Gary Chen
804361027

Homework 2

EE M146
April 17, 2019

Problem 1

Under the assumption that X^TX is singular where $X^TXy = 0$ can have solutions where, $y \neq 0$:

$$\begin{aligned} X^TXy &= 0 \\ y^TX^TXy &= 0 \\ (yX)^TXy &= 0 \\ \|Xy\|^2 &= 0 \\ Xy &= 0 \end{aligned}$$

Since $y \neq 0$ is a valid solution, X must have linearly dependent columns. Conversely, if we start off with $X^TXy = 0$ having only the trivial solution $y = 0$, then we find that $Xy = 0$ where only $y = 0$ is a valid solution. We can see that $X^TXy = 0$ iff X has linearly independent columns.

Problem 2

a) A symmetric matrix is a matrix where $A^T = A$. We first note that the inverse of a transpose is the transpose of an inverse, $(A^T)^{-1} = (A^{-1})^T$.

$$\begin{aligned} H^T &= (X(X^TX)^{-1}X^T)^T \\ &= (X^T)^T((XX^T)^{-1})^TX^T \\ &= X((XX^T)^{T)}-1X^T \\ &= X(X^TX)^{-1}X^T \\ &= H \end{aligned}$$

Where we see that H is symmetric.

b) Suppose that $H^K = H$, if we then consider H^{K+1}:

$$\begin{aligned} H^{K+1} &= HH \\ &= X(X^TX)^{-1}X^TX(X^TX)^{-1}X^T \end{aligned}$$

Where we find $X^T X (X^T X)^{-1} = 1$

$$
\begin{aligned}
&= X(X^T X)^{-1} X^T \\
&= H
\end{aligned}
$$

We know that this is true for $K = 1$, and thus, is true for $K = 2, 3, \ldots$ for all positive integer K.

c) Suppose that $(I - H)^K = I - H$, if we then consider $(1 - H)^{K+1}$:

$$
\begin{aligned}
(1 - H)^{K+1} &= (1 - H)(1 - H) \\
&= II - IH - HI + HH \\
&= I - 2H + HH
\end{aligned}
$$

Where we have already proved that $H^n = H$

$$= I - H$$

We know that this is true for $K = 1$, and thus, is true for $K = 2, 3, \ldots$ for all positive integer K.

d)

$$
\begin{aligned}
Tr(H) &= Tr(X(X^T X)^{-1} X^T) \\
&= Tr((X^T X)^{-1} X^T X) \\
&= Tr(I)
\end{aligned}
$$

Where I has dimension $M \times M$

$$= M$$

Problem 3

Given:

$$J(w_0, w_1) = \sum_{n=1}^{N} \alpha_n (w_0 + w_1 x_{n,1} - y_n)^2$$

We take the derivatives with respect to w_0 and w_1:

$$\begin{aligned}
\frac{\partial J}{\partial w_0} &= \frac{\partial}{\partial w_0}\sum_{n=1}^{N}\alpha_n(w_0 + w_1x_{n,1} - y_n)^2 \\
&= \sum_{n=1}^{N}\alpha_n\frac{\partial}{\partial w_0}(w_0 + w_1x_{n,1} - y_n)^2 \\
&= \boxed{2w_0\sum_{n=1}^{N}\alpha_n(w_0 + w_1x_{n,1} - y_n)}
\end{aligned}$$

$$\begin{aligned}
\frac{\partial J}{\partial w_1} &= \frac{\partial}{\partial w_1}\sum_{n=1}^{N}\alpha_n(w_0 + w_1x_{n,1} - y_n)^2 \\
&= \sum_{n=1}^{N}\alpha_n\frac{\partial}{\partial w_1}(w_0 + w_1x_{n,1} - y_n)^2 \\
&= \boxed{2\sum_{n=1}^{N}\alpha_n x_{n,1}(w_0 + w_1x_{n,1} - y_n)}
\end{aligned}$$

Having each index n have different weights in the form of α_n, items with lower α_n values do not contribute as strongly to the summation, while items with higher α_n values affect the derivative more.

Problem 4

First we take the derivative of $h_w(x)$:

$$\begin{aligned}
\frac{\partial h_w(x)}{\partial w_j} &= \frac{\partial}{\partial w_j}\left(\frac{1}{1+e^{-w^Tx}}\right) \\
&= (-x_je^{-w^Tx})(1+e^{-w^Tx})^{-2} \\
&= x_jh_w(x)(1-h_w(x))
\end{aligned}$$

Then taking the derivative of $J(w)$:

$$
\begin{aligned}
\frac{\partial J(w)}{\partial w_j} &= \frac{\partial}{\partial w_j}\left[-\sum_{n=1}^{N}[y_n \log(h_w(x_n)) + (1-y_n)\log(1-h_w(x_n))] + \frac{1}{2}\sum_i w_i^2\right] \\
&= -\sum_{n=1}^{N}\left[y_n \frac{\partial \log(h_w(x))}{\partial w_j} + (1-y_n)\frac{\partial \log(1-h_w(x_n))}{\partial w_j}\right] + w_j \\
&= -\sum_{n=1}^{N}\left[y_n \frac{x_j h_w(x_n)(1-h_w(x_n))}{h_w(x_n)} + (1-y_n)\frac{-x_j h_w(x_n)(1-h_w(x_n))}{(1-h_w(x_n))}\right] + w_j \\
&= -\sum_{n=1}^{N}[y_n x_j(1-h_w(x_n)) - (1-y_n)x_j h_w(x_n)] + w_j \\
&= \boxed{x_j \sum_{n=1}^{N}[h_w(x_n) - y_n] + w_j}
\end{aligned}
$$

Where we define the change in error to be

$$
\begin{aligned}
\nabla E_{in}(w_t)_j &= \frac{\partial J(w)}{\partial w_j} \\
&= x_j \sum_{n=1}^{N}[h_w(x_n) - y_n] + w_j
\end{aligned}
$$

For the update rule:

$$
w_{t+1} = w_t - \eta \frac{\nabla E_{in}(w_t)}{\|\nabla E_{in}(w_t)\|}
$$

Problem 5

We first want to take the maximum likelihood product to the log space. Since log is a monotonically increasing function, finding the maximum of the log is the same as finding the maximum of the function itself.

$$
\text{argmax}_w \prod_{i=1}^{n} P(y_i|x_i, w)f(w) \rightarrow \text{argmax}_w \sum_{i=1}^{n} \log(P(y_i|x_i, w)f(w))
$$

For the binary classification done in logistic regression, the conditional probabilities can be written as probabilities of Bernoulli random variables:

$$P(y_i|x_i, w) = \sigma(w^T x_i)^{y_i}(1-\sigma(w^T x_i))^{1-y_i}$$

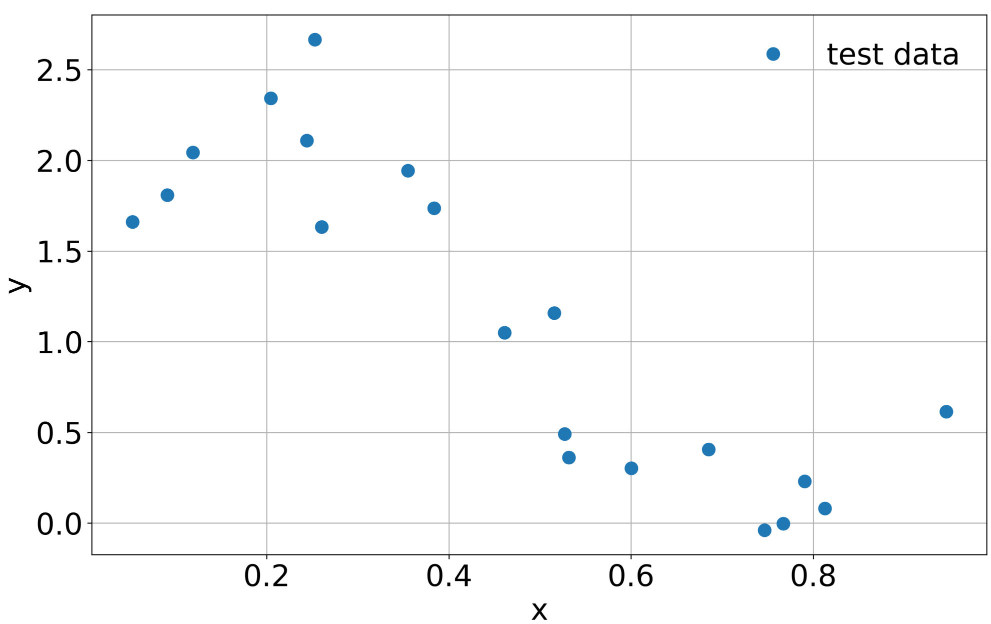
Considering $\text{argmax}_w \sum_{i=1}^{n} \log(P(y_i|x_i, w)f(w))$, we find:

$$\begin{aligned}
&\to \text{argmax}_w \sum_{i=1}^{n} \log\left[\sigma(w^T x_i)^{y_i}(1-\sigma(w^T x_i))^{1-y_i}\frac{1}{(2\pi)^{\frac{m}{2}}}e^{-\sum_{j=1}^{m}\frac{w_j^2}{2}}\right] \\
&= \text{argmax}_w \sum_{i=1}^{n}\left[y_i\log(\sigma(w^T x_i)) + (1-y_i)\log(1-\sigma(w^T x_i)) - \frac{m}{2}\log(2\pi) - \sum_{j=1}^{m}\frac{w_j^2}{2}\right] \\
&= \text{argmin}_w \left(-\sum_{i-1}^{n}[y_i\log(h_w(x_i)) + (1-y_i)\log(1-h_w(x_i))] + \frac{1}{2}\sum_{j=1}^{m} w_j^2\right)
\end{aligned}$$

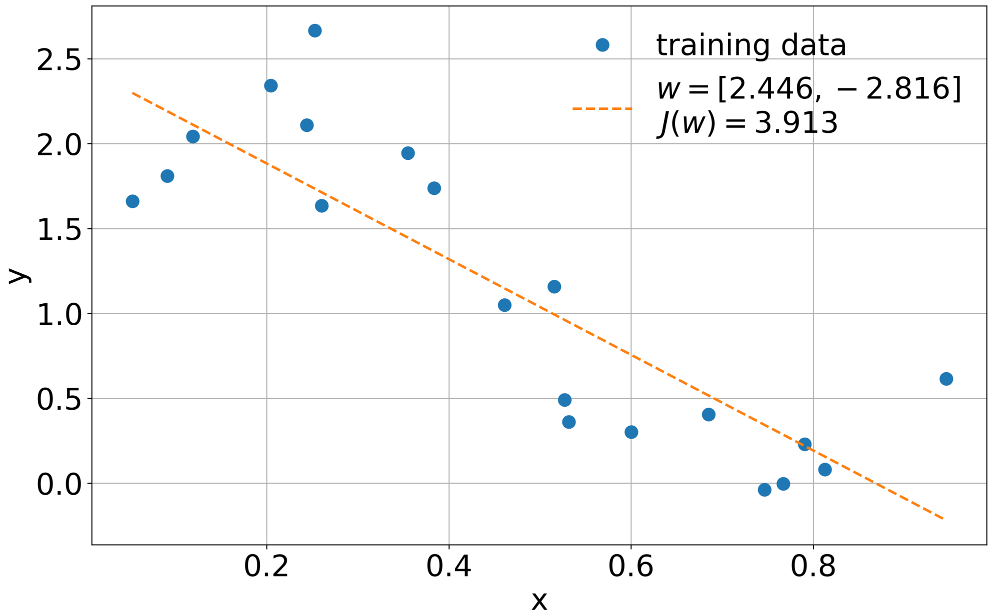
Where we see that the minimization of the argument is the same as the logistic regression object with L2 regularization.

Problem 6

a) The training data looks like it has a distinctive downward linear trend. I would believe that the linear regression will make a good prediction on this data set. As long as the test data is not vastly different, it will be a good model.



b) Plotted data shown here

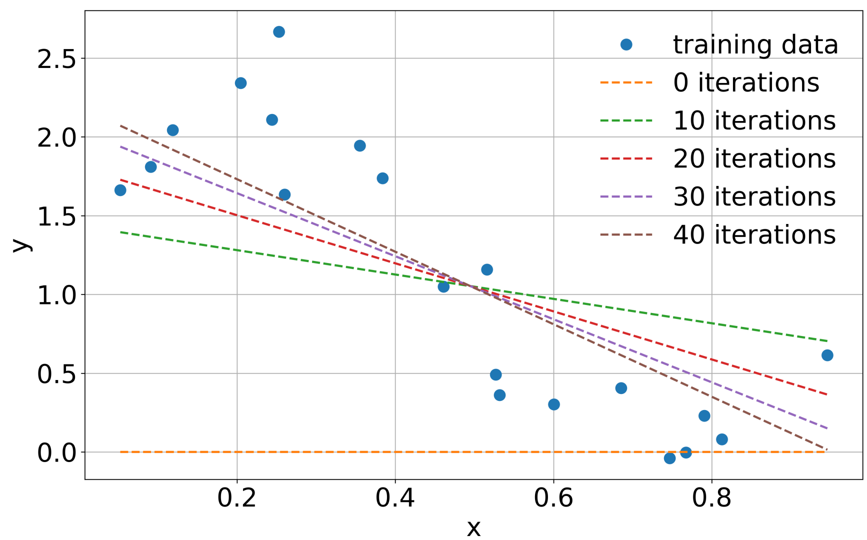


c) As the learning rate η is decreased, the gradient descent fares worse and worse to the point that it does not converge. The non-convergence just means more iterations are needed, but the ones that do converge still are farther away than ones of larger η values. I believe this is due to the fact that making the algorithm stop when the absolute value of the $J(w)$ difference is not a good stopping condition; this terminates the program when the change is small, not when an equilibrium has been reached.

η	iterations	$J(w)$
0.0407	104	3.914
0.01	364	3.917
0.001	2609	3.958
0.0001	10000	5.494

d) We find that the greater number of iterations used, the closer the values of w and $J(w)$ tend towards the closed form values found in part b).

iterations	w	$J(w)$
0	[0, 0]	40.234
10	[1.436, -0.773]	9.646
20	[1.808, -1.526]	6.199
30	[2.043, -2.002]	4.824
40	[2.192, -2.302]	4.276



e) From the plot of RMSE, I would say that the $m = 5$ results in the best fit to the data as it yields the lowest RMSE on the test data. For $m < 3$, we find that in general, adding more degrees lowers the RMSE for both training and test data, which seems like under-fitting. When $m > 8$, we see the test data's RMSE drastically rise, while the test data's RMSE continues to fall, which is indicative of overfitting.