Homework 6

Karan Sarkar
sarkak2@rpi.edu

October 16, 2019

Exercise 3.4.

(a) Note that $\hat{\mathbf{y}} = \mathbf{H}\mathbf{y}$. Therefore, $\hat{\mathbf{y}} = \mathbf{H}(\mathbf{X}\mathbf{w}^* + \epsilon) = \mathbf{H}\mathbf{X}\mathbf{w}^* + \mathbf{H}\epsilon = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}\mathbf{w}^* + \mathbf{H}\epsilon$. This simplifies to $\mathbf{X}\mathbf{w}^* + \mathbf{H}\epsilon$ as required.

(b) Note that $\mathbf{y} = \mathbf{X}\mathbf{w}^* + \epsilon$. From (a), we get that $\hat{\mathbf{y}} - \mathbf{y} = \mathbf{X}\mathbf{w}^* + \mathbf{H}\epsilon - (\mathbf{X}\mathbf{w}^* + \epsilon) = \mathbf{H}\epsilon - \epsilon = (\mathbf{H} - \mathbf{I})\epsilon$. Thus the matrix is $\mathbf{H} - \mathbf{I}$.

(c) Note that the in-sample error mean-square average of the residual vector. Moreover note that $\mathbf{H} - \mathbf{I}$ is symmetric.

$$\begin{aligned}
E_{\text{in}} &= \frac{1}{N}((\mathbf{H} - \mathbf{I})\epsilon)^T(\mathbf{H} - \mathbf{I})\epsilon \\
&= \frac{1}{N}\epsilon^T(\mathbf{H} - \mathbf{I})^T(\mathbf{H} - \mathbf{I})\epsilon \\
&= \frac{1}{N}\epsilon^T(\mathbf{H} - \mathbf{I})^2\epsilon \\
&= \frac{1}{N}\epsilon^T(\mathbf{I} - \mathbf{H})\epsilon
\end{aligned}$$

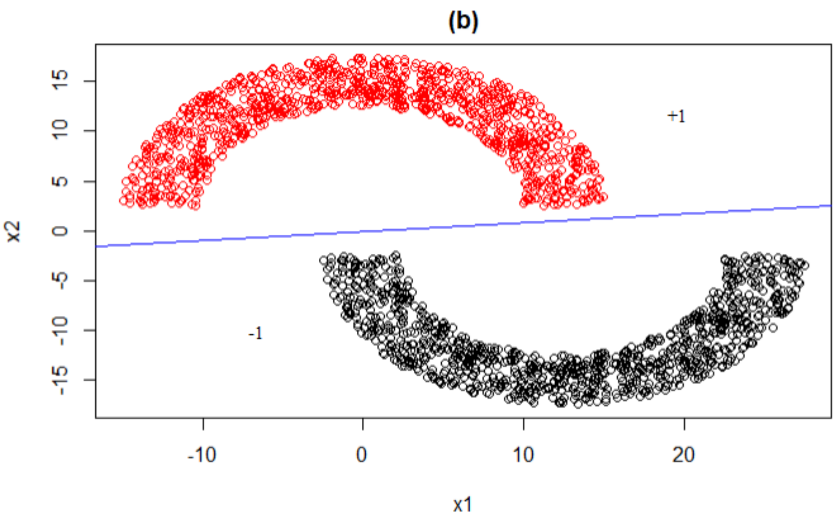
(d) We will now simplify the result of part (c) further. Note that the variance of $\epsilon = \sigma^2$.

$$\begin{aligned}
E_{\text{in}} &= \frac{1}{N}\epsilon^{\mathbf{T}}(\mathbf{I} - \mathbf{H})\epsilon \\
&= \frac{1}{N}\epsilon^T\mathbf{I}\epsilon - \frac{1}{N}\epsilon^T\mathbf{H}\epsilon \\
\mathbb{E}_{\mathcal{D}}[E_{\text{in}}] &= \mathbb{E}_{\mathcal{D}}\left[\frac{1}{N}\epsilon^T\mathbf{I}\epsilon\right] - \mathbb{E}_{\mathcal{D}}\left[\frac{1}{N}\epsilon^T\mathbf{H}\epsilon\right] \\
&= \frac{1}{N}\operatorname{Tr}(\epsilon^T\mathbf{I}\epsilon) - \frac{1}{N}\operatorname{Tr}(\epsilon^T\mathbf{H}\epsilon) \\
&= \sigma^2 - \sigma^2\frac{d+1}{N} \\
&= \sigma^2\left(1 - \frac{d+1}{N}\right)
\end{aligned}$$

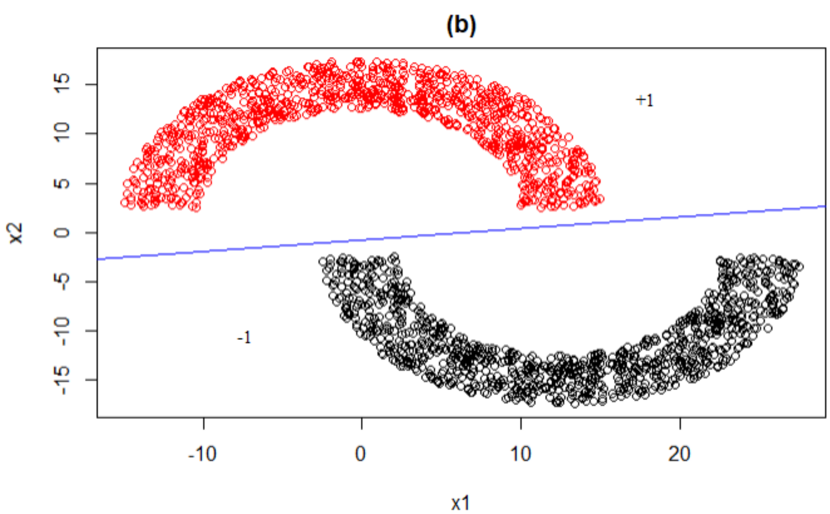
(e) Note that $\hat{\mathbf{y}} = \mathbf{X}\mathbf{w}^* + \mathbf{H}\epsilon$ holds as before. However, now we have that $\mathbf{y'} = \mathbf{X}\mathbf{w}^* + \epsilon'$ Therefore, we have $\hat{\mathbf{y}} - \mathbf{y}' = \mathbf{H}\epsilon - \epsilon'$. It follows that:

$$\begin{aligned}
E_{\text{test}} &= \frac{1}{N}(\mathbf{H}\epsilon - \epsilon')^T(\mathbf{H}\epsilon - \epsilon') \\
&= \frac{1}{N}(\epsilon^T\mathbf{H}^T - \epsilon'^T)(\mathbf{H}\epsilon - \epsilon') \\
&= \frac{1}{N}(\epsilon^T\mathbf{H}^T\mathbf{H}\epsilon - \epsilon^T\mathbf{H}^T\epsilon' - \epsilon'^T\mathbf{H}\epsilon + \epsilon'^T\epsilon') \\
&= \frac{1}{N}(\epsilon^T\mathbf{H}\epsilon - \epsilon^T\mathbf{H}\epsilon' - \epsilon'^T\mathbf{H}\epsilon + \epsilon'^T\epsilon') \\
\mathbb{E}_{\mathcal{D},\epsilon'}[E_{\text{test}}] &= \frac{1}{N}\mathbb{E}_{\mathcal{D},\epsilon'}[\epsilon^T\mathbf{H}\epsilon] - \frac{1}{N}\mathbb{E}_{\mathcal{D},\epsilon'}[\epsilon^T\mathbf{H}\epsilon'] - \frac{1}{N}\mathbb{E}_{\mathcal{D},\epsilon'}[\epsilon'^T\mathbf{H}\epsilon] + \mathbb{E}_{\mathcal{D},\epsilon'}[\epsilon'^T\epsilon] \\
&= \sigma^2\frac{d+1}{N} + 0 + 0 + \sigma^2 \\
&= \sigma^2\left(1 + \frac{d+1}{N}\right)
\end{aligned}$$

Problem 3.1.

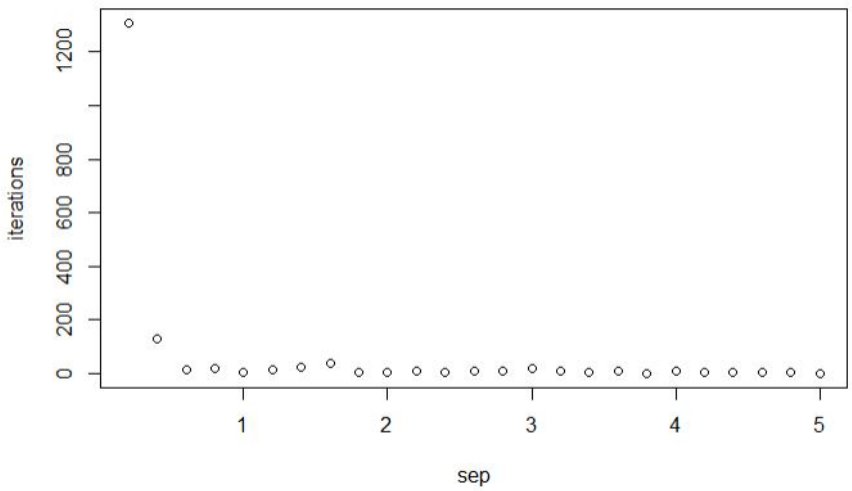


(a)



(b)

Both algorithms yield nearly identical solutions. They are both able to separate the data effectively. Thus, linear regression may be used as a classification algorithm to approximate the perceptron learning algorithm.



Problem 3.2.

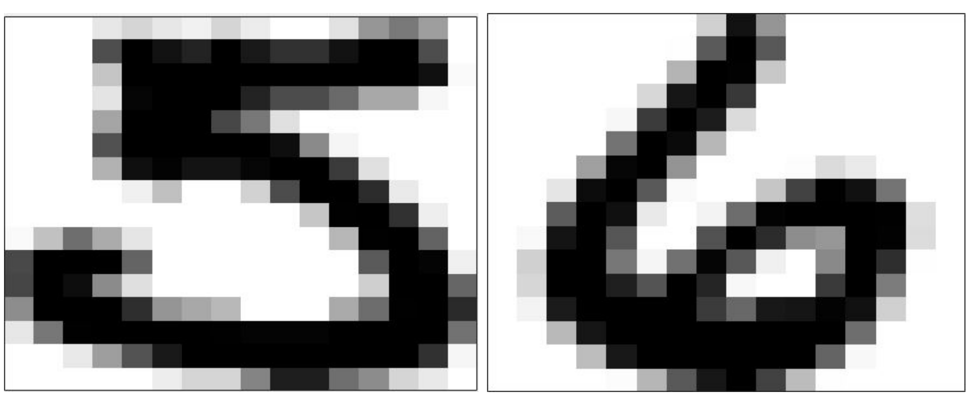
The number of iterations skyrockets as the window of separability narrows. We see that the number of iterations decreases in general as the separation increases.

Problem 3.8. Note that we can add zero creatively as $h^*(x) - h^*(x)$. It follows that:

$$\begin{aligned}
E_{\text{out}}(h) &= \mathbb{E}[(h(x) - y)^2] \\
&= \mathbb{E}[(h(x) - h^*(x) + h^*(x) - y)^2] \\
&= \mathbb{E}[(h(x) - h^*(x))^2] + \mathbb{E}[(h^*(x) - y)^2] + 2\mathbb{E}[(h(x) - h^*(x))(h^*(x) - y))] \\
&= \mathbb{E}[(h(x) - h^*(x))^2] + \mathbb{E}_x[\mathbb{E}_{y|x}[(h^*(x) - y)^2]] + 2\mathbb{E}_x[(h(x) - h^*(x))\mathbb{E}_{y|x}[h^*(x) - y])] \\
&= \mathbb{E}[(h(x) - h^*(x))^2] + \mathbb{E}_x[0^2] + 2\mathbb{E}_x[(h(x) - h^*(x)) \cdot 0)] \\
&= \mathbb{E}[(h(x) - h^*(x))^2]
\end{aligned}$$

Thus the out of sample error is clearly minimized when $h(x) = h^*(x)$. Note that $y = h^*(x) + (y - h^*(x)) = h^*(x) + \epsilon(x)$ where $\epsilon(x) = y - h^*(x)$. We have that $\mathbb{E}[\epsilon(x)] = \mathbb{E}_{y|x}[y - h^*(x)] = h^*(x) - h^*(x) = 0$.

Obtaining Features

a 

b I chose to use the first and second principal component from principal components analysis. By principal component I mean projection onto the eigenvectors of the covariance matrix.

c