CSCI 4100 Assignment 6

Boliang Yang 661541863

1. Exercise 3.4 in LFD

(a) Since $y = {w^*}^T x + \epsilon$, then $y = Xw^* + \epsilon$. It follows $\hat{y} = Hy = H(Xw^* + \epsilon) = X(X^TX)^{-1}X^TXw^* + H\epsilon = XIw^* + H\epsilon = Xw^* + H\epsilon$.

(b) $\hat{y} - y = Xw^* + H\epsilon - (Xw^* + \epsilon) = (H - I)\epsilon$, the matrix is $H - I$.

(c) $E_{in}(w_{lin}) = \frac{1}{N}\sum_{n=1}^{N}(\widehat{y_n} - y_n)^2 = \frac{1}{N}\left||\hat{y} - y|\right|^2 = \frac{1}{N}\left||(H - I)\epsilon|\right|^2 = \frac{1}{N}\left||H\epsilon - \epsilon|\right|^2$

From exercise 3.3(c) and $H^T = (X(X^TX)^{-1}X^T)^T = X((X^TX)^{-1})^TX^T = X(X^TX)^{-1}X^T = H$, we have $H^TH = H^2 = H$. Thus, $E_{in}(w_{lin}) = \frac{1}{N}\left||H\epsilon - \epsilon|\right|^2 = \frac{1}{N}(\epsilon^TH^TH\epsilon - 2\epsilon^TH^T\epsilon + \epsilon^T\epsilon) = \frac{1}{N}(\epsilon^TH\epsilon - 2\epsilon^TH\epsilon + \epsilon^T\epsilon) = \frac{1}{N}(\epsilon^T\epsilon - \epsilon^TH\epsilon)$

(d) $\mathbb{E}_D[E_{in}(w_{lin})] = \mathbb{E}_D\left[\frac{1}{N}\epsilon^T\epsilon\right] - \mathbb{E}_D\left[\frac{1}{N}\epsilon^TH\epsilon\right] = \frac{1}{N}(\mathbb{E}_D[\epsilon^T\epsilon] - \mathbb{E}_D[\epsilon^TH\epsilon])$

Since $\epsilon^TH\epsilon = \sum_{i=1}^{N}\epsilon_i(H\epsilon) = \sum_{i=1}^{N}\sum_{j=1}^{N}\epsilon_i\epsilon_jH_{i,j} = \sum_{i=1}^{N}\epsilon_i^2H_{i,i} + \sum_{i,j\geq1;i\neq j}^{N}\epsilon_i\epsilon_jH_{i,j}$, $trace(H) = d + 1$, $E(\epsilon) = 0, var(\epsilon) = \sigma^2$ and independence of $\epsilon_1, \dots, \epsilon_N$, we have $\mathbb{E}_D[\epsilon^TH\epsilon] = \mathbb{E}_D\left[\sum_{i=1}^{N}\epsilon_i^2H_{i,i}\right] + \mathbb{E}_D\left[\sum_{i,j\geq1,i\neq j}^{N}\epsilon_i\epsilon_jH_{i,ij}\right] = \sigma^2(d + 1) + 0 = \sigma^2(d + 1)$. Along with, $\mathbb{E}_D[\epsilon^T\epsilon] = \sum_{i=1}^{N}\epsilon_i^2 = N\sigma^2$, we have $\mathbb{E}_D[E_{in}(w_{lin})] = \frac{1}{N}(\mathbb{E}_D[\epsilon^T\epsilon] - \mathbb{E}_D[\epsilon^TH\epsilon]) = \frac{1}{N}\left(N\sigma^2 - \sigma^2(d + 1)\right) = \sigma^2(1 - \frac{d+1}{N})$.

(e) Since $\hat{y} = Xw^* + H\epsilon$ and $y_{test} = Xw^* + \epsilon'$, then $E_{test}(w_{lin}) = \frac{1}{N}\left||\hat{y} - y_{test}|\right|^2 = \frac{1}{N}\left||H\epsilon - \epsilon'|\right|^2 = \frac{1}{N}(\epsilon^TH\epsilon - 2\epsilon^TH\epsilon' + \epsilon'^T\epsilon')$. And it follows that $\mathbb{E}_{D,\epsilon'}[E_{test}(w_{lin})] = \frac{1}{N}(\mathbb{E}_{D,\epsilon'}[\epsilon^TH\epsilon] - 2\mathbb{E}_{D,\epsilon'}[\epsilon^TH\epsilon'] + \mathbb{E}_{D,\epsilon'}[\epsilon'^T\epsilon'])$. And by the same method as in part(d), we have $\mathbb{E}_{D,\epsilon'}[\epsilon^TH\epsilon] = \mathbb{E}_{D,\epsilon'}[\sum_{i=1}^{N}\epsilon_i^2H_{i,i}] + \mathbb{E}_{D,\epsilon'}[\sum_{i,j\geq1,i\neq j}^{N}\epsilon_i\epsilon_jH_{i,j}] = \sigma^2(d + 1)$,

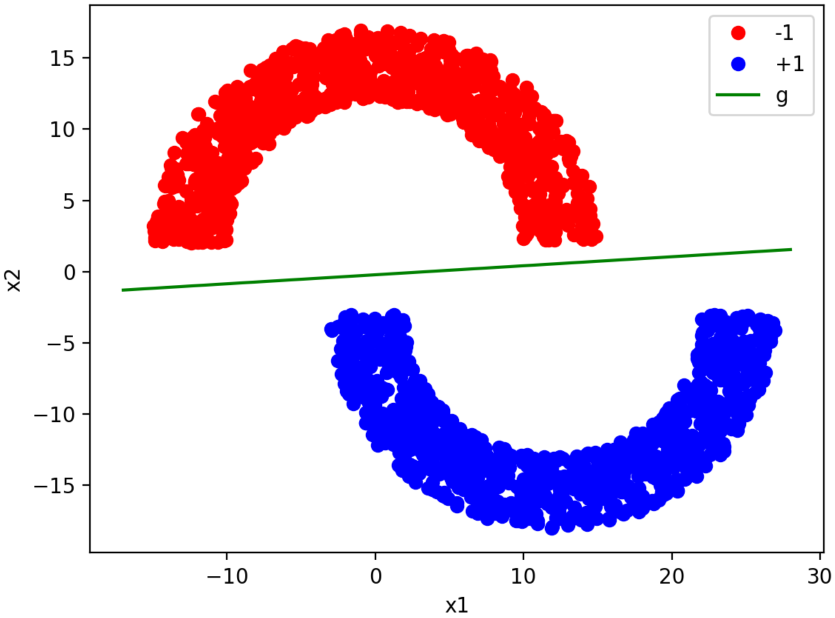
$\mathbb{E}_{D,\epsilon'}[\epsilon^TH\epsilon'] = \mathbb{E}_{D,\epsilon'}\left[\sum_{i=1}^{N}\sum_{j=1}^{N}\epsilon_i'\epsilon_jH_{i,j}\right] = \sum_{i=1}^{N}\sum_{j=1}^{N}\mathbb{E}_{D,\epsilon'}[\epsilon_i']\mathbb{E}_{D,\epsilon'}[\epsilon_j]H_{i,j} = 0$,

$\mathbb{E}_{D,\epsilon'}[\epsilon'^T\epsilon'] = \sum_{i=1}^{N}{\epsilon_i'}^2 = N\sigma^2$,.

Thus, we have $\mathbb{E}_{D,\epsilon'}[E_{test}(w_{lin})] = \frac{1}{N}\left(\mathbb{E}_{D,\epsilon'}[\epsilon^TH\epsilon] - 2\mathbb{E}_{D,\epsilon'}[\epsilon^TH\epsilon'] + \mathbb{E}_{D,\epsilon'}[\epsilon'^T\epsilon']\right) = \frac{1}{N}(\sigma^2(d + 1) - 0 + N\sigma^2) = \sigma^2(1 + \frac{d+1}{N})$.

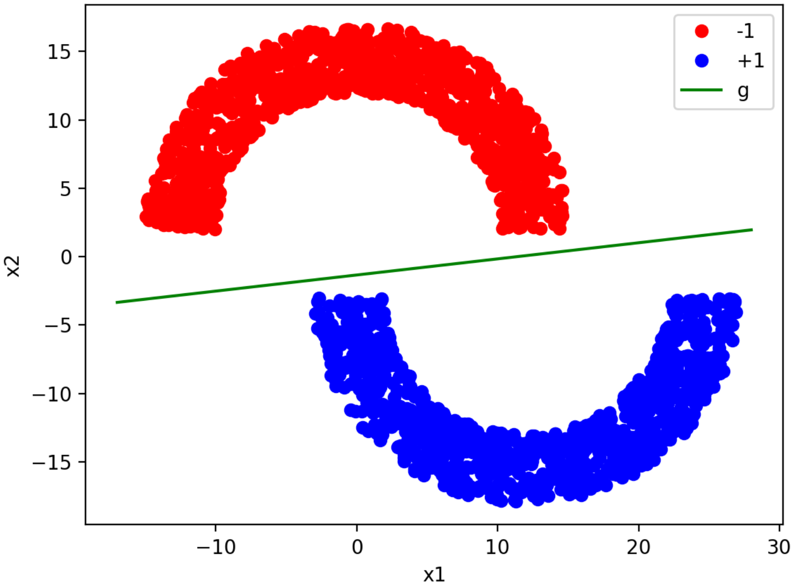
2. Problem 3.1 in LFD

(a)



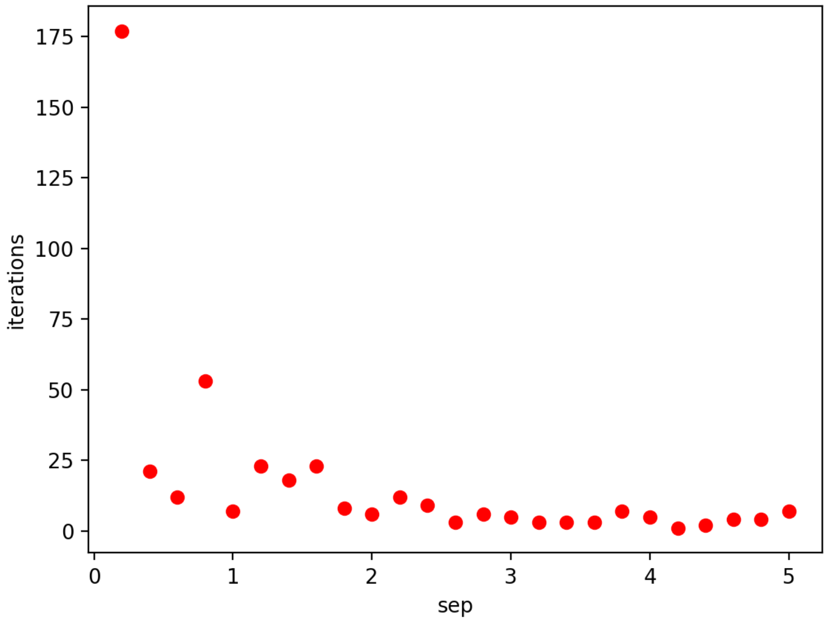
It takes 13 iterations. The final hypothesis is $g(x) = 0.063235x - 0.219676$

(b)



The final hypothesis is $g(x) = 0.117802x - 1.344676$. It runs more quickly when using the linear regression than PLA, because of the different method used: in linear regression just matrix computation but in PLA we need iteration.

3. Problem 3.2 in LFD



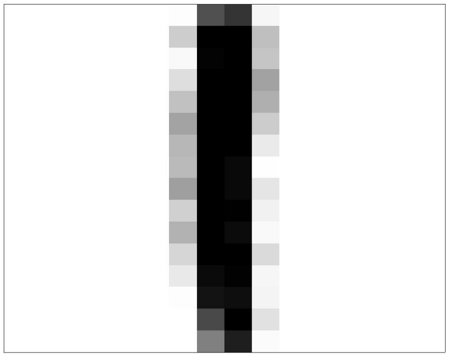
Explanation: as sep increases, the number of iteration decreases significantly and remains nearly constant. Because as sep increases, there are more possible hypotheses that fit all the data points and it is easier to separate the data. As we shown in problem 1.3, the bound is $t \leq \frac{R^2||w^*||^2}{\rho^2}$. Since R and ρ increase as sep increases, the resulting bound would approximately remain the same.

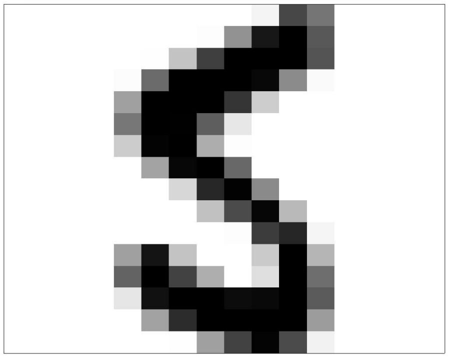
4. Problem 3.8 in LFD

From problem, we have $E_{out}(h) = \mathbb{E}[(h(x) - y)^2] = \mathbb{E}[h(x)^2 - 2h(x)y + y^2] = h(x)^2 - 2h(x)E[y|x] + E[y|x]^2$. We can now take the derivative and set to 0. $\frac{dE_{out(h)}}{dh(x)} = 2h(x) - 2E[y|x] = 0$. Then we have, $h^*(x) = E[y|x]$. Since $y = h^*(x) + \epsilon(x)$, then $E[y|x] = E[h^*(x) + \epsilon(x)] = E[h^*(x)] + E[\epsilon(x)] = h^*(x) + E[\epsilon(x)]$. Since $h^*(x) = E[y|x]$, then $E[\epsilon(x)] = 0$.

5. Handwritten Digits Data – Obtaining Features

(a) A plot of two of the digits images:

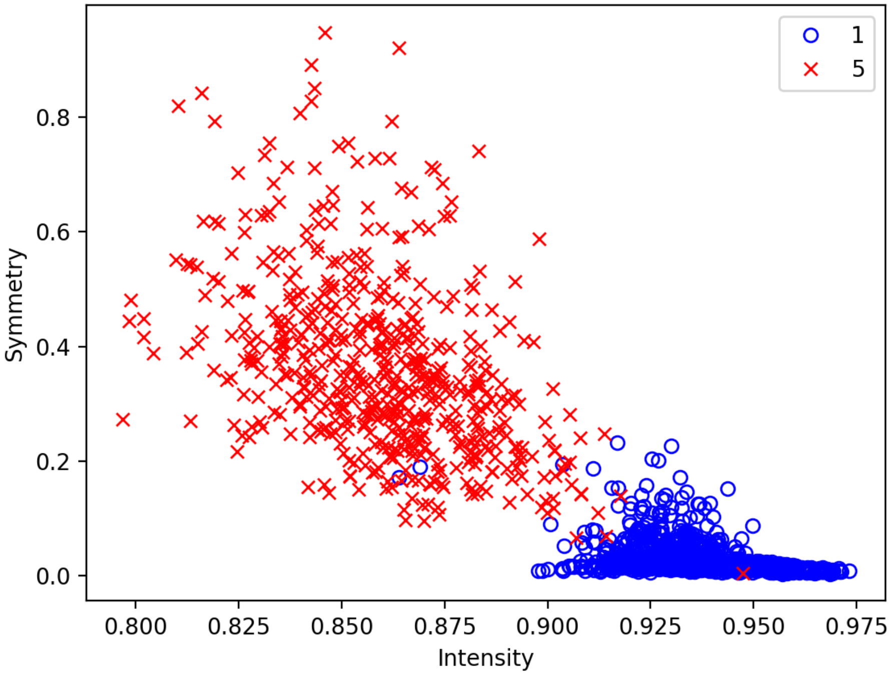




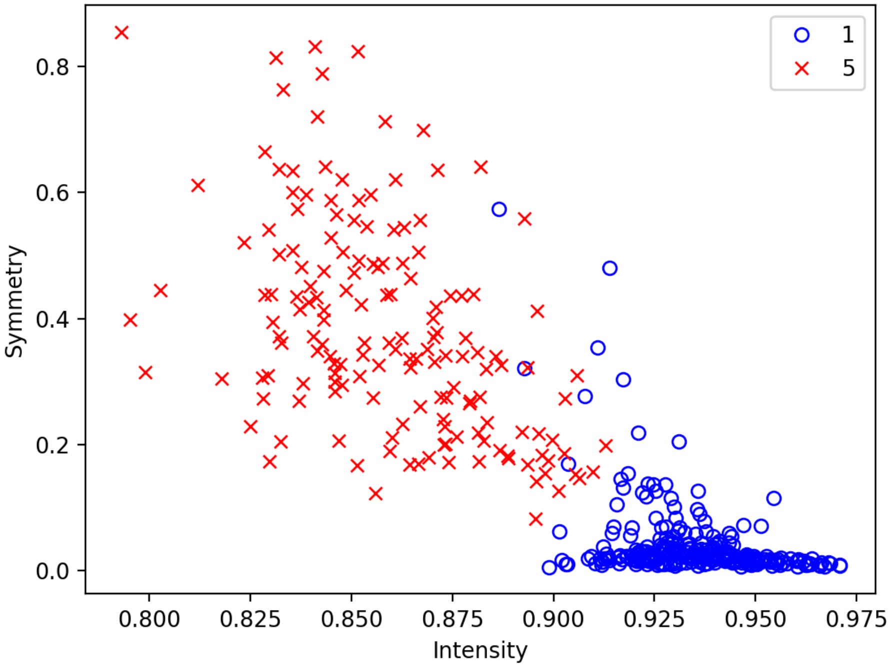
(b) Define the pixel with index i as $p[i]$, and if it is white then $p[i] = -1$, else if it is black then $p[i] = +1$. Then the average intensity is $\frac{1}{256}\sum_{i=0}^{255}|p[i]|$. Now we define symmetry as symmetry about x-axis and y-axis, then we calculate symmetry as: $\frac{1}{128}\sum_{i=0}^{7}\sum_{j=0}^{15}|p[16j+i] - p[16(j+1)-(i+1)]| \times \frac{1}{128}\sum_{i=0}^{16}\sum_{j=0}^{7}|p[16j+i] - p[16(15-j)+i]|$.

(c) 2-D scatter plot of the two features:

For training data:



For test data:



In both data sets, almost all of the 5 and 1 points are separated. However, there are a few outliers in both cases.