# Machine Learning

Answer Sheet for Homework 3

Da-Min HUANG

R04942045

*Graduate Institute of Communication Engineering, National Taiwan University*

November 16, 2015

## Problem 1

Set $\sigma = 0.1$ and $d = 8$, then we can rewrite the formula to be

$$\mathbb{E}_{\mathcal{D}}\left[E_{\text{in}}\left(\mathbf{w}_{\text{lin}}\right)\right] = 0.01\left(1 - \frac{9}{N}\right) > 0.008 \Rightarrow 0.2 > \frac{9}{N} \Rightarrow N > 45 \Rightarrow N \geq 46 \tag{1}$$

$\square$

---

## Problem 2

(a) $\mathbf{H}$ is positive semi-definite $\Leftrightarrow$ All eigenvalues is non-negative. Refer to choice (c), we have shown the properties.

(b) Refer to (e), $\mathbf{H}$ is idempotent matrix. Suppose $\mathbf{H}^{-1}$ exists, we have

$$\mathbf{H}^{-1}\left(\mathbf{H}^2\right) = \mathbf{H}^{-1}\left(\mathbf{H}\right) \Rightarrow \mathbf{H} = \mathbf{I} \tag{2}$$

Hence, $\mathbf{H}$ is invertible if and only if $\mathbf{H} = \mathbf{I}$. Hence, $\mathbf{H}$ is not always invertible.

(c) Refer to choice (e), we have $\mathbf{H}^2 = \mathbf{H}$. Suppose $\lambda$ is the eigenvalue of some non-zero vector $\vec{v}$,

$$\mathbf{H}\vec{v} = \lambda\vec{v} = \mathbf{H}^2\vec{v} = \mathbf{H}\left(\lambda\vec{v}\right) = \lambda\left(\mathbf{H}\vec{v}\right) = \lambda^2\vec{v} \Rightarrow \lambda^2 = \lambda \tag{3}$$

Hence, the possible results of $\lambda$ is 1 or 0.

(d) For a symmetric and idempotent matrix $\mathbf{H}$, $\operatorname{rank}(\mathbf{H}) = \operatorname{trace}(\mathbf{H})$, the number of non-zero eigenvalues of $\mathbf{H}$.

$$\begin{aligned}\operatorname{trace}(\mathbf{H}) &= \operatorname{trace}\left(\mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\right) = \operatorname{trace}\left(\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{X}\right) &(4)\\ &= \operatorname{trace}\left(\mathbf{I}_{d+1}\right) = d+1 &(5)\end{aligned}$$

where we have used $\operatorname{trace}(\mathbf{AB}) = \operatorname{trace}(\mathbf{BA})$.

Since eigenvalue can only be 0 or 1, so there are $d+1$ eigenvalues of 1.

(e) By the definition of $\mathbf{H}$, we have

$$\begin{aligned}\mathbf{H}^2 &= \left(\mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\right)\left(\mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\right) &(6)\\ &= \mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\left(\left(\mathbf{X}^T\mathbf{X}\right)\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\right)\mathbf{X}^T &(7)\\ &= \mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T = \mathbf{H} &(8)\end{aligned}$$

So

$$\mathbf{H}^2 = \mathbf{H} \Rightarrow \mathbf{H}^{1126} = \mathbf{H} \tag{9}$$

□

---

## Problem 3

If $\operatorname{sign}\left(\mathbf{w}^T\mathbf{x}\right) \neq y$, then $y\mathbf{w}^T\mathbf{x} < 0$ since the sign of $y$ and $\mathbf{w}^T\mathbf{x}$ are different. Similarly, if $\operatorname{sign}\left(\mathbf{w}^T\mathbf{x}\right) = y$, then $y\mathbf{w}^T\mathbf{x} \geq 0$.
Claim: $\left(\max\left(0, 1 - y\mathbf{w}^T\mathbf{x}\right)\right)^2$ is an upper bound.
Proof of claim:
Consider the following cases.

1. $\left[\operatorname{sign}\left(\mathbf{w}^T\mathbf{x}\right) \neq y\right] = 0$

   Then $y\mathbf{w}^T\mathbf{x} \geq 0$. Hence $\left(\max\left(0, 1 - y\mathbf{w}^T\mathbf{x}\right)\right)^2 \geq 0$, which bounds $\left[\operatorname{sign}\left(\mathbf{w}^T\mathbf{x}\right) \neq y\right]$.

2. $\left[\operatorname{sign}\left(\mathbf{w}^T\mathbf{x}\right) \neq y\right] = 1$

   Then $y\mathbf{w}^T\mathbf{x} < 0$. Hence $\left(\max\left(0, 1 - y\mathbf{w}^T\mathbf{x}\right)\right)^2 = \left(1 - y\mathbf{w}^T\mathbf{x}\right)^2 > 1$, which bounds $\left[\operatorname{sign}\left(\mathbf{w}^T\mathbf{x}\right) \neq y\right]$.

□

---

## Problem 4

Set $y\mathbf{w}^T\mathbf{x} := z$. Consider $\max\left(0, -y\mathbf{w}^T\mathbf{x}\right) = \max\left(0, -z\right) := f\left(z\right)$. We have $f\left(z\right) = -z$ if $z \leq 0$, else $f\left(z\right) = 0$. So

$$\lim_{z\to 0^-} \frac{f\left(z\right) - f\left(0\right)}{z - 0} = \frac{-z - 0}{z - 0} = -1, \quad \lim_{z\to 0^+} \frac{f\left(z\right) - f\left(0\right)}{z - 0} = \frac{0 - 0}{z - 0} = 0 \tag{10}$$

Hence, $f\left(z\right)$ is not differentiable at $z = 0$.

$\square$

---

## Problem 5

<u>Calim</u>: $\max\left(0, -y\mathbf{w}^T\mathbf{x}\right)$ results in PLA.
<u>Proof of claim</u>:
Consider following cases.

1. $y = \text{sign}\left(\mathbf{w}^T\mathbf{x}\right)$. PLA update term is 0.

   Then we have $y\mathbf{w}^T\mathbf{x} > 0$. So

   $$\max\left(0, -y\mathbf{w}^T\mathbf{x}\right) = 0 \tag{11}$$

2. $y \neq \text{sign}\left(\mathbf{w}^T\mathbf{x}\right)$. PLA update term is $y\mathbf{x}$.

   Then we have $y\mathbf{w}^T\mathbf{x} < 0$. So

   $$\max\left(0, -y\mathbf{w}^T\mathbf{x}\right) = -y\mathbf{w}^T\mathbf{x} \Rightarrow -\nabla_{\mathbf{w}} \max\left(0, -y\mathbf{w}^T\mathbf{x}\right) = y\mathbf{x} \tag{12}$$

$\square$

---

## Problem 6

$$\begin{aligned}
\nabla E(0,0) &= \left(\frac{\partial E}{\partial u}, \frac{\partial E}{\partial v}\right)\bigg|_{(0,0)} && (13)\\
&= \left(e^u + ve^{uv} + 2u - 2v - 3, 2e^{2v} + ue^{uv} - 2u + 4v - 2\right)\big|_{(0,0)} && (14)\\
&= (-2, 0) && (15)
\end{aligned}$$

$\square$

---

## Problem 7

I write a program Q07.py to calculate the result by using

$$(u_{t+1}, v_{t+1}) = (u_t, v_t) - 0.01\nabla E\left(u_t, v_t\right) \tag{16}$$

iteratively.

$$\begin{aligned}
(u_1, v_1) &= (0,0) - 0.01\nabla E\left(0,0\right) = (0.02, 0) && (17)\\
(u_2, v_2) &= (0.02,0) - 0.01\nabla E\left(0.02,0\right) \approx (0.039398, 0.0002) && (18)\\
(u_3, v_3) &\approx (0.039398, 0.0002) - 0.01\nabla E\left(0.039398, 0.0002\right) && (19)\\
&\approx (0.0582102, 0.000577975) && (20)\\
(u_4, v_4) &\approx (0.0764524, 0.00111381) && (21)\\
(u_5, v_5) &\approx (0.09414, 0.00178911) && (22)\\
E\left(u_5, v_5\right) &\approx 2.825 && (23)
\end{aligned}$$

□

---

## Problem 8

$$\begin{aligned}
\nabla E(0,0) &= \left(\frac{\partial E}{\partial u}, \frac{\partial E}{\partial v}\right) && (24)\\
&= \left(e^u + ve^{uv} + 2u - 2v - 3, 2e^{2v} + ue^{uv} - 2u + 4v - 2\right) && (25)
\end{aligned}$$

From this we compute the Hessian matrix

$$\nabla^2 E\left(u,v\right) = \begin{pmatrix} e^u + v^2 e^{uv} + 2 & \left(uv+1\right)e^{uv} - 2 \\ \left(uv+1\right)e^{uv} - 2 & 4e^{2v} + u^2 e^{uv} + 4 \end{pmatrix} \tag{26}$$

So

$$\begin{aligned}
\hat{E}\left(\Delta u, \Delta v\right) &= E\left(0,0\right) + \nabla E\left(0,0\right)\cdot\left(\Delta u, \Delta v\right) + \frac{1}{2}\left(\Delta u, \Delta v\right)\nabla^2 E\left(0,0\right)\begin{pmatrix}\Delta u \\ \Delta v\end{pmatrix} && (27)\\
&= 3 - 2\Delta u + \frac{1}{2}\left(\Delta u, \Delta v\right)\begin{pmatrix} 3 & -1 \\ -1 & 8\end{pmatrix}\begin{pmatrix}\Delta u \\ \Delta v\end{pmatrix} && (28)\\
&= \frac{3}{2}\left(\Delta u\right)^2 + 4\left(\Delta v\right)^2 - \Delta u \Delta v - 2\Delta u + 0\Delta v + 3 && (29)
\end{aligned}$$

□

---

## Problem 9

Claim: $-\left(\nabla^2 E(u,v)\right)^{-1} \nabla E(u,v)$ is the Newton direction.
Proof of claim:

$$\frac{\partial \hat{E}(\Delta u, \Delta v)}{\partial (\Delta u, \Delta v)} = \nabla E(u,v) + \nabla^2 E(u,v)(\Delta u, \Delta v) = 0 \tag{30}$$

$$\Rightarrow (\Delta u, \Delta v) = -\left(\nabla^2 E(u,v)\right)^{-1} \nabla E(u,v) \tag{31}$$

□

---

## Problem 10

I write a program Q10.py to calculate the result by using

$$(u_{t+1}, v_{t+1}) = (u_t, v_t) - \left(\nabla^2 E(u_t, v_t)\right)^{-1} \nabla E(u_t, v_t) \tag{32}$$

iteratively.

$$(u_1, v_1) \approx (0.695652173913, 0.0869565217391) \tag{33}$$

$$(u_2, v_2) \approx (0.613762221112, 0.0711078990173) \tag{34}$$

$$(u_3, v_3) \approx (0.611812859879, 0.0705000613365) \tag{35}$$

$$(u_4, v_4) \approx (0.611811717261, 0.0704995471019) \tag{36}$$

$$(u_5, v_5) \approx (0.61181171726, 0.0704995471016) \tag{37}$$

$$E(u_5, v_5) \approx 2.36082334564 \tag{38}$$

This equals to the value of Problem 7 after 746 updates.

□

---

## Problem 11

Write a program Q11.py to test, $\{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4, \mathbf{x}_5, \mathbf{x}_6\}$ is the biggest subset that can be shattered by the union of quadratic, linear, or constant hypotheses of $\mathbf{x}$ by feature form of

$$\Phi(\mathbf{x}) = \left(1, x_1, x_2, x_1^2, x_1 x_2, x_2^2\right) \tag{39}$$

Then I ran the PLA and all cases are stasified. The $\mathbf{w}$ for $2^6 = 64$ cases are

$$
\begin{array}{llll}
[0,0,0,0,0,0]\,, & [0,1,0,1,0,-4]\,, & [1,0,1,-2,-1,-1]\,, & [2,0,0,0,0,-2]\,, \\
[-1,-2,0,0,-2,0]\,, & [0,0,1,2,-1,-3]\,, & [1,-1,2,-1,-2,0]\,, & [2,-1,1,1,-1,-3]\,, \\
[0,0,-2,0,2,0]\,, & [0,0,-3,2,3,-3]\,, & [1,-1,-2,-1,2,0]\,, & [2,0,-2,0,2,-2]\,, \\
[-1,-4,0,0,0,0]\,, & [0,-2,-1,4,1,-3]\,, & [1,-2,-1,0,1,1]\,, & [3,-2,0,2,0,-4]\,, \\
[-1,1,0,-1,-2,0]\,, & [0,2,0,0,-2,0]\,, & [1,1,-2,-3,-2,0]\,, & [1,1,-1,-1,-3,-1]\,, \\
[-1,0,1,0,-3,1]\,, & [0,1,0,1,-4,0]\,, & [1,-1,1,-1,-3,1]\,, & [1,0,0,0,-4,0]\,, \\
[0,0,-3,0,1,1]\,, & [0,1,-4,1,0,0]\,, & [1,-1,-3,-1,1,1]\,, & [1,0,-4,0,0,0]\,, \\
[0,-1,-2,1,0,2]\,, & [0,0,-3,2,-1,1]\,, & [1,-1,-1,-1,-1,1]\,, & [1,-1,-3,1,-1,1]\,, \\
[0,0,2,0,2,0]\,, & [0,2,3,0,1,-1]\,, & [1,0,3,-2,1,-1]\,, & [1,1,3,-1,1,-1]\,, \\
[-1,0,4,0,0,0]\,, & [0,1,4,1,0,0]\,, & [1,-1,4,-1,0,0]\,, & [1,0,4,0,0,0]\,, \\
[0,0,0,0,4,0]\,, & [0,1,0,1,4,0]\,, & [1,-1,0,-1,4,0]\,, & [1,0,0,0,4,0]\,, \\
[0,-1,1,1,3,1]\,, & [0,0,1,2,3,1]\,, & [1,-2,1,0,3,1]\,, & [1,-1,1,1,3,1]\,, \\
[-2,2,0,-2,0,4]\,, & [0,4,0,0,0,0]\,, & [1,2,1,-4,-1,3]\,, & [2,4,0,0,0,0]\,, \\
[-1,0,2,0,-2,2]\,, & [0,1,3,1,-1,1]\,, & [1,0,3,-2,-3,3]\,, & [1,1,3,1,-1,1]\,, \\
[-1,1,-1,-1,1,3]\,, & [0,3,-1,1,1,1]\,, & [1,0,-1,-2,1,3]\,, & [1,2,0,0,2,0]\,, \\
[-1,0,0,0,0,2]\,, & [0,1,1,1,1,1]\,, & [1,-1,0,-1,0,4]\,, & [1,1,1,1,1,1]
\end{array}
\tag{40}
$$

□

---

## Problem 12

By the transform, we have $(\Phi(\mathbf{x}))_i = z_i = \left(0, \ldots, \underbrace{1}_{i\text{-th term}}, \ldots, 0\right)$. To shatter the original $N$ points, we can assign $w_i$ to be positive or negative to get $\mathbf{x}_i$ $\circ$ or $\times$.
So, this transform shatter any $N$ points. Hence $d_{vc}(\mathcal{H}_\Phi) = \infty$.

□

---

## Problem 13

The average $E_{\text{in}}$ is 0.503979.
After feature transform, we have

$$
\begin{aligned}
\tilde{\mathbf{w}} = \big[&-0.991720295, -3.00273423 \times 10^{-4}, -1.31851902 \times 10^{-3}, \\
&1.09524038 \times 10^{-3}, 1.55703088, 1.55594342\big]
\end{aligned}
\tag{41}
$$

with average $E_{\text{out}} = 0.124126$.

□

---

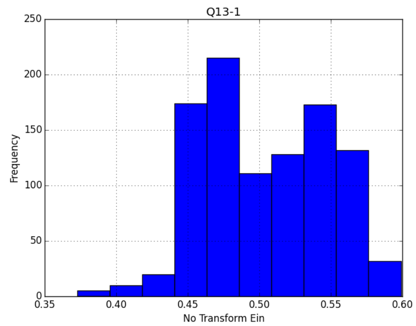


Figure 1: Q13 histogram

## Problem 14

The average $\tilde{\mathbf{w}}_3 = 0.06306984$.

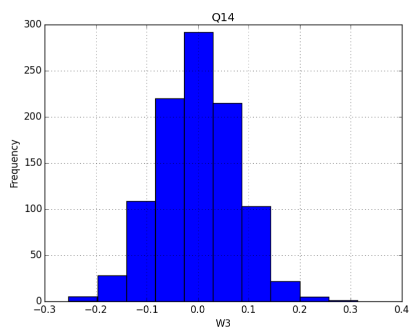


Figure 2: Q14 histogram

□

---

## Problem 15

The average $E_{\text{out}} = 0.123968$.

□

---

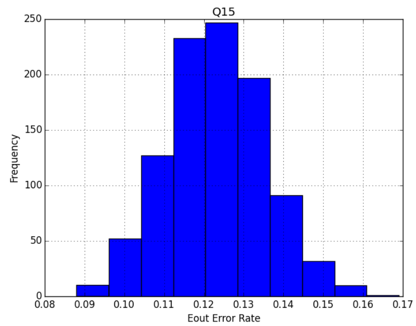


Figure 3: Q15 histogram

## Problem 16

Sum the minimized negative log likelihood $h_y$, which is $\min_y(-\ln(h_y))$, we have

$$E_{\text{in}} = \frac{1}{N}\sum_{n=1}^{N}\left(-\ln\left(\frac{\exp\left(\mathbf{w}_{y_n}^T\mathbf{x}_n\right)}{\sum_{i=1}^{K}\exp\left(\mathbf{w}_i^T\mathbf{x}_n\right)}\right)\right) \tag{42}$$

$$= \frac{1}{N}\sum_{n=1}^{N}\left(\ln\left(\sum_{i=1}^{K}\exp\left(\mathbf{w}_i^T\mathbf{x}_n\right)\right) - \mathbf{w}_{y_n}^T\mathbf{x}_n\right) \tag{43}$$

□

---

## Problem 17

$$\frac{\partial E_{\text{in}}}{\partial \mathbf{w}_i} = \frac{1}{N}\sum_{n=1}^{N}\frac{\partial}{\partial \mathbf{w}_i}\left(\ln\left(\sum_{i=1}^{K}\exp\left(\mathbf{w}_i^T\mathbf{x}_n\right)\right) - \mathbf{w}_{y_n}^T\mathbf{x}_n\right) \tag{44}$$

$$= \frac{1}{N}\sum_{n=1}^{N}\left(\frac{1}{\sum_{i=1}^{K}\exp\left(\mathbf{w}_i^T\mathbf{x}_n\right)}\frac{\partial}{\partial \mathbf{w}_i}\exp\left(\mathbf{w}_i^T\mathbf{x}_n\right) - \frac{\partial}{\partial \mathbf{w}_i}\mathbf{w}_{y_n}^T\mathbf{x}_n\right) \tag{45}$$

$$= \frac{1}{N}\sum_{n=1}^{N}\left(\frac{\exp\left(\mathbf{w}_i^T\mathbf{x}_n\right)}{\sum_{i=1}^{K}\exp\left(\mathbf{w}_i^T\mathbf{x}_n\right)}\mathbf{x}_n - [\![y_n = i]\!]\,\mathbf{x}_n\right) \tag{46}$$

$$= \frac{1}{N}\sum_{n=1}^{N}\left(h_i\left(\mathbf{x}_n\right) - [\![y_n = i]\!]\right)\mathbf{x}_n \tag{47}$$

□

---

## Problem 18

The $E_{\text{out}} = 0.475$ with

$$\begin{aligned}\tilde{\mathbf{w}} = [&0.01878417, -0.01260595, 0.04084862, -0.03266317, 0.01502334, \\ &-0.03667437, 0.01255934, 0.04815065, -0.02206419, 0.02479605, \\ &0.06899284, 0.0193719, -0.01988549, -0.0087049, 0.04605863, \\ &0.05793382, 0.061218, -0.04720391, 0.06070375, -0.01610907, -0.03484607]\end{aligned} \tag{48}$$

$\square$

---

## Problem 19

The $E_{\text{out}} = 0.220$ with

$$\begin{aligned}\tilde{\mathbf{w}} = [&-0.00385379, -0.18914564, 0.26625908, -0.35356593, 0.04088776, \\ &-0.3794296, 0.01982783, 0.33391527, -0.26386754, 0.13489328, \\ &0.4914191, 0.08726107, -0.25537728, -0.16291797, 0.30073678, \\ &0.40014954, 0.43218808, -0.46227968, 0.43230193, -0.20786372, -0.36936337]\end{aligned} \tag{49}$$

$\square$

---

## Problem 20

The $E_{\text{out}} = 0.473$ with

$$\begin{aligned}\tilde{\mathbf{w}} = [&0.01826899, -0.01308051, 0.04072894, -0.03295698, 0.01498363, \\ &-0.03691042, 0.01232819, 0.04791334, -0.02244958, 0.02470544, \\ &0.06878235, 0.01897378, -0.02032107, -0.00901469, 0.04589259, \\ &0.05776824, 0.06102487, -0.04756147, 0.06035018, -0.01660574, -0.03509342]\end{aligned} \tag{50}$$

$\square$

---

## Problem 21

$$\mathbf{h}^T \mathbf{y} = \sum_{i=1}^{N} h\left(\mathbf{x}_i\right) y_i \tag{51}$$

We just need two times queries to obtain $\mathbf{h}^T\mathbf{y}$.

First, take some $h'$ such that $h'(\mathbf{x}_i) = 0$, $\forall i$. Then we have

$$\mathrm{RMSE}(h') = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - h'(\mathbf{x_i}))^2} = \sqrt{\frac{1}{N}\sum_{i=1}^{N} y_i^2} \Rightarrow \sum_{i=1}^{N} y_i^2 = N(\mathrm{RMSE}(h'))^2 \tag{52}$$

Second, query for some $h$,

$$\mathrm{RMSE}(h) = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - h(\mathbf{x_i}))^2} \tag{53}$$

$$N(\mathrm{RMSE}(h))^2 = \sum_{i=1}^{N}(y_i - h(\mathbf{x_i}))^2 = \sum_{i=1}^{N} y_i^2 - 2h(\mathbf{x}_i)y_i + h^2(\mathbf{x_i}) \tag{54}$$

$$\Rightarrow \mathbf{h}^T\mathbf{y} = -\frac{1}{2}\left(N(\mathrm{RMSE}(h) - \mathrm{RMSE}(h'))^2 - \sum_{i=1}^{N} h^2(\mathbf{x_i})\right) \tag{55}$$

where we have known the value of $\sum_{i=1}^{N} h^2(\mathbf{x_i})$ since we have already known all $\mathbf{x}_i$.

□

---

## Problem 22

To find $\min_{\mathbf{w}} \mathrm{RMSE}(H)$, we need to find $\nabla\mathrm{RMSE}(H) = 0$.

$$\nabla\mathrm{RMSE}(H) = 0 \Rightarrow \frac{\partial}{\partial w_k}\sum_{i=1}^{N}\left(y_i - \sum_{k=1}^{K} w_k h_k(\mathbf{x_i})\right)^2 = 0, \forall k \tag{56}$$

$$\Rightarrow 2\sum_{i=1}^{N}\left(h_k(\mathbf{x}_i)y_i - h_k(\mathbf{x}_i)\sum_{k=1}^{K} w_k h_k(\mathbf{x_i})\right) = 0, \forall k \tag{57}$$

Hence, we need to know the value of $\mathbf{h}_k^T\mathbf{y}$ for all $k$. Hence, follow the conclusion above, we need to query for $K+1$ times, where the $+1$ is for querying the value of $\sum_{i=1}^{N} y_i^2$. The derivation steps are

1. First, use some hypothesis $h'$ such that $h'(\mathbf{x}_i) = 0$, $\forall i$ to get the value of $\sum_{i=1}^{N} y_i^2$.

2. Second, query for $\mathrm{RMSE}(h_k)$, $\forall k$, costs $K$ times query. Then we have all $\mathbf{h}_k^T\mathbf{y}$.

Hence, we have

$$\mathbf{h}_k^T\mathbf{y} - \sum_{i=1}^{N}\left(h_k(\mathbf{x}_i)\sum_{k=1}^{K} w_k h_k(\mathbf{x}_i)\right) = 0,\ \forall k \tag{58}$$

Then we can solve the value of $\mathbf{w}$ to get the minimized value.

□

---

# Reference

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