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

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

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 $1 \le n_1, n_2 \le 32$, 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 if we have no $d^{(3)}$ layer (which means $n_2 = 0$), then $N_w|_{k=3}$ can be larger.

4. If $18 \ge k = s \ge 4$ with with $(36-s) - \sum_{i=1}^{s-1} n_i$ units (not included $x_0^{(1)}$) in $d^{(1)}$, n_1 units (not included $x_0^{(2)}$) in $d^{(2)}$, n_2 units (not included $x_0^{(3)}$) in $d^{(3)}$,..., n_{s-1} units (not included $x_0^{(s)}$) in $d^{(s)}$ where $n_i \in \mathbb{N}$ and $1 \le n_i \le (35-s)$, $\forall i$, we have

$$N_w|_{k=s} = (37-s) + 9 \times \left((36-s) - \sum_{i=1}^{s-1} n_i\right) + \left(\left((36-s) - \sum_{i=1}^{s-1} n_i\right) \times n_1 + \cdots + n_{s-2} \times n_{s-1}\right) + n_{s-1} \tag{18}$$

We can find that there is no $n_1 n_2$ term, only $n_2 n_3$ term exists and no other terms contains n_2. We have

$$\frac{\partial \left. N_w \right|_{k=s}}{\partial n_2} = n_3 = 0 \text{ as } \left. N_w \right|_{k=s} \text{ reaches maximum} \tag{19}$$

This implies $\left. N_w \right|_{k=s}$ can be larger without $d^{(4)}$ layer. So $k = s \geq 4$ cannot maximize N_w.

Hence, we have $N_w \geq 510$.

$\square$

Problem 4

$$\begin{aligned}
\nabla_{\mathbf{w}} \mathrm{err}_n\left(\mathbf{w}\right) &= \frac{\partial}{\partial \mathbf{w}} \left\| \mathbf{x}_n - \mathbf{w}\mathbf{w}^T\mathbf{x}_n \right\|^2 = \frac{\partial}{\partial \mathbf{w}} \left(\mathbf{x}_n - \mathbf{w}\mathbf{w}^T\mathbf{x}_n\right)^T \left(\mathbf{x}_n - \mathbf{w}\mathbf{w}^T\mathbf{x}_n\right) && (20)\\
&= \frac{\partial}{\partial \mathbf{w}} \left(\mathbf{x}_n^T - \mathbf{x}_n^T\mathbf{w}^T\mathbf{w}\right)\left(\mathbf{x}_n - \mathbf{w}\mathbf{w}^T\mathbf{x}_n\right) && (21)\\
&= \left(-2\mathbf{x}_n^T\mathbf{w}\right)\left(\mathbf{x}_n - \mathbf{w}\mathbf{w}^T\mathbf{x}_n\right) + \left(\mathbf{x}_n^T - \mathbf{x}_n^T\mathbf{w}^T\mathbf{w}\right)\left(-2\mathbf{w}\mathbf{x}_n\right) && (22)\\
&= -2\left(\mathbf{x}_n^T\mathbf{w}\right)\mathbf{x}_n + 2\left(\mathbf{x}_n^T\mathbf{w}\right)\mathbf{w}\left(\mathbf{x}_n^T\mathbf{w}\right)^T - 2\left(\mathbf{x}_n^T\mathbf{w}\right)\mathbf{x}_n + 2\mathbf{x}_n^T\left(\mathbf{w}^T\mathbf{w}\right)\mathbf{w}\mathbf{x}_n && (23)\\
&= -4\left(\mathbf{x}_n^T\mathbf{w}\right)\mathbf{x}_n + 2\left(\mathbf{x}_n^T\mathbf{w}\right)^2\mathbf{w} + 2\mathbf{x}_n^T\left(\mathbf{w}^T\mathbf{w}\right)\mathbf{w}\mathbf{x}_n && (24)
\end{aligned}$$

where we have used

$$\begin{aligned}
\left(\mathbf{x}_n^T\mathbf{w}\right)\mathbf{w}\left(\mathbf{x}_n^T\mathbf{w}\right)^T &= \left(\begin{pmatrix} x_1 & x_2 & \cdots & x_n \end{pmatrix}\begin{pmatrix} w_1 \\ w_2 \\ \vdots \\ w_n \end{pmatrix}\right)\begin{pmatrix} w_1 \\ w_2 \\ \vdots \\ w_n \end{pmatrix}\left(\mathbf{x}_n^T\mathbf{w}\right)^T && (25)\\
&= c\begin{pmatrix} w_1 \\ w_2 \\ \vdots \\ w_n \end{pmatrix} c = \begin{pmatrix} c^2 w_1 \\ c^2 w_2 \\ \vdots \\ c^2 w_n \end{pmatrix} = \left(\mathbf{x}_n^T\mathbf{w}\right)^2\mathbf{w} && (26)
\end{aligned}$$

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

$\square$

Problem 5

$$E_{\text{in}}\left(\mathbf{w}\right) = \frac{1}{N}\sum_{n=1}^{N}\left(\mathbf{x}_n - \mathbf{w}\mathbf{w}^T\left(\mathbf{x}_n + \boldsymbol{\epsilon}_n\right)\right)^T\left(\mathbf{x}_n - \mathbf{w}\mathbf{w}^T\left(\mathbf{x}_n + \boldsymbol{\epsilon}_n\right)\right) \tag{27}$$

$$= \frac{1}{N}\sum_{n=1}^{N}\left(\mathbf{x}_n^T - \left(\mathbf{x}_n + \boldsymbol{\epsilon}_n\right)^T\mathbf{w}^T\mathbf{w}\right)\left(\mathbf{x}_n - \mathbf{w}\mathbf{w}^T\left(\mathbf{x}_n + \boldsymbol{\epsilon}_n\right)\right) \tag{28}$$

$$\begin{aligned}= \frac{1}{N}\sum_{n=1}^{N}&\left\|\mathbf{x}_n - \mathbf{w}\mathbf{w}^T\mathbf{x}_n\right\|^2 - \boldsymbol{\epsilon}_n^T\mathbf{w}^T\mathbf{w}\left(\mathbf{x}_n - \mathbf{w}\mathbf{w}^T\mathbf{x}_n\right)\\ &- \mathbf{w}\mathbf{w}^T\boldsymbol{\epsilon}_n\left(\mathbf{x}_n^T - \mathbf{x}_n^T\mathbf{w}^T\mathbf{w}\right) + \left(\boldsymbol{\epsilon}_n\right)^2\left(\mathbf{w}^T\mathbf{w}\right)^2\end{aligned} \tag{29}$$

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

$$\mathcal{E}\left(E_{\text{in}}\left(\mathbf{w}\right)\right) = \frac{1}{N}\sum_{n=1}^{N}\left\|\mathbf{x}_n - \mathbf{w}\mathbf{w}^T\mathbf{x}_n\right\|^2 + \left(\mathbf{w}^T\mathbf{w}\right)^2 \tag{30}$$

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

□

Problem 6

<u>Claim</u>: $\mathbf{w} = 2\left(\mathbf{x}_+ - \mathbf{x}_-\right)$, $b = -\left\|\mathbf{x}_+\right\|^2 + \left\|\mathbf{x}_-\right\|^2$.

<u>Proof of Claim</u>:

Consider the following cases.

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

$$\mathbf{w}^T\mathbf{x} + b = 2\left(\mathbf{x}_+^T - \mathbf{x}_-^T\right)\mathbf{x}_+ + \left(-\left\|\mathbf{x}_+\right\|^2 + \left\|\mathbf{x}_-\right\|^2\right) \tag{31}$$

$$= \left\|\mathbf{x}_+\right\|^2 - 2\mathbf{x}_-^T\mathbf{x}_+ + \left\|\mathbf{x}_-\right\|^2 = \left\|\mathbf{x}_+\right\|^2 - \mathbf{x}_-^T\mathbf{x}_+ - \mathbf{x}_+^T\mathbf{x}_- + \left\|\mathbf{x}_-\right\|^2 \tag{32}$$

$$= \left\|\mathbf{x}_+ - \mathbf{x}_-\right\|^2 > 0 \tag{33}$$

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

$$\mathbf{w}^T\mathbf{x} + b = 2\left(\mathbf{x}_+^T - \mathbf{x}_-^T\right)\mathbf{x}_- + \left(-\left\|\mathbf{x}_+\right\|^2 + \left\|\mathbf{x}_-\right\|^2\right) \tag{34}$$

$$= -\left\|\mathbf{x}_+\right\|^2 + 2\mathbf{x}_-^T\mathbf{x}_+ - \left\|\mathbf{x}_-\right\|^2 = -\left\|\mathbf{x}_+\right\|^2 + \mathbf{x}_-^T\mathbf{x}_+ + \mathbf{x}_+^T\mathbf{x}_- - \left\|\mathbf{x}_-\right\|^2 \tag{35}$$

$$= -\left\|\mathbf{x}_+ - \mathbf{x}_-\right\|^2 < 0 \tag{36}$$

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

$$\mathbf{w}^T\mathbf{x} + b = 2\left(\mathbf{x}_+^T - \mathbf{x}_-^T\right)(\mathbf{x}_+ + \mathbf{x}_-)/2 + \left(-\|\mathbf{x}_+\|^2 + \|\mathbf{x}_-\|^2\right) \tag{37}$$

$$= \left(\|\mathbf{x}_+\|^2 - \|\mathbf{x}_-\|^2\right) + \left(-\|\mathbf{x}_+\|^2 + \|\mathbf{x}_-\|^2\right) = 0 \tag{38}$$

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

$$\mathbf{w}^T\mathbf{x} + b = 2\left(\mathbf{x}_+^T - \mathbf{x}_-^T\right)((\mathbf{x}_+ + \mathbf{x}_-)/2 + \mathbf{x}') + \left(-\|\mathbf{x}_+\|^2 + \|\mathbf{x}_-\|^2\right) \tag{39}$$

$$= 2\left(\mathbf{x}_+^T - \mathbf{x}_-^T\right)\mathbf{x}' \tag{40}$$

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

$$\beta_+ \exp\left(-\|\mathbf{x} - \boldsymbol{\mu}_+\|^2\right) + \beta_- \exp\left(-\|\mathbf{x} - \boldsymbol{\mu}_-\|^2\right) > 0 \tag{41}$$

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

$$\left|\frac{\beta_+}{\beta_-}\right| \frac{\exp\left(-\|\mathbf{x} - \boldsymbol{\mu}_+\|^2\right)}{\exp\left(-\|\mathbf{x} - \boldsymbol{\mu}_-\|^2\right)} < 1 \tag{42}$$

$$\left|\frac{\beta_+}{\beta_-}\right| \exp\left(-\|\mathbf{x} - \boldsymbol{\mu}_+\|^2 + \|\mathbf{x} - \boldsymbol{\mu}_-\|^2\right) < 1 \tag{43}$$

$$\exp\left(-\|\mathbf{x} - \boldsymbol{\mu}_+\|^2 + \|\mathbf{x} - \boldsymbol{\mu}_-\|^2\right) < \left|\frac{\beta_-}{\beta_+}\right| \tag{44}$$

$$2(\boldsymbol{\mu}_+ - \boldsymbol{\mu}_-)^T\mathbf{x} < \ln\left|\frac{\beta_-}{\beta_+}\right| + \|\boldsymbol{\mu}_+\|^2 - \|\boldsymbol{\mu}_-\|^2 \tag{45}$$

$$2(\boldsymbol{\mu}_- - \boldsymbol{\mu}_+)^T\mathbf{x} + \left(\ln\left|\frac{\beta_-}{\beta_+}\right| + \|\boldsymbol{\mu}_+\|^2 - \|\boldsymbol{\mu}_-\|^2\right) > 0 \tag{46}$$

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

$$2(\boldsymbol{\mu}_- - \boldsymbol{\mu}_+)^T\mathbf{x} + \left(\ln\left|\frac{\beta_-}{\beta_+}\right| + \|\boldsymbol{\mu}_+\|^2 - \|\boldsymbol{\mu}_-\|^2\right) < 0 \tag{47}$$

Hence

$$2(\boldsymbol{\mu}_- - \boldsymbol{\mu}_+),\ b = \ln\left|\frac{\beta_-}{\beta_+}\right| + \|\boldsymbol{\mu}_+\|^2 - \|\boldsymbol{\mu}_-\|^2 \tag{48}$$

□

Problem 8

$$\text{optimal } \beta_n = \left(\left(\mathrm{Z}^T\mathrm{Z}\right)^{-1}\mathrm{Z}^T\mathbf{y}\right)_n \tag{49}$$

where Z is

$$\mathrm{Z} = \begin{pmatrix} [\![\mathbf{x}_1 \neq \mathbf{x}_1]\!] & [\![\mathbf{x}_1 \neq \mathbf{x}_2]\!] & \cdots & [\![\mathbf{x}_1 \neq \mathbf{x}_N]\!] \\ [\![\mathbf{x}_2 \neq \mathbf{x}_1]\!] & [\![\mathbf{x}_2 \neq \mathbf{x}_2]\!] & \cdots & [\![\mathbf{x}_2 \neq \mathbf{x}_N]\!] \\ \vdots & \vdots & \ddots & \vdots \\ [\![\mathbf{x}_N \neq \mathbf{x}_1]\!] & [\![\mathbf{x}_N \neq \mathbf{x}_2]\!] & \cdots & [\![\mathbf{x}_N \neq \mathbf{x}_N]\!] \end{pmatrix} = \begin{pmatrix} 0 & 1 & \cdots & 1 \\ 1 & 0 & \cdots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \cdots & 0 \end{pmatrix} \tag{50}$$

so

$$\left(\mathrm{Z}^T\mathrm{Z}\right)^{-1}\mathrm{Z}^T = \frac{1}{N-1}\begin{pmatrix} -(N-2) & 1 & \cdots & 1 \\ 1 & -(N-2) & \cdots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \cdots & -(N-2) \end{pmatrix} \tag{51}$$

Hence we have

$$\beta_n = \frac{1}{N-1}\left(\sum_{i\neq n} y_i - (N-2)\, y_n\right) \tag{52}$$

□

Problem 9

V is initialized as

$$\mathrm{V}_{\tilde{d}\times N} = \begin{pmatrix} 1 & 1 & \cdots & 1 \end{pmatrix} \tag{53}$$

so

$$\min_{w_m} \frac{1}{N}\sum_{n}^{N}\left(r_{nm} - w_m v_n\right)^2 = \min_{w_m} \frac{1}{N}\sum_{n}^{N}\left(r_{nm} - w_m\right)^2 \tag{54}$$

and we have

$$\frac{\partial}{\partial w_m}\frac{1}{N}\sum_{n}^{N}\left(r_{nm} - w_m\right)^2 = -\frac{2}{N}\sum_{n}^{N}\left(r_{nm} - w_m\right) = -2\left(\left(\frac{1}{N}\sum_{n}^{N} r_{nm}\right) - w_m\right) = 0 \tag{55}$$

Hence,

$$w_m = \frac{1}{N}\sum_{n}^{N} r_{nm} = \text{average rating of the } m\text{-th movie} \tag{56}$$

□

Problem 10

$$\mathbf{v}_{N+1}^T \mathbf{w}_m = \frac{1}{N}\left(\sum_{n=1}^{N} \mathbf{v}_n^T\right)\mathbf{w}_m = \frac{1}{N}\sum_{n=1}^{N} \mathbf{v}_n^T \mathbf{w}_m = \frac{1}{N}\sum_{n=1}^{N} r_{nm} \tag{57}$$

Hence, we have

$$\max_m \mathbf{v}_{N+1}^T \mathbf{w}_m = \max_m \frac{1}{N}\sum_{n=1}^{N} r_{nm} = \text{the movie with largest average rating} \tag{58}$$

□

Problem 11

□

Problem 12

□

Problem 13

□

Problem 14

□

Problem 15

□

Problem 16

□

Problem 17

$\square$

Problem 18

$\square$

Problem 19

$\square$

Problem 20

$\square$

Reference

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