

## Team Evaluation

1. Brendan Rizzo
  - a. Worked on Question 1 and came up with a great answer
2. Richard Huang
  - a. Richard solved question 2 with little help by Eden.
3. Jiashang Cao
  - a. I don't know where Jiashang went.
4. **Eden Seo**
  - a. Helped Richard little bit on solving question 2.

## 1. Question 1 (Exercise 3.6 in LFD)

a.

3.6 (a)  $y_n = +1 \rightarrow P(y_n/x_n) = h(x_n)$   
 $y_n = -1 \rightarrow P(y_n/x_n) = 1 - h(x_n)$

Maximum likelihood

$$\prod_{n=1}^N P(y_n/x_n) = \sum_{n=1}^N \ln(P(y_n/x_n)) = - \sum_{n=1}^N \underbrace{P(y_n/x_n)}_{\text{minimize}}$$

$$E_{in}(w) = - \sum_{n=1}^N \ln(P(y_n/x_n))$$

$$= - \sum_{n=1}^N I(y_n = +1) \ln h(x_n) + I(y_n = -1) \ln(1 - h(x_n))$$

$$= \sum_{n=1}^N I(y_n = +1) \ln \frac{1}{h(x_n)} + I(y_n = -1) \ln \frac{1}{1 - h(x_n)}$$

(b)  $h(x) = \theta(w^T x) = \frac{e^{w^T x}}{1 + e^{w^T x}}$

$$\ln \frac{1}{h(x_n)} = \ln(1 + e^{w^T x_n})$$

$$\ln \frac{1}{1 - h(x_n)} = \ln(1 + e^{w^T x_n})$$

$$E_{in}(w) = \sum_{n=1}^N I(y_n = +1) \ln(1 + e^{-w^T x_n}) + I(y_n = -1) \ln(1 + e^{w^T x_n})$$

$$= \sum_{n=1}^N \ln(1 + e^{-y_n w^T x_n})$$

Minimizing this ~~sample~~ sample is the same as minimizing 3.9  $\rightarrow E_{in}(w) = \frac{1}{N} \sum_{n=1}^N \ln(1 + e^{-y_n w^T x_n})$

## 2. Question 2 (Exercise 3.7 in LFD)

3.7

$$E(w) = \frac{1}{N} \sum_{n=1}^N \ln(1 + \exp(-y_n w^T x_n))$$

taking the derivative wrt "w" we get

$$-\frac{1}{N} \sum_{n=1}^N \frac{y_n x_n \exp(-y_n w^T x_n)}{1 + \exp(-y_n w^T x_n)} \times \frac{1 + \exp(y_n w^T x_n)}{1 + \exp(y_n w^T x_n)}$$

$$= -\frac{1}{N} \sum_{n=1}^N \frac{y_n x_n}{1 + \exp(-y_n w^T x_n)}$$

$$= -\frac{1}{N} \sum_{n=1}^N y_n x_n \theta(-y_n w^T x_n)$$

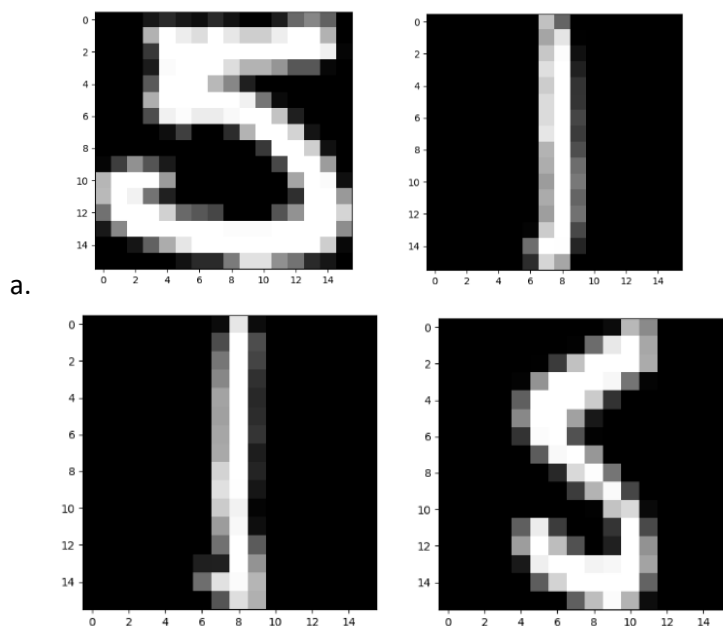
$$= \frac{1}{N} \sum_{n=1}^N -y_n x_n \theta(-y_n w^T x_n)$$

$\theta(s) = \frac{e^s}{1+e^s} = \frac{1}{1+e^{-s}}$   
 $[\ln(u)]' = \frac{1}{u} \cdot u'$   
 $[e^u]' = e^u \cdot u'$

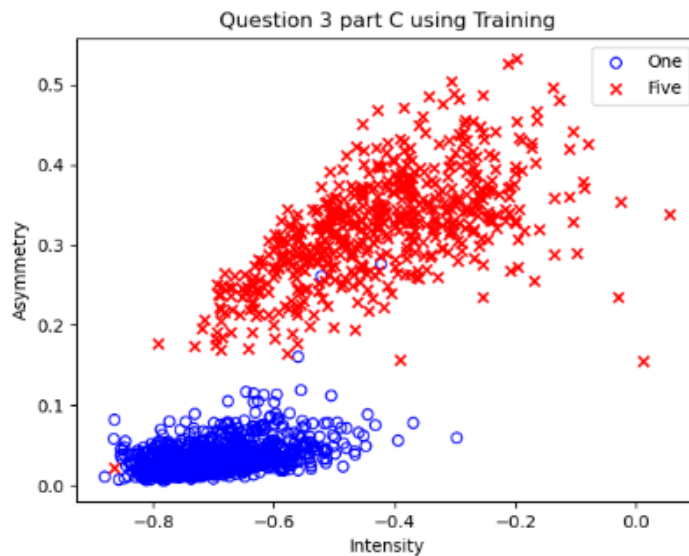
The reason why a misclassified sample contributes more to the gradient is because when misclassified  $y_n w^T x_n < 0 \therefore \theta(-y_n w^T x_n) > 0.5$  while a sample that is correctly classified will have  $\theta(-y_n w^T x_n) < 0.5$ . So the misclassified sample contributes more to the gradient.

a.

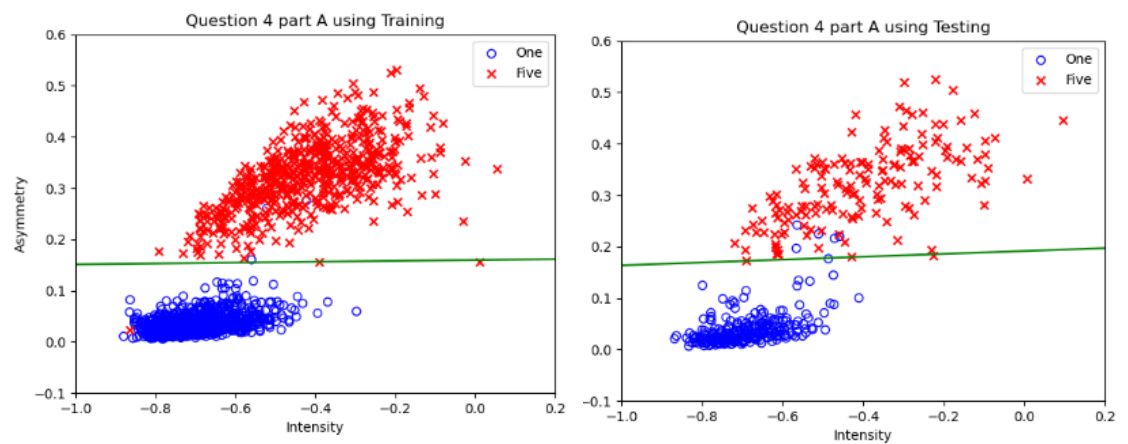
## 3. Question 3



- b.
- Intensity of One: [-0.7539140625, -0.77228125, ... -0.44755859374999996]  
 Intensity of Five: [-0.11173828124999999, -0.56403515625, ... -0.53423828125]  
 Symmetry of One: [0.029765625, 0.035273437500000004, ... 0.04910546875]  
 Symmetry of Five: [0.42023828124999996, 0.21845703125, ... 0.29608203125]



- c.  
4. Question 4  
a.



- b.

```
E_in: 0.00107905544053368
E_test: 0.006148133462409647
```

- c.

```
E_train_3rd: 0.0019084895331687268
E_test_3rd: 0.009466150225192
```

- d. We should not use 3<sup>rd</sup> order transform since it has higher E values than the original