Assignment 6

Name: Mohaiminul Al Nahian
RIN: 662026703
Course no.: CSCI 6100

Exercise 3.4

a) Given $y = Xw^* + \epsilon$.
We know that $\hat{y} = Hy$. where $H = X(X^TX)^{-1}X^T$. So, We have:
$\hat{y} = Hy = X(X^TX)^{-1}X^T(Xw^* + \epsilon) = X(X^TX)^{-1}X^TXw^* + H\epsilon = Xw^* + H\epsilon$ (Showed)

b) The in sample error $\hat{y} - y$ can be expressed by:

$$\begin{aligned}\hat{y} - y &= (Xw^* + H\epsilon) - (Xw^* + \epsilon)\\ &= (H - I_N)\epsilon\end{aligned}$$

So, in sample error can be expressed by the matrix $(H - I_N)$ times ϵ where $I_N = N \times N$ dimensional identity matrix.

c)

$$\begin{aligned}E_{in}(w_{lin}) &= \frac{1}{N}||\hat{y} - y||^2\\ &= \frac{1}{N}||(H - I_N)\epsilon||^2\\ &= \frac{1}{N}\epsilon^T(H - I_N)^T(H - I_N)\epsilon\end{aligned}$$

Here, $H - I_N$ is symmetric, so $(H - I_N)^T(H - I_N) = (H - I_N)^2 = (I_N - H)^2 = I_N - H$ [From exercise 3.3(c)]. So,

$$E_{in}(w_{lin}) = \frac{1}{N}\epsilon^T(I_N - H)\epsilon$$

d)The expected in-sample error of linear regression is given by,

$$
\begin{aligned}
E_{\mathcal{D}}[E_{in}(w_{lin})] &= E_{\mathcal{D}}\left[\frac{1}{N}\epsilon^T\epsilon - \frac{1}{N}\epsilon^T H\epsilon\right] \\
&= \frac{1}{N}E_{\mathcal{D}}\left[\epsilon^T\epsilon\right] - \frac{1}{N}E_{\mathcal{D}}\left[\epsilon^T H\epsilon\right] \\
&= \frac{1}{N}(E_{\mathcal{D}}\left[\epsilon^T\epsilon\right] - E_{\mathcal{D}}\left[\epsilon^T H\epsilon\right])
\end{aligned}
$$

Here, we can say $E_{\mathcal{D}}\left[\epsilon^T\epsilon\right] = N\sigma^2$, since it is a zero mean noise term with σ^2
And, $E_{\mathcal{D}}\left[\epsilon^T H\epsilon\right]$ is a diagonal matrix, where $E_{\mathcal{D}}\left[\epsilon^T H\epsilon\right] = trace(H) * \sigma^2 = (d+1) * \sigma^2$ [From exercise 3.4(d), $trace(H) = d+1$]
So,

$$
\begin{aligned}
E_{\mathcal{D}}[E_{in}(w_{lin})] &= \frac{1}{N}(N * \sigma^2 - (d+1)\sigma^2) \\
&= \sigma^2(1 - \frac{d+1}{N})(Showed)
\end{aligned}
$$

e) Expected out-sample error

$$
\begin{aligned}
E_{\mathcal{D},\epsilon'}[E_{test}(w_{lin})] &= E_{\mathcal{D},\epsilon'}\left[\ \frac{1}{N}||Hy - y'||^2\right] \\
&= \frac{1}{N}E_{\mathcal{D},\epsilon'}\left[\ ||X(X^TX)^{-1}X^T(Xw^* + \epsilon) - (Xw^* + \epsilon')||^2\right] \\
&= \frac{1}{N}E_{\mathcal{D},\epsilon'}\left[\ ||Xw^* + H\epsilon - (Xw^* + \epsilon')||^2\right] \\
&= \frac{1}{N}E_{\mathcal{D},\epsilon'}\left[\ ||H\epsilon - \epsilon'||^2\right] \\
&= \frac{1}{N}E_{\mathcal{D},\epsilon'}\left[\ ||\epsilon^T H^T H\epsilon - 2\epsilon^T H^T\epsilon' + \epsilon'^T\epsilon'||\right] \\
&= \frac{1}{N}[\sigma^2(d+1) + N\sigma^2] \\
&= \sigma^2(1 + \frac{d+1}{N})
\end{aligned}
$$

Problem 3.1

In this Problem, the followwing has been selected: separation=5, radius=10 and thickness=5. The first semi-circle's center is chosen $(0, separation/2)$ and the second center to be $(radius + thickness/2, -separation/2)$ to satisfy center of the top semicircle align with middle of edge of bottom semicircle

a) PLA is run starting with w = [0 0 0]. The final hypothesis after 4 iterations give $W_{PLA} = [4.0, -1.50697623, 24.52773534]$

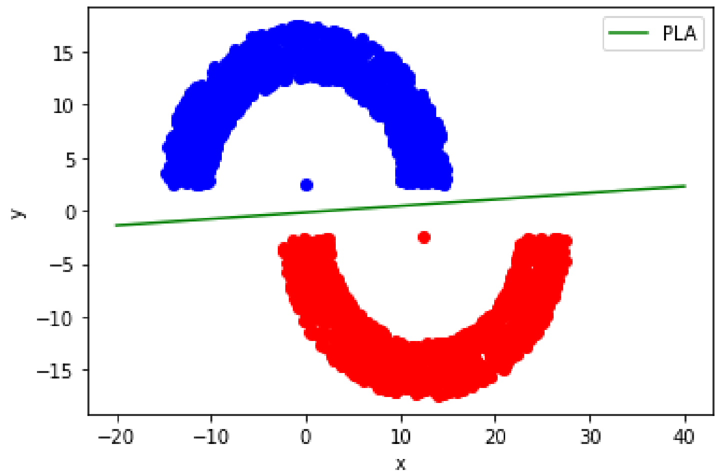


Figure 1: PLA final hypothesis result

b) The Linear regression obtains W_{lin} to be [0.07301111, -0.01052355, 0.07854483]

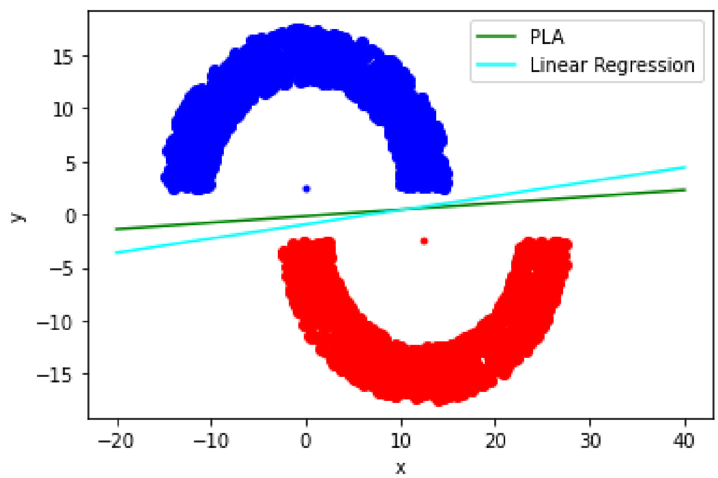


Figure 2: PLA and Linear Regression

Both PLA and linear regression separates the two types perfectly although the results are different.

Problem 3.2

separation is varied in the range $\{0.2, 0.4, ..., 5\}$ and the PLA is run get the number of iterations to converge. For each different separation, PLA is run on randomly generated data to get a clear picture about how number of iteration change if only the separation is changed.

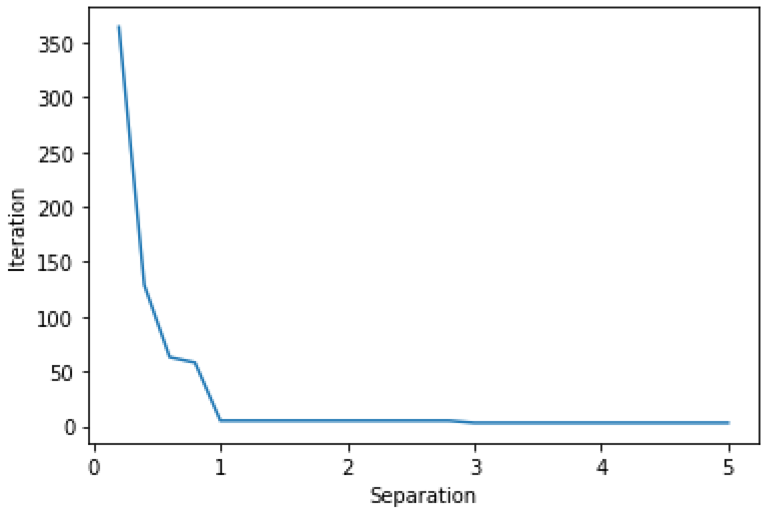


Figure 3: Iterations vs Separation

As *separation* increases, the number of iterations decreases. According to problem 1.3 PLA converges quicker than the bound $\frac{R^2||w^*||^2}{\rho^2}$. Here, ρ increases with the increase in separation and and PLA should converge inversely propotional to square of ρ. Although the experiment does not clearly show this relationship, but the resulting trend shows that increase in separation results in much faster convergence.

Problem 3.8

$$\begin{aligned}
E_{out} = E[(h(x)-y)^2] &= \int_x \int_y (h(x)-y)^2 p(x)p(y|x)dydx \\
&= \int_x \int_y (h(x)^2 - 2h(x) + y^2)p(x)p(y|x)dydx \\
&= \int_x h(x)^2 p(x) \int_y p(y|x)dydx - 2\int_x h(x)p(x) \int_y yp(y|x)dydx + \int_x p(x) \int_y y^2 p(y|x)dyd \\
&= \int_x [h(x)^2 - 2h(x)E[y|x] + (E[y|x])^2 + variance[y|x]]p(x)dx \\
&= \int_x (h(x) - E[y|x])^2 p(x)dx + \int_x variance[y|x]p(x)dx \\
&= \int_x (h(x) - E[y|x])^2 p(x)dx + E[variance[y|x]]
\end{aligned}$$

Here, $E[variance[y|x]]$ is constant. So, to minimize E_{out}, we need to minimize, $\int_x (h(x) - E[y|x])^2 p(x)dx$.
This the minimum value is achieved if $h(x) - E[y|x] = 0$ or, $h^*(x) = E[y|x]$ (Showed)

Now,

$$
\begin{aligned}
y(x) &= h^*(x) + \epsilon(x) \\
or, E[y(x)] &= E[h^*(x) + \epsilon(x)] = E[h^*(x)] + E[\epsilon(x)]
\end{aligned}
$$

Now, $E[y(x)] = E[y|x]$ and $E[h^*(x)] = h^*(x) = E[y|x]$.
So,

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

So,, $E[\epsilon(x)] = 0$ (Showed)

Problem 3.6

a)For linearly separable data, the sign of y_n and $w^T x_n$ must be same. So, $y_n * w^T x_n \geq a$, where $a \geq 0$. So, we may scale w with a number $\frac{1}{a}$ so that new W=$\frac{w}{a}$ and we get $y_n * (W^T x_n) \geq 1$ for all n.

b) We know that

$$
\begin{aligned}
y(i)(w^T x(i)) &\geq 1 \\
=> (y(i)x(i))^T w &\geq 1 \\
=> -(y(i)x(i))^T w &\leq -1
\end{aligned}
$$

So, the linear programm can be written as:

$$
A = -\left[y_1 x_1^T, y_2 x_2^T, ..., y_n x_n^T\right]_{n \times d}^T, \; z = [w1, w2, ..., w_d]_{d\times 1}^T, \; b = [-1, -1, ..., -1]_{n\times 1}^T,
$$
$$
c = [0, 0, ..., 0]_{d\times 1}^T. \text{ Here } y_i x_i = y_i [x_{i1}, x_{i2}, ..., x_{id}]^T \text{ a d dimensional vector.}
$$

For $\min c^T z$ subject to $Az \leq b$

c) For data that are not linearly searable, we may write

$$
\begin{aligned}
y_n(w^T x_n) &\geq 1 - \xi_n \\
\xi_n + (y_n x_n^T)w &\geq 1 \\
-\xi_n + (-y_n x_n^T)w &\leq -1
\end{aligned}
$$

Now, construct the linear program as:

$$A = \begin{bmatrix} -1 & 0 & \cdots & 0 & -y_1x_1^T \\ 0 & -1 & \cdots & 0 & -y_2x_2^T \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & -1 & -y_nx_n^T \\ -1 & 0 & \cdots & 0 & [0]_d \\ 0 & -1 & \cdots & 0 & [0]_d \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & -1 & [0]_d \end{bmatrix}_{2n\times(n+d)} \quad z = \begin{bmatrix} \xi_1 \\ \xi_2 \\ \vdots \\ \xi_n \\ w1 \\ w2 \\ \vdots \\ wd \end{bmatrix}_{n+d} \quad b = \begin{bmatrix} -1 \\ -1 \\ \vdots \\ -1 \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}_{2n} \quad c = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}_{n+d}$$

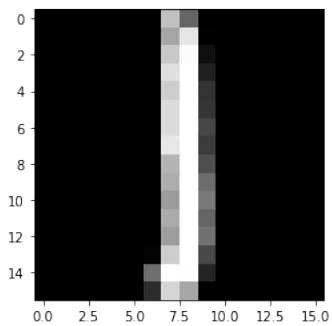
d) We have for non-separable data

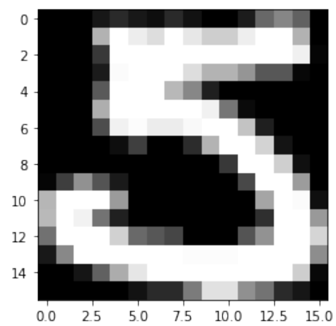
$$\begin{aligned} y_n(w^Tx_n) &\geq 1 - \xi_n \\ or, \xi_n &\geq 1 - y_n(w^Tx_n) \end{aligned}$$

Note that, subject to $\xi \geq 0$, we have ξ_n should be the bound, $\xi_n = max(0, 1 - y_n(w^Tx_n))$
So we can minimize: $\min_w \sum_{n=1}^{N} \max(0, 1 - y_nw^Tx_n)$ Which is the same problem as 3.5

Problem 6: Handwritten Digits- Obtaining Features

a) Two of the digit images:



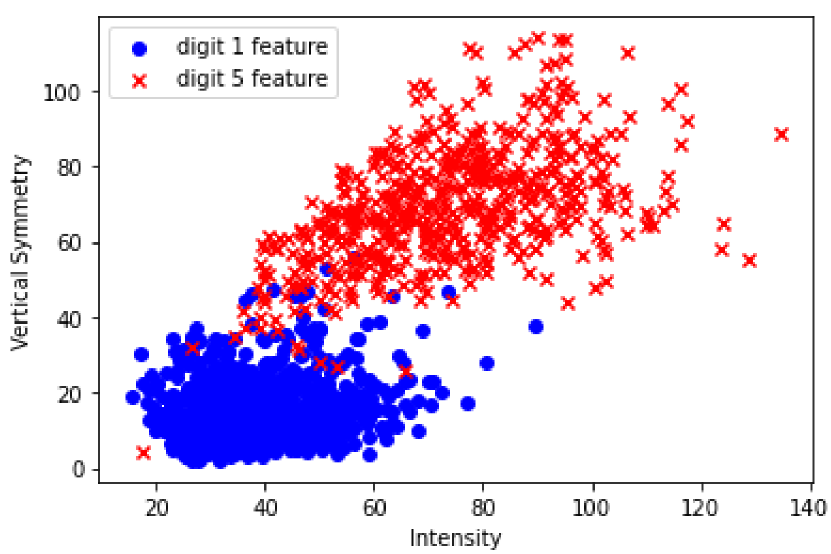


b)The two chosen features are 'intensity' and 'vertical symmetry'. The pixel values between -1 and 1 has been mapped into the range [0,255]. Let pix(i,j) denotes the pixel values of the pixel (i,j), where i=[0,15] and j=[0,15]. Then we have, for N=256 pixels of an image,

$$(i) Intensity = \frac{1}{N}\sum_{i=0}^{15}\sum_{j=0}^{15} pix(i,j)$$

$$(ii) VerticalSymmetry = \frac{1}{N}\sum_{i=0}^{15}\sum_{j=0}^{15} abs[pix(15-i,j) - pix(i,j)]$$

c) The Intensity and Vertical Symmetry has been calculated for all training examples of Digit1 and Digit5. The resulting plot of Intensity vs Vertical Symmetry is as follows:



The Digit5 displays significantly higher Vertical Symmetry and also the Intensity is higher than Digit1 in most of the cases.