Assignment 6

Ruochen Kong, 661947549, CSCI 4100

1. **Exercise 3.4**

   (a) Since $X = [\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, ...\mathbf{x}_N]^\top$, $\mathbf{y} = [y_1, y_2, y_3, ..., y_N]^\top$
   we can rewrite $y = \mathbf{w}^{*\top}\mathbf{x} + \epsilon$ to $\mathbf{y} = X\mathbf{w}^* + \epsilon$.
   As $\mathbf{w}_{lin} = (X^\top X)^{-1}X^\top \mathbf{y}$ and $H = X(X^\top X)^{-1}X^\top$,

$$\begin{aligned}
\hat{\mathbf{y}} &= X\mathbf{w}_{lin} \\
&= X(X^\top X)^{-1}X^\top \mathbf{y} \\
&= X(X^\top X)^{-1}X^\top (X\mathbf{w}^* + \epsilon) \\
&= X(X^\top X)^{-1}(X^\top X)\mathbf{w}^* + X(X^\top X)^{-1}X^\top \epsilon \\
&= X\mathbf{w}^* + H\epsilon
\end{aligned}$$

   (b)

$$\begin{aligned}
\hat{\mathbf{y}} - \mathbf{y} &= X\mathbf{w}^* + H\epsilon - (X\mathbf{w}^* + \epsilon) \\
&= H\epsilon - \epsilon \\
&= (H - I)\epsilon
\end{aligned}$$

   (c) From exercises 3.3 we have:

$$(I - H)^K = I - H$$

   So,

$$\begin{aligned}
E_{in}(\mathbf{w}_{lin}) &= \frac{1}{N}||x\mathbf{w} - \mathbf{y}||^2 \\
&= \frac{1}{N}||\hat{\mathbf{y}} - \mathbf{y}||^2 \\
&= \frac{1}{N}||(H - I)\epsilon||^2 \\
&= \frac{1}{N}((H - I)\epsilon)^\top((H - I)\epsilon) \\
&= \frac{1}{N}\epsilon^\top(H - I)^2\epsilon, \quad \textit{Since H is symmetric} \\
&= \frac{1}{N}\epsilon^\top(I - H)\epsilon
\end{aligned}$$

(d) In exercise 3.3,

$$
\begin{aligned}
trace(AB) &= trace(BA) \\
trace(H) &= trace(X(X^\top X)^{-1}X^\top) \\
&= trace(X^\top X(X^\top X)^{-1}) \\
&= trace(I_{d+1}) \\
&= d+1 \\
\mathbb{E}_{\mathcal{D}}[E_{in}(\mathbf{w}_{lin})] &= \frac{1}{N}\mathbb{E}_{\mathcal{D}}[\epsilon^T(I-H)\epsilon] \\
&= \frac{1}{N}\mathbb{E}_{\mathcal{D}}[\epsilon^T\epsilon - \epsilon^T H\epsilon] \\
&= \frac{1}{N}\mathbb{E}_{\mathcal{D}}(\sum_{i=1}^{N}\epsilon_i^2) - \mathbb{E}_{\mathcal{D}}(\sum_{i=1}^{N}\sum_{j=1}^{N}\epsilon_i H_{ij}\epsilon_j) \\
&= \frac{1}{N}\mathbb{E}_{\mathcal{D}}(\sum_{i=1}^{N}\epsilon_i^2) - \mathbb{E}_{\mathcal{D}}(\sum_{i=1}^{N}\epsilon_i^2 H_{ii}) \\
&= \frac{1}{N}(N\sigma^2 - \sum_{i=1}^{N}H_{ii}\sigma^2) \\
&= \sigma^2 - \frac{trace(H)}{N}\sigma^2 \\
&= \sigma^2 - \frac{d+1}{N}\sigma^2 \\
&= \sigma^2(1-\frac{d+1}{N})
\end{aligned}
$$

(e) Similarly, $\mathbf{y}' = [y_1',\ y_2',\ y_3',\ ...,\ y_N']^\top$, and $y' = X\mathbf{w}^* + \epsilon'$, and $\hat{y} = X\mathbf{w}^* + H\epsilon$.
First we have

$$
\begin{aligned}
E_{test}(\mathbf{w}_{lin}) &= Average\ squared\ error\ on\ \mathcal{D}_{test} \\
&= \frac{1}{N}||\hat{y}-y'||^2 \\
&= \frac{1}{N}||(Xw^* + H\epsilon) - (Xw^* + \epsilon')||^2 \\
&= \frac{1}{N}||H\epsilon - \epsilon'||^2 \\
&= \frac{1}{N}(H\epsilon-\epsilon')^\top(H\epsilon-\epsilon') \\
&= \frac{1}{N}(\epsilon^\top H - \epsilon'^\top)(H\epsilon-\epsilon'),\quad Symmetric\ H \\
&= \frac{1}{N}(\epsilon^\top HH\epsilon - 2\epsilon'^\top H\epsilon + \epsilon'^\top\epsilon') \\
&= \frac{1}{N}(\epsilon^\top H\epsilon - 2\epsilon'^\top H\epsilon + \epsilon'^\top\epsilon'),\quad H^K = H
\end{aligned}
$$

And in previous part, we get

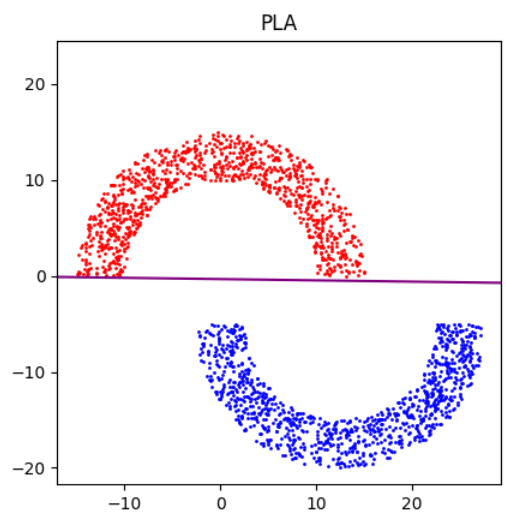
$$\mathbb{E}_{\mathcal{D}}[\epsilon^T \epsilon] = N\sigma^2$$
$$\mathbb{E}_{\mathcal{D}}[\epsilon^T H \epsilon] = (d+1)\sigma^2$$

Thus,

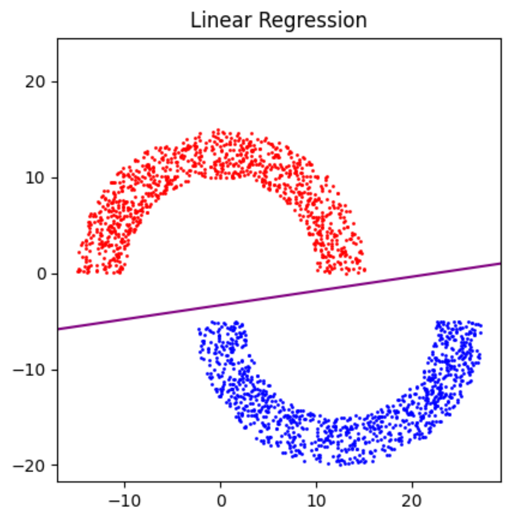
$$\begin{aligned}
\mathbb{E}_{\mathcal{D},\epsilon}[E_{test}(\mathbf{w}_{lin})] &= \mathbb{E}_{\mathcal{D},\epsilon}[\frac{1}{N}(\epsilon^\top H\epsilon - 2\epsilon'^\top H\epsilon + \epsilon'^\top \epsilon')] \\
&= \frac{1}{N}[\mathbb{E}_{\mathcal{D},\epsilon}(\epsilon^\top H\epsilon) - 2\mathbb{E}_{\mathcal{D},\epsilon}(\epsilon'^\top H\epsilon) + \mathbb{E}_{\mathcal{D},\epsilon}(\epsilon'^\top \epsilon')] \\
&= \sigma^2(1 + \frac{d+1}{N}) - \frac{2}{N}\mathbb{E}_{\mathcal{D},\epsilon}(\epsilon'^\top H\epsilon) \\
&= \sigma^2(1 + \frac{d+1}{N}) - \frac{2}{N}(\mathbb{E}_{\mathcal{D},\epsilon}(trace(\epsilon'^\top H\epsilon))) \\
&= \sigma^2(1 + \frac{d+1}{N}) - \frac{2}{N}(\sum_{i=1}^{N} \epsilon'_i H_{ii}\epsilon_i) \\
&= \sigma^2(1 + \frac{d+1}{N})
\end{aligned}$$

2. **Problem 3.1**

   (a) Plotting:



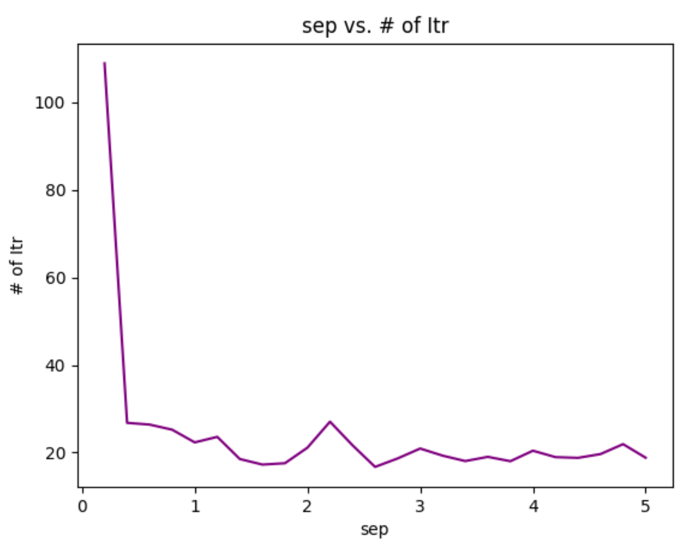
(b) Plotting:



Explanation: Since linear regression is seeking to minimize the squared error between $h(\mathbf{x})$ and y^2, the line will at the "middle" of the two classes, which is different to PLA.

3. **Problem 3.2**

Plotting:



Explanation: From problem 1.3, we have $\rho = min_{1 \le n \le N} y_n(w^{*\top} x_n)$, and $t \le (\frac{R||w^*||}{\rho})^2$. Here R is a constant, so the larger the $\frac{||w^*||}{\rho}$, the smaller the t. Also, let $w = [w_0, w_1, w_2, ..., w_m]^\top$, then $\frac{||w^\top x_n||}{||w||}$ represents the distance from x_n to plane $w^\top x = 0$. Since $\frac{||w^*||}{\rho} \approx \frac{1}{min(dis)}$, the larger the sep, the larger the $distance$, then the smaller the $\frac{||w^*||}{\rho}$, finally the smaller the t.

4. **Problem 3.8**

$$
\begin{aligned}
(h(\mathbf{x}) - y)^2 &= (h(\mathbf{x}) - y - \mathbb{E}[y|\mathbf{x}] + \mathbb{E}[y|\mathbf{x}])^2 \\
&= [(\mathbb{E}[y|\mathbf{x}] - y) + (h(\mathbf{x}) - \mathbb{E}[y|\mathbf{x}])]^2 \\
&= (h(\mathbf{x}) - \mathbb{E}[y|\mathbf{x}])^2 + (\mathbb{E}[y|\mathbf{x}] - y)^2 + 2(h(\mathbf{x}) - \mathbb{E}[y|\mathbf{x}])(\mathbb{E}[y|\mathbf{x}] - y)
\end{aligned}
$$

For $h^*(\mathbf{x}) = \mathbb{E}[y|\mathbf{x}])$, we have

$$
\begin{aligned}
E_{out}(h) &= \mathbb{E}[(h(\mathbf{x}) - y)^2] \\
&= \mathbb{E}[(h(\mathbf{x}) - \mathbb{E}[y|\mathbf{x}])^2] + \mathbb{E}[(\mathbb{E}[y|\mathbf{x}] - y)^2] + 2\mathbb{E}[(h(\mathbf{x}) - \mathbb{E}[y|\mathbf{x}])(\mathbb{E}[y|\mathbf{x}] - y)] \\
&= \mathbb{E}[(h(\mathbf{x}) - h^*(\mathbf{x}))^2] + \mathbb{E}[(h^*(\mathbf{x}) - y)^2] + 2\mathbb{E}[(h(\mathbf{x}) - h^*(\mathbf{x}))(h^*(\mathbf{x}) - y)]
\end{aligned}
$$

Consider $\mathbb{E}[(h(\mathbf{x}) - h^*(\mathbf{x}))(h^*(\mathbf{x}) - y)]$ with $\mathbb{E}[\mathbb{E}[y|\mathbf{x}]] = \mathbb{E}[y]$,

$$
\begin{aligned}
\mathbb{E}[(h(\mathbf{x}) - h^*(\mathbf{x}))(h^*(\mathbf{x}) - y)] &= \mathbb{E}[\mathbb{E}[(h(\mathbf{x}) - h^*(\mathbf{x}))(h^*(\mathbf{x}) - y)|x]] \\
&= \mathbb{E}[((h(\mathbf{x}) - h^*(\mathbf{x}))\mathbb{E}[(h^*(\mathbf{x}) - y)|x]] \\
&= \mathbb{E}[((h(\mathbf{x}) - h^*(\mathbf{x}))(\mathbb{E}[(h^*(\mathbf{x})|x] - \mathbb{E}[y|x])] \\
&= \mathbb{E}[((h(\mathbf{x}) - h^*(\mathbf{x}))((h^*(\mathbf{x}) - (h^*(\mathbf{x}))] \\
&= 0
\end{aligned}
$$

Then

$$
E_{out}(h) = \mathbb{E}[(h(\mathbf{x}) - h^*(\mathbf{x}))^2] + \mathbb{E}[(h^*(\mathbf{x}) - y)^2]
$$

Since $h(\mathbf{x})$ contains all the hypothesis,

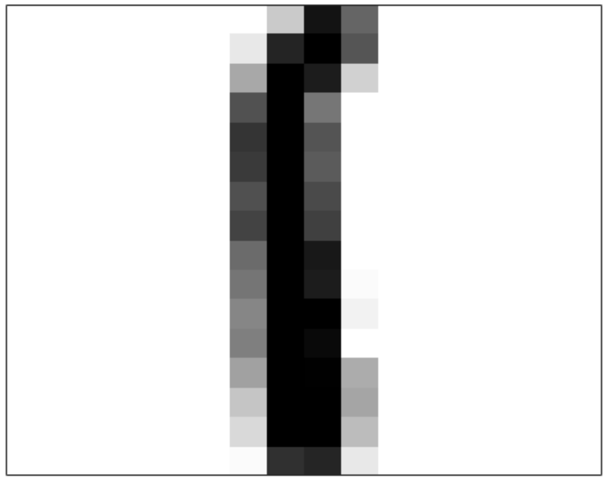
$$
E_{out}(h) \geq \mathbb{E}[(h^*(\mathbf{x}) - y)^2]
$$

And this $h^*(\mathbf{x})$ minimizes $E_{out}(h)$.
For $y = h^*(\mathbf{x}) + \epsilon(\mathbf{x})$, $\epsilon(\mathbf{x}) = y - h^*(\mathbf{x})$, then

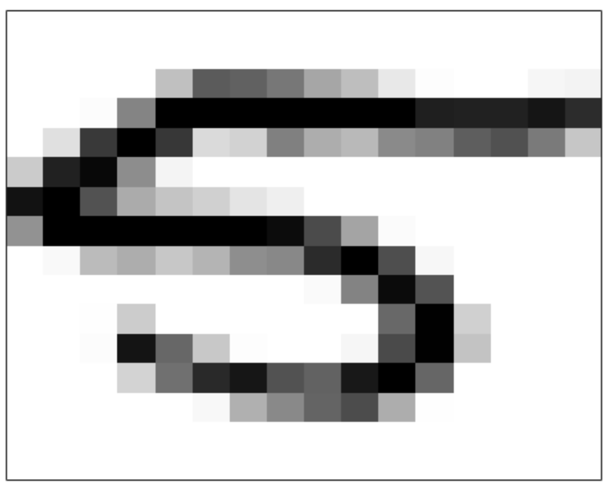
$$
\begin{aligned}
\mathbb{E}[\epsilon(\mathbf{x})|\mathbf{x}] &= \mathbb{E}[y|\mathbf{x}] - \mathbb{E}[h^*(\mathbf{x})|\mathbf{x}] \\
&= h^*(\mathbf{x}) - h^*(\mathbf{x}) \\
&= 0
\end{aligned}
$$

5. **Handwritten Digits Data - Obtaining Features**

(a) Plot 1:



Plot 2:

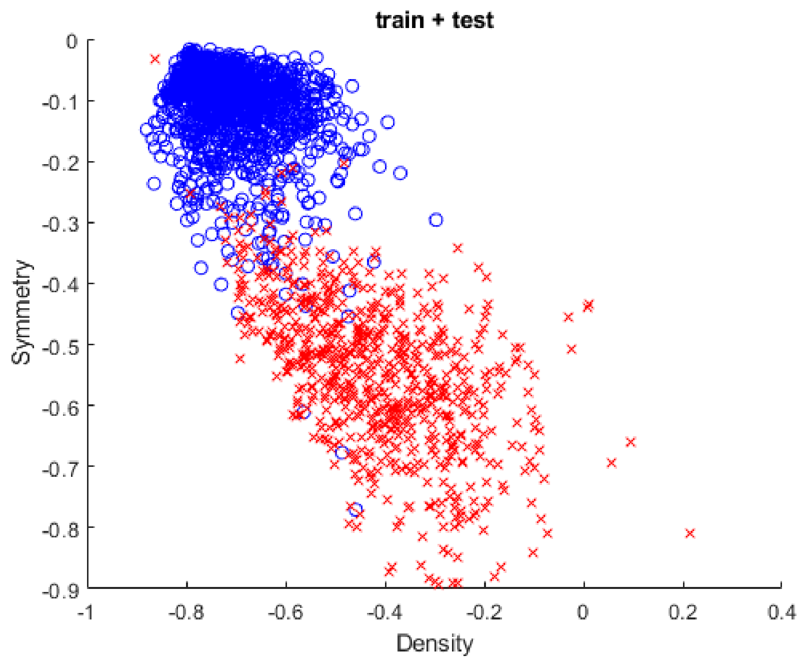


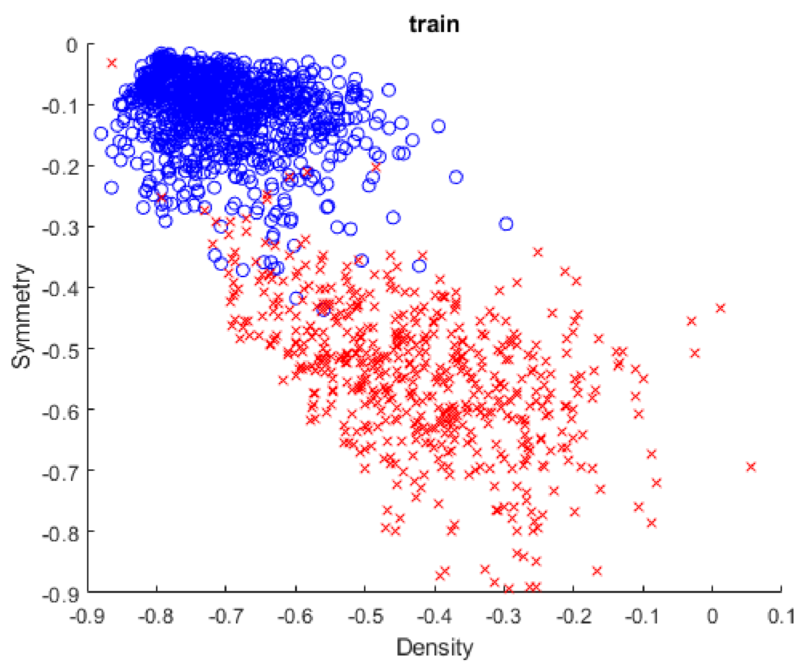
(b) Using the features symmetry and density.

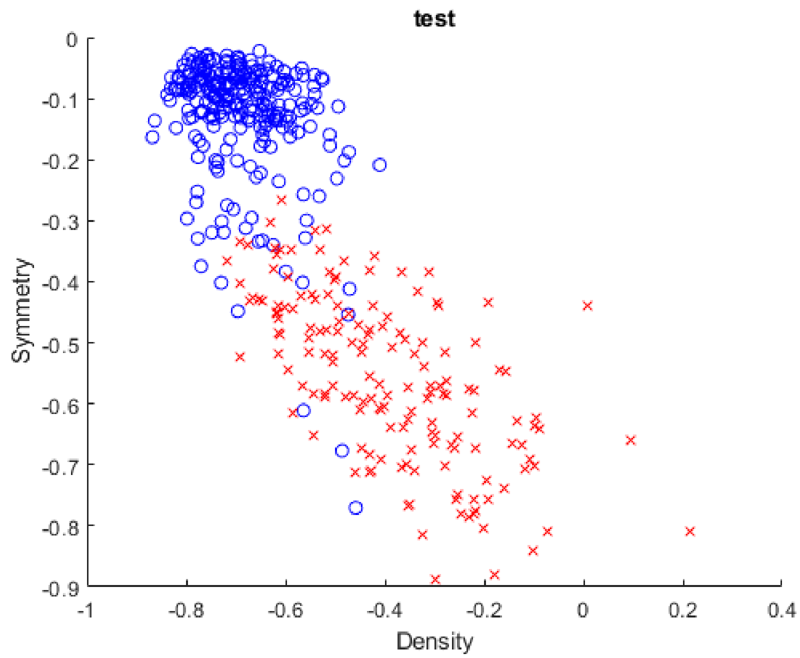
Symmetry $= -\frac{1}{256}\sum_{i=1}^{16}\sum_{j=1}^{16}|X_{ij} - X_{i(17-j)}|$

Density $= \frac{1}{256}\sum_{i=1}^{16}\sum_{j=1}^{16}X_{ij}$

(c) Plotting with Density vs. Symmetry






test
Symmetry
Density
0
-0.1
-0.2
-0.3
-0.4
-0.5
-0.6
-0.7
-0.8
-0.9
-1
-0.8
-0.6
-0.4
-0.2
0
0.2
0.4