CSCI 6100 Machine Learning From Data
Fall 2018

HOMEWORK 6

Daniel Southwick
661542908
southd@rpi.edu

Exercise 3.4

a) The in sample error is defined as:

$$E_{in}(w) = \frac{1}{N}||Xw - y||^2$$

To minimize the sample error, we set the gradient of the sample error to 0:

$$\nabla E_{in}(w) = \frac{2}{N}(X^T X w - X^T y) = 0$$
$$w = (X^T X)^{-1} X^T y$$

$$\begin{aligned}\hat{y} &= Xw = X(X^T X)^{-1} X^T y \\ &= X(X^T X)^{-1} X^T (Xw^* + \epsilon) \\ &= Xw^* + X(X^T X)^{-1} X^T \epsilon \\ &= Xw^* + H\epsilon\end{aligned}$$

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

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

Where I_N denotes an $N \times N$identity matrix.

c)

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

We know that $(H) - I_N$ is symmetric and $(H - I_N)^2 = I_N - H$ by 3.3(c), thus:

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

d)

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

Since ϵ is a noise term with zero means and σ^2, so $\mathbb{E}_{\mathcal{D}}\left[\epsilon^T\epsilon\right] = N\sigma^2$, $\mathbb{E}_{\mathcal{D}}\left[\epsilon^T H\epsilon\right]$ is a diagonal matrix, and $\mathbb{E}_{\mathcal{D}}\left[\epsilon^T H\epsilon\right] = trace(H) * \sigma^2$ and

$$\begin{aligned}trace(H) &= trace(X(X^TX)^-1X^T) = trace((X^TX)^-1X^TX) \\ &= trace(I_{d+1}) = d+1\end{aligned}$$

Where I_{d+1} is a $d+1 \times \dot{+}1$ dimension identity matrix. So

$$\begin{aligned}\mathbb{E}_{\mathcal{D}}[E_{in}(w)] &= \frac{1}{N}(N * \sigma^2 - (d+1)\sigma^2) \\ &= \sigma^2(1 - \frac{d+1}{N})\end{aligned}$$

e)

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

Problem 3.1

In this Problem, we took the first semi-circle's coordinate center to be $(0, 2.5)$ and the second center to be $(10, -2.5)$ to satisfy $rad = 10$, $thk = 5$ and $sep = 5$.

a) We first run the PLA starting from w = 0 until it converges, the final hypothesis came out to be $y = -0.001172x + 0.02945$

b) Then, we use the linear regression to obtain by $w = (X^T X)^{-1} X^T Y = [-0.0416; 0.0083; -0.0790]$ and the hypothesis came out to be $-0.0416 + 0.0083x - 0.0790y = 0$

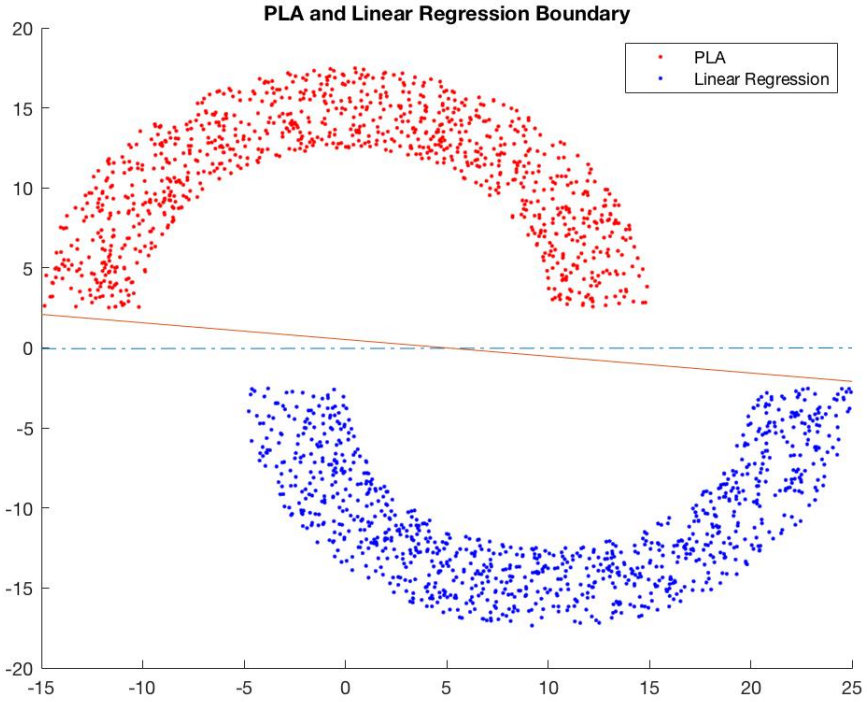

Figure 1: PLA and Linear Regression final hypothesis result

Both PLA and linear regression hypothesis obtained in part a) and b) found solution to separate two regions. But the solutions generated from the different hypothesis are different. The result obtained by linear regression has the smallest E_{in} comparing to PLA.

Problem 3.2

We varied *sep* in the range $0.2, 0.4, ..., 5$ in the program from Problem 3.1 and run the PLA 25 times to record the number of iterations PLA takes to converge. Again, points are randomly sampled each time.

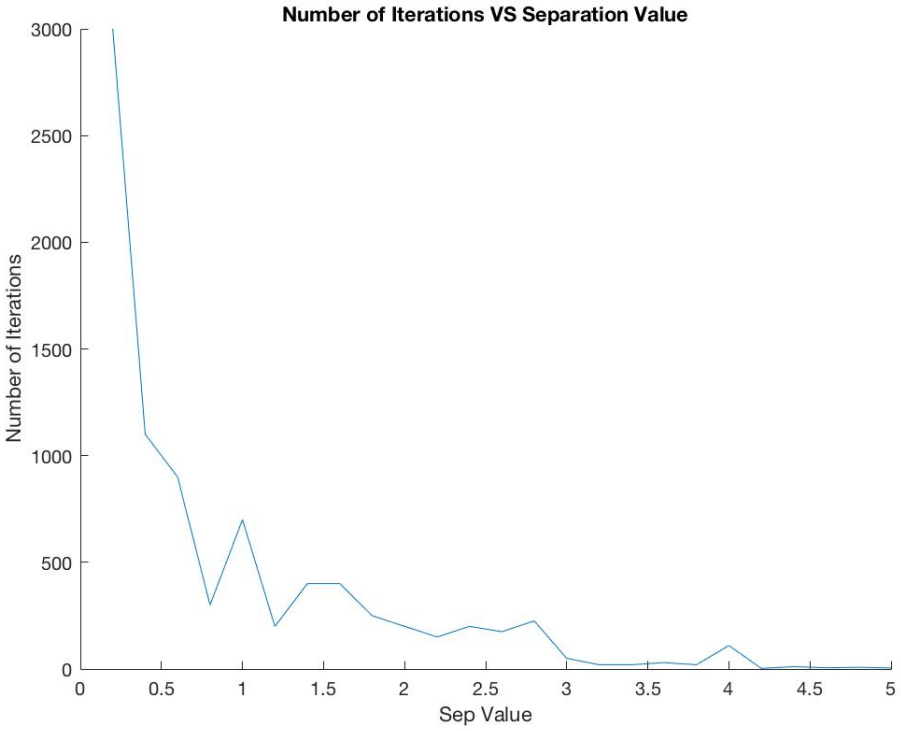

Figure 2: Number of iterations PLA vs varies Sep Value

As *sep* increases, the number of iterations decreases, we can interpret this intuitively, since two regions will be further away from each other, so PLA is much easier to converge. If one line can separate two regions for the smallest *sep*, it can also separate all regions with a larger parameter *sep*. So R and $||w^*||$ are both constant, and ρ increases with the increase of *sep*, the upper bound of iteration is $\frac{R^2||w^*||^2}{\rho^2}$, thus the number of iterations decrease as *sep* increases.

Problem 3.8

We know the out of sample error is: $E_{out}(h) = \mathbb{E}[(h(x) - y)^2]$, to minimize $E_{out}(h)$, we can set it's gradient to 0, by $\nabla E_{out}(h^*) = \mathbb{E}[h^*(x) - y] = 0$ and for any fixed x, $h^*(x)$ can be treated as a fixed value, thus $h^*(x) = \mathbb{E}[y|x]$ by:

$$\begin{aligned}
\mathbb{E}[\mathbb{E}[X\big|Y]] &= \sum_y \mathbb{E}[X\big|Y = y] \times P(Y = y) \\
&= \sum_y \left(\sum_x x \times P\left(X = x\big|Y = y\right) \right) \times P(Y = y) \\
&= \sum_y \sum_x x \times P\left(X = x\big|Y = y\right) \times P(Y = y) \\
&= \sum_y \sum_x x \times P\left(Y = y\big|X = x\right) \times P(Y = y) \\
&= \sum_x x \times P(X = x) \times \left(\sum_y P\left(Y = y\big|X = x\right) \right) \\
&= \sum_x x \times P(X = x) \\
&= \mathbb{E}[X]
\end{aligned}$$

Thus :

$$\begin{aligned}
\nabla E_{out}(h^*) &= \mathbb{E}[h^*(x) - y] = 0 \\
\mathbb{E}[y|x] &= \mathbb{E}[h^*(x) + \epsilon(x)|x] \\
&= \mathbb{E}[h^*(x)|x] + \mathbb{E}[\epsilon(x)|x] \\
\text{So } \mathbb{E}[\epsilon(x)|x] &= 0
\end{aligned}$$

Thus $\Rightarrow E_{out}$ is minimized at $h^*(x)$, the expected value $\epsilon(x)$ is zero.

Problem 3.6

a)For linearly separable data, there must exist some w that satisfy (separate the data correctly):

$$\begin{aligned}
&\text{i) } y(i) = +1, \text{ and } h(x(i)) = sign(w^T x(i)) = +1 \\
&\text{ii)} y(i) = -1, \text{ and } h(x(i)) = sign(w^T x(i)) = -1
\end{aligned}$$

Either cases, $y(i)(w^T x(i)) > 0$. Then we must be able to find an ϵ such that $y(i)(w^T x(i)) \geq \epsilon$ for all data within the dataset. Now, we let $\hat{w} = \dfrac{w}{\epsilon}$, then:

$$y(i)(w^T x(i)) = y(i)(\frac{w^T x(i)}{\epsilon}) \geq \frac{\epsilon}{\epsilon} = 1$$

Thus for some w, ie. $\hat{w} = \frac{w}{\epsilon}$, $y(i)(w^T x(i)) \geq 1$ for all data.

b) Now we know that $y(i)(w^T x(i)) \geq 1$, then $(y(i)x(i))^T w \geq 1$ and $(-y(i)x(i))^T w \leq -1$. We can construct the linear program as:

$$A = -\left[y_1 * x_1^T, y_2 * x_2^T, ..., y_n * x_n^T\right]_{n+1}^T, z = w,\ b = [-1, -1, ..., -1]_{n+1}^T$$

$$\text{For } \min_z c^T z \text{ subject to } Az \leq b$$

So if the given data set is linearly separable, then the LP problem is feasible, then the obtained w can separate the dataset.

c) We can derive

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

Now, combining $-\xi_n \leq 0$, we can construct the linear program as:

$$A = \begin{bmatrix} -y_1 * x_1^T & -1 & 0 & \cdots & 0 \\ -y_2 * x_2^T & 0 & -1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ -y_n * x_n^T & 0 & 0 & \cdots & -1 \\ 0 & -1 & 0 & \cdots & 0 \\ 0 & 0 & -1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & -1 \end{bmatrix}_{(n+1)\times 2n} \quad z = \begin{bmatrix} w \\ \xi_1 \\ \xi_2 \\ \vdots \\ \xi_n \end{bmatrix}_{n+1} \quad b = \begin{bmatrix} -1 \\ -1 \\ \vdots \\ -1 \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}_{2n} \quad c = \begin{bmatrix} 0 \\ 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}_{n+1}$$

d) Now we derive from the result from part b) again:
Note that ξ_n should be the lower bound, $\xi_n = max(0, 1 - y_n(w^T x_n))$
Then:

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

$$\text{So we can minimize: } \min_w \sum_{i=1}^{N} \max(0, 1 - y_n w^T x_n)$$

Which is the same problem as 3.5

Handwritten Digits Data - Obtaining Features

a)We've sorted out the number images and chose to display a number 1 and a number 5 here:

(a) Digit 1

(b) Digit 5

Figure 3: 1 and 5

b)Here we choose the features of intensity and symmetry, where symmetry in this case means the whether the image is vertical symmetry. Let $f(i, j)$ denotes the grayscale values from -1 to 1 for pixel (i, j) as given. And i, j ranges from 1 to 16. Then the intensity is defined as:

$$I_{avg} = \frac{1}{256} \sum_{i=1}^{16} \sum_{i=1}^{16} f(i, j)$$

Then the symmetric is defined as

$$I_{sym} = \frac{1}{256} \times \frac{1}{256} \sum_{i=1}^{16} \sum_{i=1}^{16} |f(i, j) - f(17 - i, j)|$$

c) We calculated I_{avg} and I_{sym} for all the 1 and 5 training data, the result is as follows, we can see that digit 5 has higher intensity and symmetrisity than digit 1.

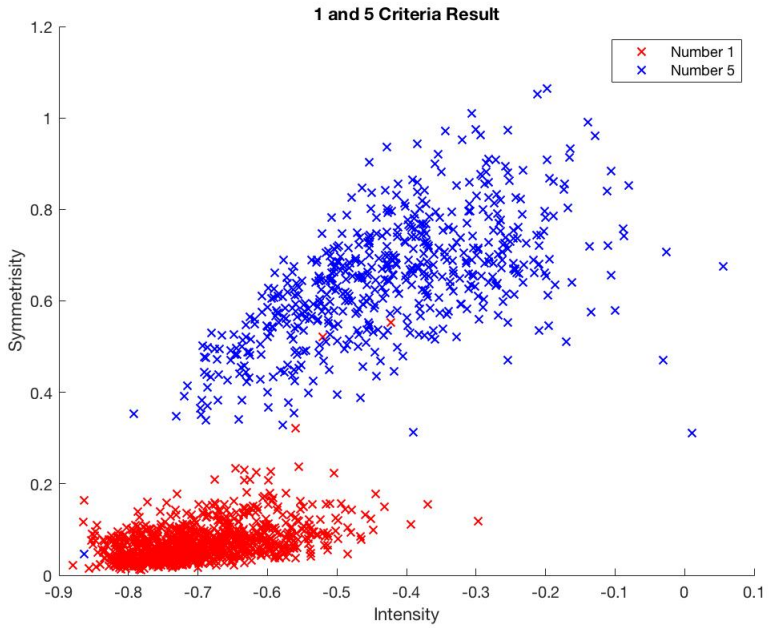

Figure 4: Intensity and Symmetrisity Result for digit 1 and 5