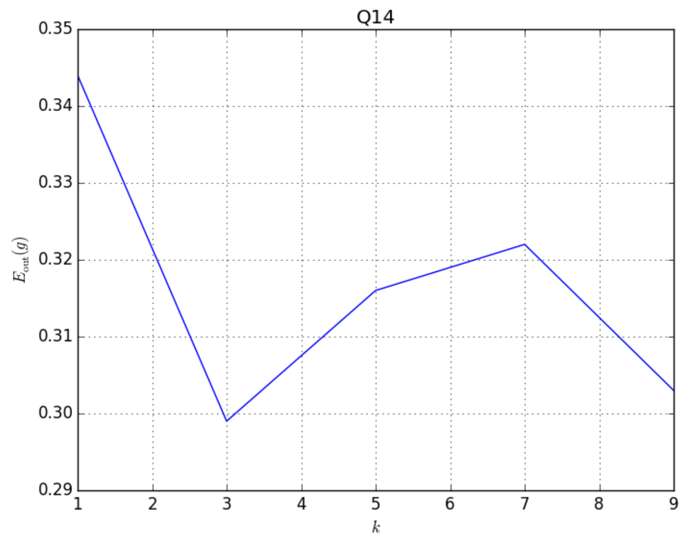
Problem 12



$E_{\text{in}}\left(g_{k\text{-nbor}}\right)$ reaches maximum at $k = 5$. And $E_{\text{in}}\left(g_{k\text{-nbor}}\right) = 0$ at $k = 1$, as expect.

□

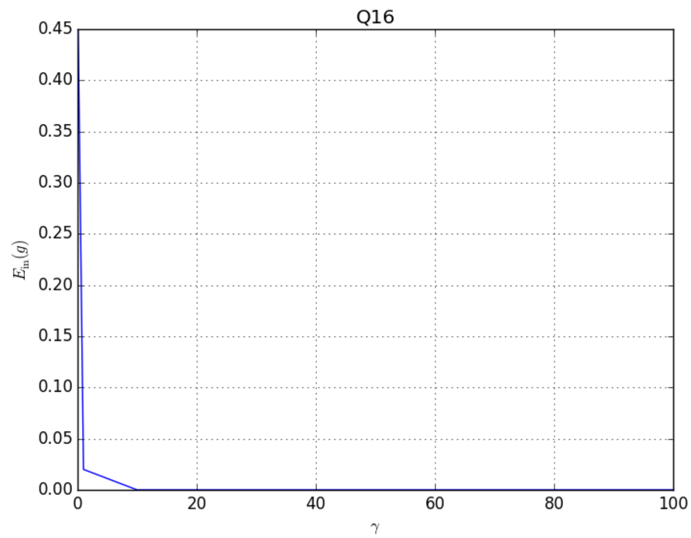
Problem 14

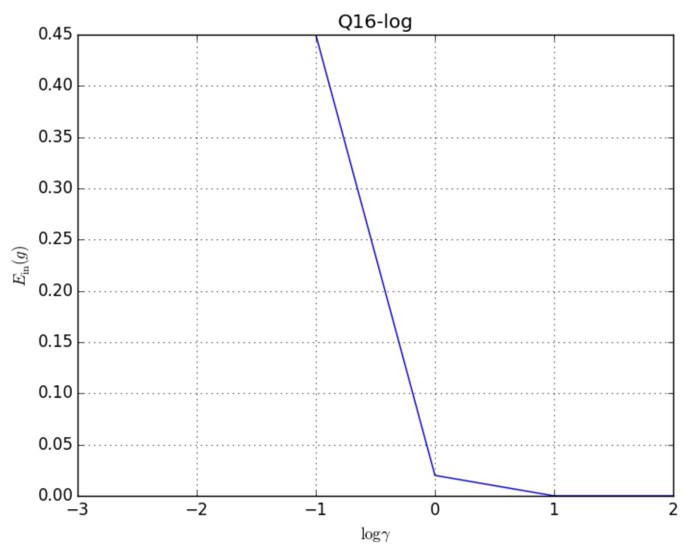


$E_{\text{in}}\left(g_{k\text{-nbor}}\right)$ reaches minimum at $k = 3$. After $k = 7$, $E_{\text{in}}\left(g_{k\text{-nbor}}\right)$ decreases.

□

Problem 16

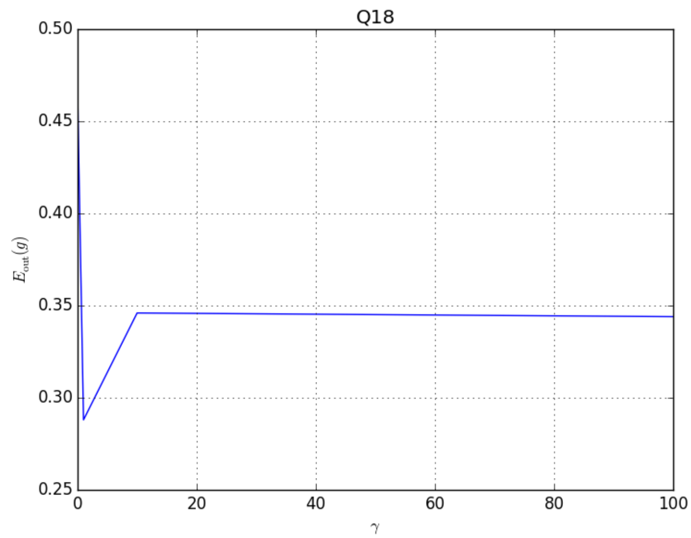


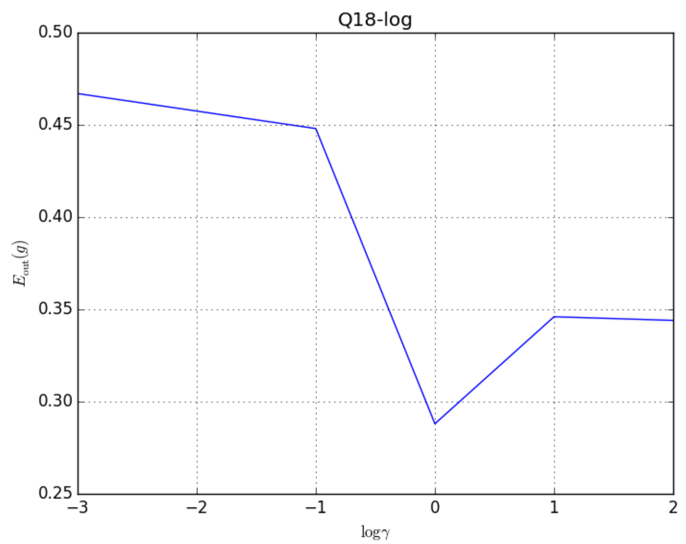


As $\log \gamma > -1$, $E_{\text{in}}\left(g_{\text{uniform}}\right)$ decreases dramatically.

$\square$

Problem 18

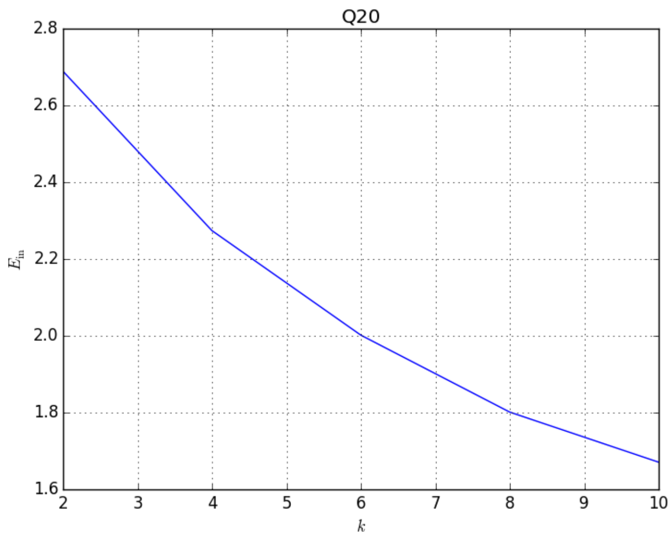




As $\log \gamma = 0$, $E_{\text{out}}\left(g_{\text{uniform}}\right)$ reaches minimum and then increases. Combine the results of Problem 16, $\log \gamma > 0$ may cause overfitting in this case.

$\square$

Problem 20



As k increases, E_{in} decreases. This is reasonable since $\|\mathbf{x}_n - \boldsymbol{\mu}_m\|^2$ should be smaller if k is larger.

□

Problem 21

Now we have $\Delta \geq 2$ and $N \geq 3\Delta \log_2 \Delta \geq 6$, so $2^N \geq 2^6 = 64$ and $N^\Delta + 1 \geq 6^2 + 1 = 37$. For fixed Δ, we have

$$\frac{\partial}{\partial N}\left(2^N\right) = \ln(2)\, 2^N \tag{71}$$

$$\frac{\partial}{\partial N}\left(N^\Delta + 1\right) = \Delta N^{\Delta-1} \tag{72}$$

so

$$\frac{\partial_N\left(2^N\right)}{\partial_N\left(N^\Delta + 1\right)} = \frac{\ln(2)\, 2^N}{\Delta N^{\Delta-1}} \tag{73}$$

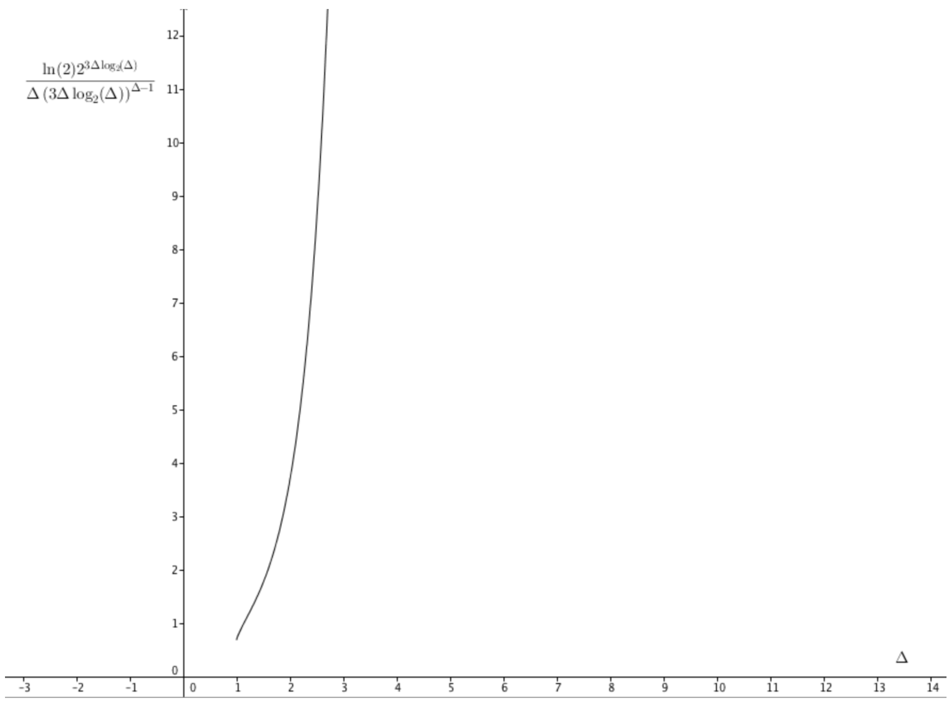
We have

$$\lim_{N\to\infty} \frac{\ln(2)\, 2^N}{\Delta N^{\Delta-1}} = \infty \tag{74}$$

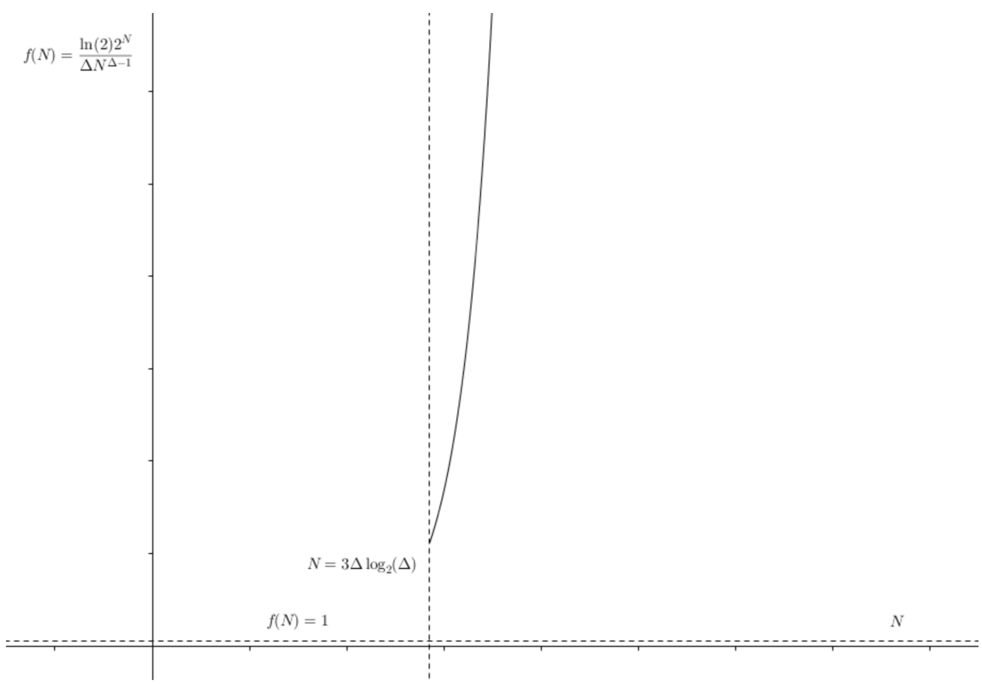
which means $\ln(2)\, 2^N$ grows faster than $\Delta N^{\Delta-1}$ if N is large. Consider $\min N = 3\Delta \log_2 \Delta$, we have

$$\frac{\ln(2)\, 2^N}{\Delta N^{\Delta-1}} = \frac{\ln(2)\, 2^{3\Delta \log_2 \Delta}}{\Delta\left(3\Delta \log_2 \Delta\right)^{\Delta-1}} \tag{75}$$

This graphs is like



and this function is greater than 1 for $\Delta \geq 2$. For some fixed Δ, we have



We can see that this function is greater than 1 for $N \geq 3\Delta \log_2 \Delta$.
As N becomes larger, 2^N always grows faster than $N^\Delta + 1$ and $\min\left(2^N\right) > \min\left(N^\Delta + 1\right)$, so $N^\Delta + 1 < 2^N$.

$\square$

Problem 22

Since $\left(w_{01}^{(2)}, w_{11}^{(2)}, w_{21}^{(2)}, w_{31}^{(2)}\right) = (-2.5, 1, 1, 1)$, the function of output layer is

$$\text{AND}\left(x_1^{(1)}, x_2^{(1)}, x_3^{(1)}\right) \tag{76}$$

And we have $(d+1) \times 3 = 3\,(d+1)$ weights between input and hidden layer.

$$\Delta = (3\,(d+1)+1),\ \ N = 3\Delta \log_2 \Delta \tag{77}$$

$$\Delta \log_2 (N) < \log_2 \left(N^{\Delta} + 1\right) < N \tag{78}$$

□

Reference

[1] Lecture Notes by Hsuan-Tien LIN, Department of Computer Science and Information Engineering, National Taiwan University, Taipei 106, Taiwan.