
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:

$$(A + 1) \times B + (B + 1) \times 1 = (A + 2) B + 1 \quad (1)$$

2. Backward:

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

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

with

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

So one backward counts

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

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

$$((A + 2) B + 1) + ((A + 3) B + 1) = (2A + 5) B + 2 \quad (5)$$

□

Problem 2

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

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

with

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

So we have

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

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

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

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

$$N_w|_{k=18-m} = (19 + m) + 9d'^{(1)} + \left(\sum_{i=1}^{17-m} d'^{(i)} d'^{(i+1)} \right) + d'^{(18-m)} \quad (10)$$

$$\geq (19 + m) + 9 + (17 - m) + 1 \quad (11)$$

$$= 19 + 9 + 17 + 1 = N_w|_{k=18} \quad (12)$$

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 $N_w \geq 46$.

□

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 $d^{(2)}$ maximizes 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

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

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 $1 \leq m \leq 33$, we have

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

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

2. If $k = 1$.

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

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 $1 \leq n_1, n_2 \leq 31$, we have

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

$$= -(n_1 - 12)^2 - 8n_2 + 475 \leq -8n_2 + 475 \leq 467 < 510 \quad (17)$$

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

Claim: The structure of neural network of s hidden layers and maximum number of weights is constructed by adding single neuron hidden layer to neural network of $s - 1$ hidden layers with maximum number of weights if $18 \geq s \geq 3$ 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 $\max(N_w|_{k=\ell}) = N_w^{(\ell)}$, if neural network of $\ell + 1$ hidden layers is constructed by adding single neuron hidden layer to neural network, we have

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

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

$$N_w|_{k=\ell+1} = -1 - 9 - d^{(1)} - d^{(2)} + 1 - d^{(3)} - d^{(\ell)} + d^{(\ell)}d^{(\ell+1)} + d^{(\ell+1)} + N_w^{(\ell)} \quad (19)$$

$$= -9 - d^{(1)} - d^{(2)} - d^{(3)} - d^{(\ell)} + d^{(\ell)}d^{(\ell+1)} + d^{(\ell+1)} + N_w^{(\ell)} \quad (20)$$

$$= -9 - d^{(1)} - d^{(2)} + N_w^{(\ell)} \quad (21)$$

(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

$$N_w|_{k=\ell+1} = (37 - (\ell + 1)) + 9(d^{(1)} - 1) + ((d^{(1)} - 1)(d^{(2)} - 2) + (d^{(2)} - 2)d^{(3)}) \\ + \left(\sum_{i=3}^{\ell-1} d^{(i)}d^{(i+1)}\right) + d^{(\ell)}d^{(\ell+1)} + d^{(\ell+1)} \quad (22)$$

$$= -1 - 9 - 2d^{(1)} - d^{(2)} + 2 - 2d^{(3)} - d^{(\ell)} + d^{(\ell)}d^{(\ell+1)} + d^{(\ell+1)} + N_w^{(\ell)} \quad (23)$$

$$= -8 - 2d^{(1)} - d^{(2)} - 2d^{(3)} - d^{(\ell)} + d^{(\ell)}d^{(\ell+1)} + d^{(\ell+1)} + N_w^{(\ell)} \quad (24)$$

$$= -7 - 2d^{(1)} - d^{(2)} + N_w^{(\ell)} < -9 - d^{(1)} - d^{(2)} + N_w^{(\ell)} \quad (25)$$

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

$$N_w|_{k=\ell+1} = (37 - (\ell + 1)) + 9(d^{(1)} - 2) + ((d^{(1)} - 2)(d^{(2)} - 1) + (d^{(2)} - 1)d^{(3)}) \\ + \left(\sum_{i=3}^{\ell-1} d^{(i)}d^{(i+1)}\right) + d^{(\ell)}d^{(\ell+1)} + d^{(\ell+1)} \quad (26)$$

$$= -1 - 18 - d^{(1)} - 2d^{(2)} + 2 - d^{(3)} - d^{(\ell)} + d^{(\ell)}d^{(\ell+1)} + d^{(\ell+1)} + N_w^{(\ell)} \quad (27)$$

$$= -17 - d^{(1)} - 2d^{(2)} - d^{(3)} - d^{(\ell)} + d^{(\ell)}d^{(\ell+1)} + d^{(\ell+1)} + N_w^{(\ell)} \quad (28)$$

$$= -15 - d^{(1)} - 2d^{(2)} + N_w^{(\ell)} < -9 - d^{(1)} - d^{(2)} + N_w^{(\ell)} \quad (29)$$

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 $N_w \leq 510$.

□

Problem 4

$$\nabla_{\mathbf{w}} \text{err}_n(\mathbf{w}) = \frac{\partial}{\partial \mathbf{w}} \|\mathbf{x}_n - \mathbf{w}\mathbf{w}^T \mathbf{x}_n\|^2 = \frac{\partial}{\partial \mathbf{w}} (\mathbf{x}_n - \mathbf{w}\mathbf{w}^T \mathbf{x}_n)^T (\mathbf{x}_n - \mathbf{w}\mathbf{w}^T \mathbf{x}_n) \quad (30)$$

$$= \frac{\partial}{\partial \mathbf{w}} \left(\mathbf{x}_n^T - \mathbf{x}_n^T (\mathbf{w}\mathbf{w}^T)^T \right) (\mathbf{x}_n - \mathbf{w}\mathbf{w}^T \mathbf{x}_n) \quad (31)$$

$$= (-2\mathbf{x}_n^T \mathbf{w}) (\mathbf{x}_n - \mathbf{w}\mathbf{w}^T \mathbf{x}_n) + (\mathbf{x}_n^T - \mathbf{x}_n^T (\mathbf{w}\mathbf{w}^T)) (-2\mathbf{w}\mathbf{x}_n) \quad (32)$$

$$= -2(\mathbf{x}_n^T \mathbf{w}) \mathbf{x}_n + 2(\mathbf{x}_n^T \mathbf{w}) \mathbf{w} (\mathbf{x}_n^T \mathbf{w})^T - 2(\mathbf{x}_n^T \mathbf{w}) \mathbf{x}_n + 2\mathbf{x}_n^T (\mathbf{w}^T \mathbf{w}) \mathbf{w} \mathbf{x}_n \quad (33)$$

$$= -4(\mathbf{x}_n^T \mathbf{w}) \mathbf{x}_n + 2(\mathbf{x}_n^T \mathbf{w})^2 \mathbf{w} + 2\mathbf{x}_n^T (\mathbf{w}^T \mathbf{w}) \mathbf{w} \mathbf{x}_n \quad (34)$$

where we have used

$$(\mathbf{x}_n^T \mathbf{w}) \mathbf{w} (\mathbf{x}_n^T \mathbf{w})^T = \begin{pmatrix} x_1 & x_2 & \cdots & x_n \end{pmatrix} \begin{pmatrix} w_1 \\ w_2 \\ \vdots \\ w_n \end{pmatrix} \begin{pmatrix} w_1 \\ w_2 \\ \vdots \\ w_n \end{pmatrix} (\mathbf{x}_n^T \mathbf{w})^T \quad (35)$$

$$= c \begin{pmatrix} w_1 \\ w_2 \\ \vdots \\ w_n \end{pmatrix} c = \begin{pmatrix} c^2 w_1 \\ c^2 w_2 \\ \vdots \\ c^2 w_n \end{pmatrix} = (\mathbf{x}_n^T \mathbf{w})^2 \mathbf{w} \quad (36)$$

and

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

and

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

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

□

Problem 5

$$E_{\text{in}}(\mathbf{w}) = \frac{1}{N} \sum_{n=1}^N (\mathbf{x}_n - \mathbf{w}\mathbf{w}^T(\mathbf{x}_n + \boldsymbol{\epsilon}_n))^T (\mathbf{x}_n - \mathbf{w}\mathbf{w}^T(\mathbf{x}_n + \boldsymbol{\epsilon}_n)) \quad (39)$$

$$= \frac{1}{N} \sum_{n=1}^N \left(\mathbf{x}_n^T - (\mathbf{x}_n + \boldsymbol{\epsilon}_n)^T \mathbf{w}^T \mathbf{w} \right) (\mathbf{x}_n - \mathbf{w}\mathbf{w}^T(\mathbf{x}_n + \boldsymbol{\epsilon}_n)) \quad (40)$$

$$= \frac{1}{N} \sum_{n=1}^N \left\| \mathbf{x}_n - \mathbf{w}\mathbf{w}^T \mathbf{x}_n \right\|^2 - \boldsymbol{\epsilon}_n^T \mathbf{w}^T \mathbf{w} (\mathbf{x}_n - \mathbf{w}\mathbf{w}^T \mathbf{x}_n) - \mathbf{w}\mathbf{w}^T \boldsymbol{\epsilon}_n (\mathbf{x}_n^T - \mathbf{x}_n^T \mathbf{w}^T \mathbf{w}) + (\boldsymbol{\epsilon}_n)^2 (\mathbf{w}^T \mathbf{w})^2 \quad (41)$$

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

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

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

□

Problem 6

Claim: $\mathbf{w} = 2(\mathbf{x}_+ - \mathbf{x}_-)$, $b = -\|\mathbf{x}_+\|^2 + \|\mathbf{x}_-\|^2$.

Proof of Claim:

Consider the following cases.

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

$$\mathbf{w}^T \mathbf{x} + b = 2(\mathbf{x}_+^T - \mathbf{x}_-^T) \mathbf{x}_+ + (-\|\mathbf{x}_+\|^2 + \|\mathbf{x}_-\|^2) \quad (43)$$

$$= \|\mathbf{x}_+\|^2 - 2\mathbf{x}_-^T \mathbf{x}_+ + \|\mathbf{x}_-\|^2 = \|\mathbf{x}_+\|^2 - \mathbf{x}_-^T \mathbf{x}_+ - \mathbf{x}_+^T \mathbf{x}_- + \|\mathbf{x}_-\|^2 \quad (44)$$

$$= \|\mathbf{x}_+ - \mathbf{x}_-\|^2 > 0 \quad (45)$$

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

$$\mathbf{w}^T \mathbf{x} + b = 2(\mathbf{x}_+^T - \mathbf{x}_-^T) \mathbf{x}_- + (-\|\mathbf{x}_+\|^2 + \|\mathbf{x}_-\|^2) \quad (46)$$

$$= -\|\mathbf{x}_+\|^2 + 2\mathbf{x}_-^T \mathbf{x}_+ - \|\mathbf{x}_-\|^2 = -\|\mathbf{x}_+\|^2 + \mathbf{x}_-^T \mathbf{x}_+ + \mathbf{x}_+^T \mathbf{x}_- - \|\mathbf{x}_-\|^2 \quad (47)$$

$$= -\|\mathbf{x}_+ - \mathbf{x}_-\|^2 < 0 \quad (48)$$

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

$$\mathbf{w}^T \mathbf{x} + b = 2 (\mathbf{x}_+^T - \mathbf{x}_-^T) (\mathbf{x}_+ + \mathbf{x}_-)/2 + (-\|\mathbf{x}_+\|^2 + \|\mathbf{x}_-\|^2) \quad (49)$$

$$= (\|\mathbf{x}_+\|^2 - \|\mathbf{x}_-\|^2) + (-\|\mathbf{x}_+\|^2 + \|\mathbf{x}_-\|^2) = 0 \quad (50)$$

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 $\|\mathbf{x}_+ - \mathbf{x}'\|^2 > \|\mathbf{x}_- - \mathbf{x}'\|^2$.

$$\mathbf{w}^T \mathbf{x} + b = 2 (\mathbf{x}_+^T - \mathbf{x}_-^T) ((\mathbf{x}_+ + \mathbf{x}_-)/2 + \mathbf{x}') + (-\|\mathbf{x}_+\|^2 + \|\mathbf{x}_-\|^2) \quad (51)$$

$$= 2 (\mathbf{x}_+^T - \mathbf{x}_-^T) \mathbf{x}' \quad (52)$$

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

Hence, we have proved the claim. \square

Problem 7

If g_{RBFNET} outputs +1, which means

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

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

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

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

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

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

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

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

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

Hence

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

\square

Problem 8

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

where Z is

$$Z = \begin{pmatrix} \llbracket \mathbf{x}_1 \neq \mathbf{x}_1 \rrbracket & \llbracket \mathbf{x}_1 \neq \mathbf{x}_2 \rrbracket & \cdots & \llbracket \mathbf{x}_1 \neq \mathbf{x}_N \rrbracket \\ \llbracket \mathbf{x}_2 \neq \mathbf{x}_1 \rrbracket & \llbracket \mathbf{x}_2 \neq \mathbf{x}_2 \rrbracket & \cdots & \llbracket \mathbf{x}_2 \neq \mathbf{x}_N \rrbracket \\ \vdots & \vdots & \ddots & \vdots \\ \llbracket \mathbf{x}_N \neq \mathbf{x}_1 \rrbracket & \llbracket \mathbf{x}_N \neq \mathbf{x}_2 \rrbracket & \cdots & \llbracket \mathbf{x}_N \neq \mathbf{x}_N \rrbracket \end{pmatrix} = \begin{pmatrix} 0 & 1 & \cdots & 1 \\ 1 & 0 & \cdots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \cdots & 0 \end{pmatrix} \quad (62)$$

so

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

Hence we have

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

□

Problem 9

V is initialized as

$$V_{\tilde{d} \times N} = \begin{pmatrix} 1 & 1 & \cdots & 1 \end{pmatrix} \quad (65)$$

so

$$\min_{w_m} \frac{1}{N} \sum_n^N (r_{nm} - w_m v_n)^2 = \min_{w_m} \frac{1}{N} \sum_n^N (r_{nm} - w_m)^2 \quad (66)$$

and we have

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

Hence,

$$w_m = \frac{1}{N} \sum_n^N r_{nm} = \text{average rating of the } m\text{-th movie} \quad (68)$$

□

Problem 10

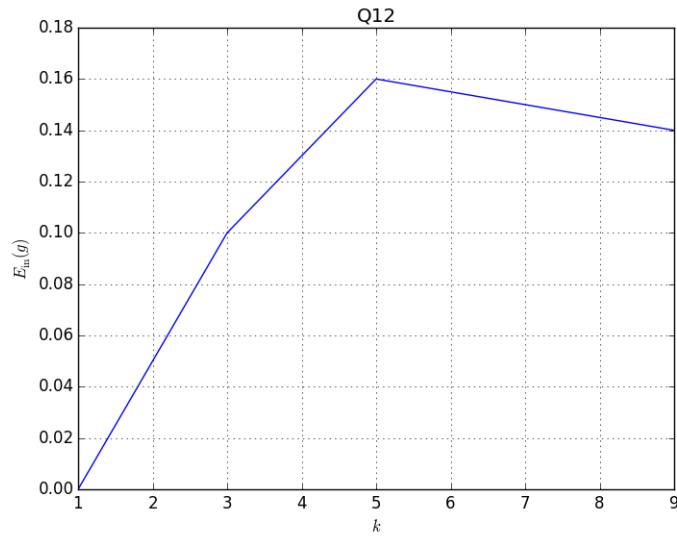
$$\mathbf{v}_{N+1}^T \mathbf{w}_m = \frac{1}{N} \left(\sum_{n=1}^N \mathbf{v}_n^T \right) \mathbf{w}_m = \frac{1}{N} \sum_{n=1}^N \mathbf{v}_n^T \mathbf{w}_m = \frac{1}{N} \sum_{n=1}^N r_{nm} \quad (69)$$

Hence, we have

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

□

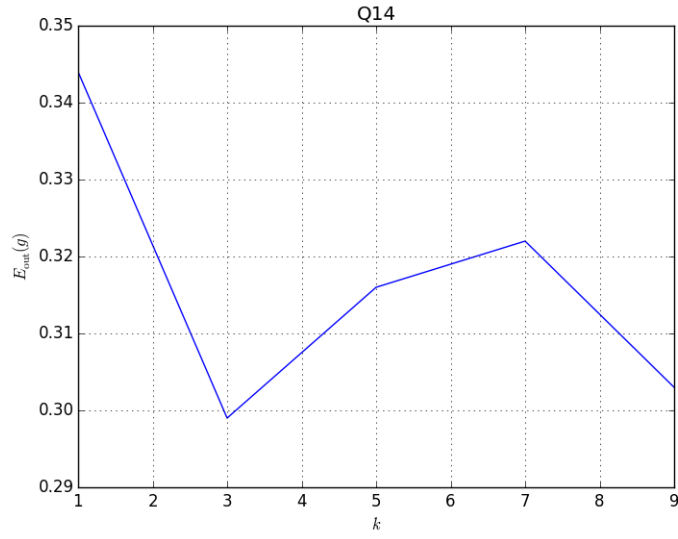
Problem 12



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

□

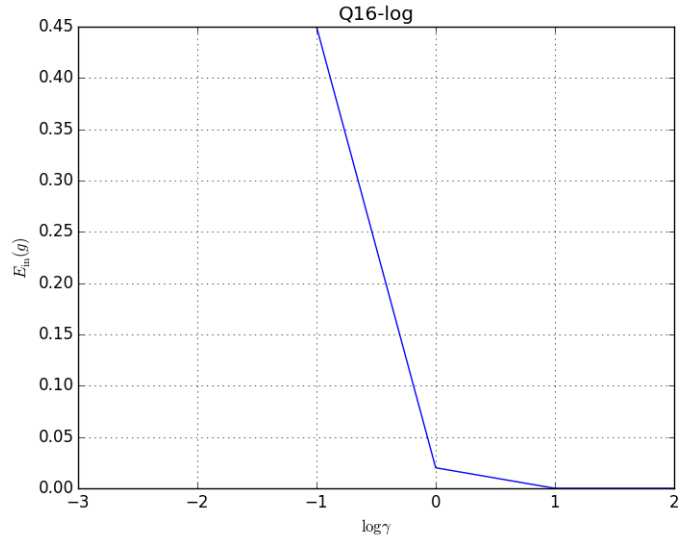
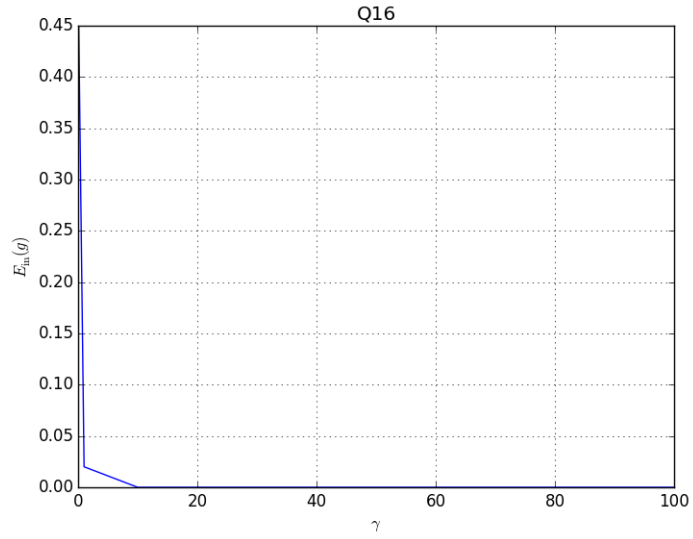
Problem 14



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

□

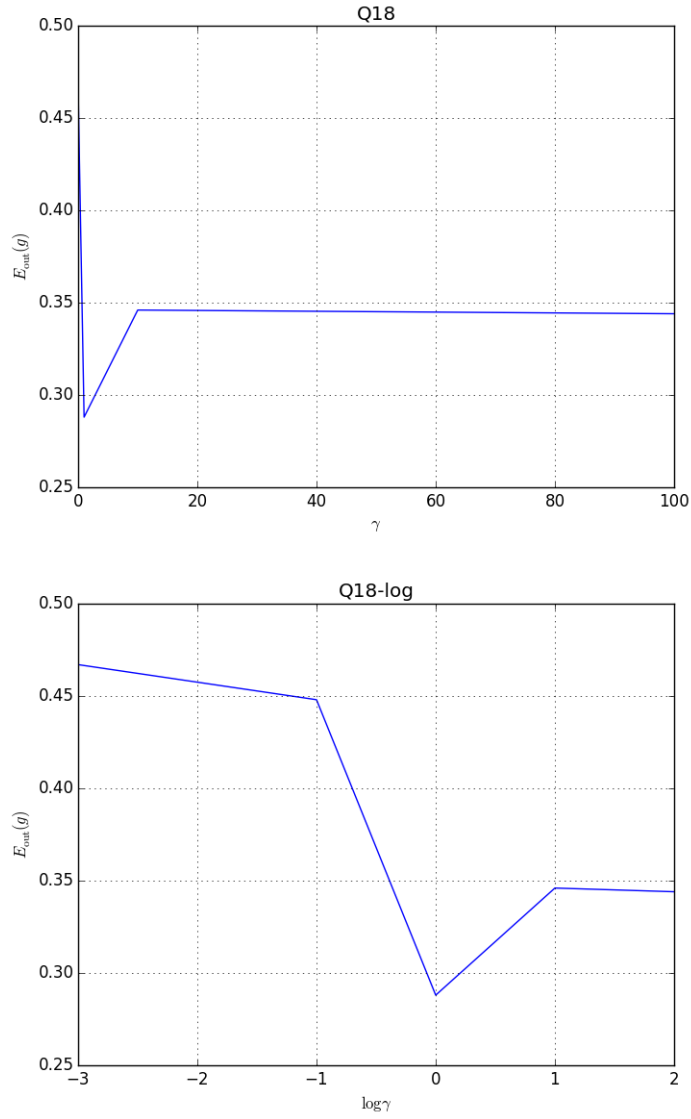
Problem 16



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

□

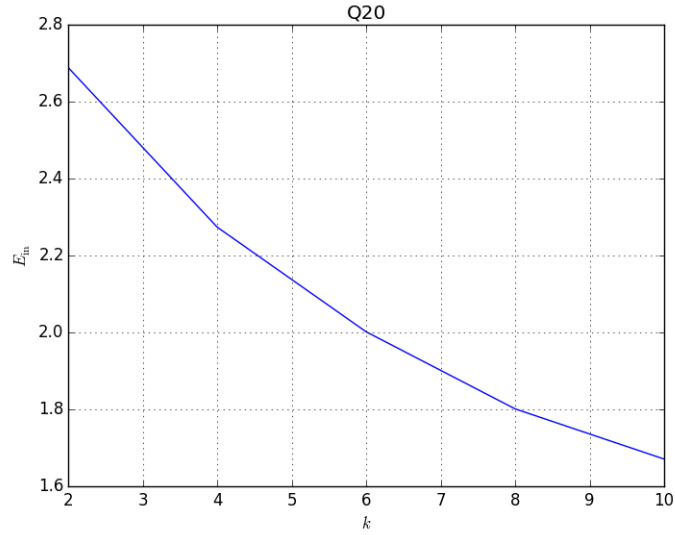
Problem 18



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

□

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

For $N = \min N = 3\Delta \log_2(\Delta)$, we have the following figure

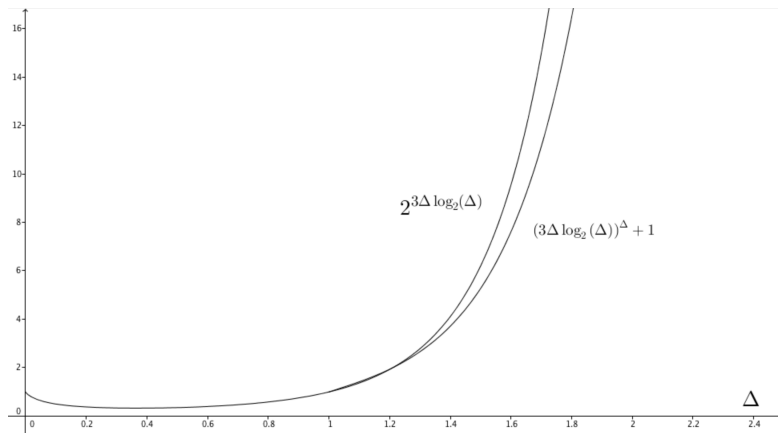


Figure 1: 2^N and $N^{\Delta} + 1$ with $N = 3\Delta \log_2(\Delta)$.

The values of Δ satisfy

$$2^N = 2^{3\Delta \log_2(\Delta)} = (3\Delta \log_2(\Delta))^{\Delta} + 1 = N^{\Delta} + 1 \quad (71)$$

are $\Delta = 1$ and $\Delta \approx 1.21033$. $\forall \Delta \geq 2$, we have

$$(3\Delta \log_2(\Delta))^\Delta + 1 < 2^{3\Delta \log_2(\Delta)} \quad (72)$$

Consider some fixed $\Delta \geq 2$, we have

$$\lim_{N \rightarrow \infty} \frac{N^\Delta + 1}{2^N} = \lim_{N \rightarrow \infty} \frac{\frac{\partial^{[\Delta]} (N^\Delta + 1)}{\partial N^{[\Delta]}}}{\frac{\partial^{[\Delta]} (2^N)}{\partial N^{[\Delta]}}} \quad (73)$$

If $\Delta \in \mathbb{N}$, we have

$$\lim_{N \rightarrow \infty} \frac{\frac{\partial^{[\Delta]} (N^\Delta + 1)}{\partial N^{[\Delta]}}}{\frac{\partial^{[\Delta]} (2^N)}{\partial N^{[\Delta]}}} = \lim_{N \rightarrow \infty} \frac{\Delta!}{(\ln(2))^\Delta 2^N} = 0 \quad (74)$$

if $\Delta \notin \mathbb{N}$, we have

$$\lim_{N \rightarrow \infty} \frac{\frac{\partial^{[\Delta]} (N^\Delta + 1)}{\partial N^{[\Delta]}}}{\frac{\partial^{[\Delta]} (2^N)}{\partial N^{[\Delta]}}} = \lim_{N \rightarrow \infty} \frac{0}{(\ln(2))^{[\Delta]} 2^N} = 0 \quad (75)$$

This result shows that 2^N must be larger than $N^\Delta + 1$ if N is enough. Plot the figure, we have

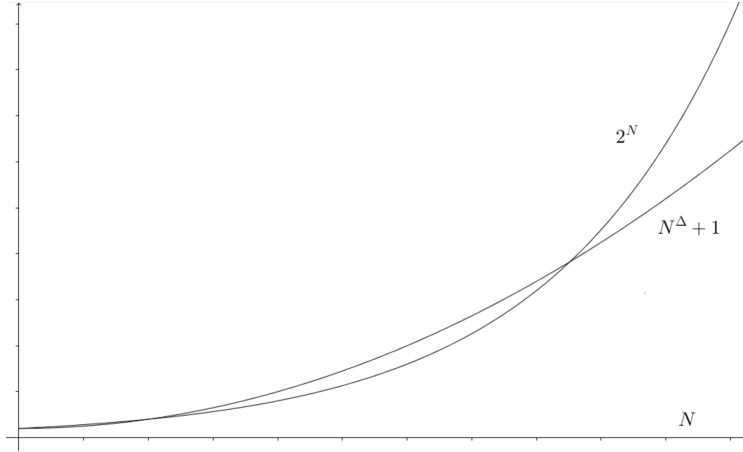


Figure 2: 2^N and $N^\Delta + 1$ for some fixed Δ .

The values of N satisfy $2^N = N^\Delta + 1$ are 1 and $f(\Delta)$ with $f(\Delta) > 1$ and $2^N > N^\Delta + 1$ if $N > f(\Delta)$, $f(\Delta)$ here is the “large enough” point for N .

Claim: $f(\Delta) \leq 3\Delta \log_2(\Delta)$.

Proof of Claim:

Suppose by contradiction, if there exists Δ such that $f(\Delta) > 3\Delta \log_2(\Delta)$, we have

$$2^{3\Delta \log_2(\Delta)} < (3\Delta \log_2(\Delta))^\Delta + 1 \Rightarrow 1 < \Delta < 1.21033... \quad (76)$$

This violates the assumption that $\Delta \geq 2$. So $f(\Delta) \leq 3\Delta \log_2(\Delta)$.

Since $N \geq 3\Delta \log_2(\Delta) \geq f(\Delta)$, we have

$$N^\Delta + 1 < 2^N \text{ if } N \geq 3\Delta \log_2(\Delta), \forall \Delta \geq 2 \quad (77)$$

□

Problem 22

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

$$\text{AND}(x_1^{(1)}, x_2^{(1)}, x_3^{(1)}) \quad (78)$$

And we have $(d+1) \times 3 = 3(d+1)$ weights between input and hidden layer. With one fixed weights of output layer, set $\Delta = 3(d+1) + 1$. The VC dimension of neural network is approximately

$$O\left(\underbrace{3}_{\text{number of layers}} \times \underbrace{(3(d+1) + 1)}_{\text{number of weights}}\right) \quad (79)$$

Set $\Delta := 3(d+1) + 1 \geq 2$ and $O(3(3(d+1) + 1)) = 3\Delta f(\Delta)$.

Claim: $3\Delta f(\Delta) < 3\Delta \log_2(\Delta)$.

Proof of Claim:

Suppose by contradiction, we have

$$3\Delta f(\Delta) \geq 3\Delta \log_2(\Delta) \quad (80)$$

which means

$$m_{\mathcal{H}}(N)|_{N=3\Delta \log_2(\Delta)} = 2^N|_{N=3\Delta \log_2(\Delta)} = 2^{3\Delta \log_2(\Delta)} \quad (81)$$

This is a contradiction since $m_{\mathcal{H}}(N) \propto N^d$ and from the conclusion of Problem 21,

$$N^d < N^\Delta + 1 < 2^N \quad (82)$$

Hence, we have $3\Delta f(\Delta) < 3\Delta \log_2(\Delta)$.

□

Reference

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