COM S 573: Machine Learning

Homework #1

Abdurahman Mohammed

February 3, 2022

Question 1

A) The function will have two cases

- When $w^T x_i \geq 0$, $h(x) = 1$ since its using sign function
- When $w^T x_i < 0$, $h(x) = -1$ since its using sign function

So we can say that there is a separating line between those two regions. The equation of the line will be

$$w_0 x_0 + w_1 x_1 + w_2 x_2 = 0$$

and we know that $x_0 = 1$. Hence, the equation will be

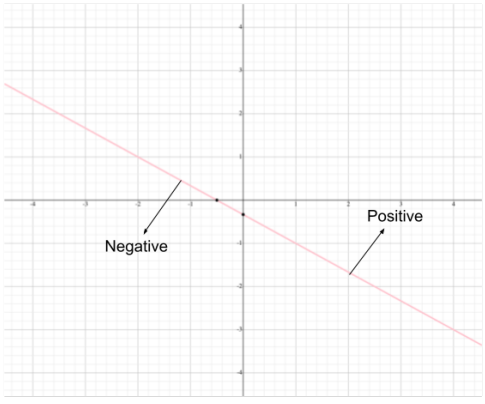
$$w_0 + w_1 x_1 + w_2 x_2 = 0$$

and it can be written as

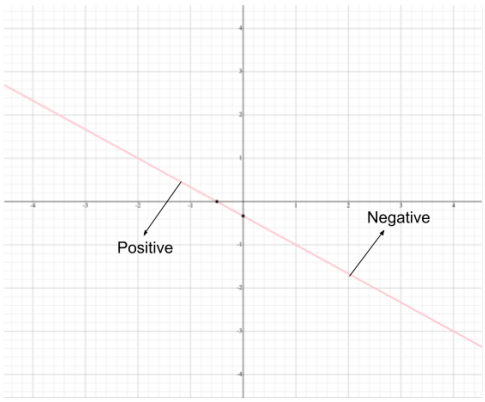
$$\begin{aligned} x_2 &= a x_1 + b \\ w_1 x_1 + w_2 x_2 + w_0 &= 0 \\ w_2 x_2 &= -w_1 x_1 - w_0 \\ x_2 &= \frac{-w_1 x_1}{w_2} - \frac{w_0}{w_2} \end{aligned}$$

Therefore, the slope a will be $-\frac{w_1}{w_2}$ and the intercept b will be $-\frac{w_0}{w_2}$

B) For $w = [1, 2, 3]^T$ the plot will look like



For $w = -[1, 2, 3]^T$ the plot will look like



Question 2

A) The following is the first order derivative $\nabla E(w)$

$$\begin{aligned}
\nabla E(w) &= \frac{1}{N}\sum_{n=1}^{N}\frac{\partial \ln(1+e^{-y_n w^T x_n})}{\partial w} \\
&= \frac{1}{N}\sum_{n=1}^{N}\frac{1}{1+e^{-y_n w^T x_n}}\left(\frac{\partial 1+e^{-y_n w^T x_n}}{\partial w}\right) \\
&= \frac{1}{N}\sum_{n=1}^{N}\frac{1}{1+e^{-y_n w^T x_n}}\left(\frac{\partial}{\partial w}(1+e^{-y_n w^T x_n})\right) \\
\nabla E(w) &= \frac{1}{N}\sum_{n=1}^{N}\frac{1}{1+e^{-y_n w^T x_n}}(-y_n x_n e^{y_n w^T x_n})
\end{aligned}$$

B) We know that $sigmoid(w^T x_i) = \frac{1}{1+e^{-w^T x_i}} = \frac{1}{2}$ where $\frac{1}{2}$ is the threshold.

$$\begin{aligned}
\frac{1}{1+e^{-w^T x_i}} &= \frac{1}{2} \\
2 &= 1+e^{-w^T x_i} \\
1 &= e^{-w^T x_i} \\
0 &= w^T x_i
\end{aligned}$$

Therefore, we can see that the equation is still linear.

C)

$$\begin{aligned}
\frac{1}{1+e^{-z}} &= \frac{9}{10} \\
\frac{1}{1+e^{-w^T x_i}} &= \frac{9}{10} \\
10 &= 9 + 9e^{-w^T x_i} \\
1 &= 9e^{-w^T x_i} \\
\frac{1}{9} &= e^{-w^T x_i} \\
\ln \frac{1}{9} &= -w^T x_i \\
-2.19722457734 &= -w^T x_i \\
w^T x_i &= 2.19722457734
\end{aligned}$$

The decision boundary is still linear. Therefore, even though we changed the threshold value to 0.9, the decision boundary stays linear.

D) Based on the observations from the above questions, we can say that since the Sigmoid function is a monotonic function, logistic regression will stay a linear model and that is the essential property of it.

Question 3

We are given $X = [x_1, x_2, ...x_n] \in \mathbb{R}^{m \times n}$ and $Y = \begin{bmatrix} y_1^T \\ y_2^T \\ . \\ . \\ . \\ y_n^T \end{bmatrix}$ and we want to show that $XY = \sum_{i=1}^{n} x_i y_i^T$.

We will perform matrix multiplication.

$$\begin{aligned}
XY &= \begin{bmatrix} x_1 & x_2 & . & . & .x_n \end{bmatrix} \begin{bmatrix} y_1^T \\ y_2^T \\ . \\ . \\ . \\ y_n^T \end{bmatrix} \\
&= x_1 y_1^T + x_2 y_2^T + x_3 y_3^T + ... + x_n y_n^T \\
&= \sum_{i=1}^{n} x_i y_i^T
\end{aligned}$$

Therefore, $XY = \sum_{i=1}^{n} x_i y_i^T$.

Question 4

A) To find optimal w^* and σ^* we will calculate first order derivatives with respect to w and σ. Let's first calculate w^*. We know that $RSS(w) = (y - w^T x)^T (y - w^T x)$

$$\begin{aligned}\frac{\partial}{\partial w} \log \mathcal{L}(w|x) &= -\frac{1}{2}\frac{\partial}{\partial w}(\frac{1}{\sigma^2} RSS(w) + n \log \sigma^2) + const \\ &= -\frac{1}{2\sigma^2}\frac{\partial}{\partial w}(y^T y - 2w^T x^T y + w^T w x^T x) \\ &= -\frac{1}{2\sigma^2}(0 - 2x^T y + w x^T x)\end{aligned}$$

We will now find optimal w

$$\begin{aligned}0 &= -\frac{1}{2\sigma^2}(0 - 2x^T y + w x^T x) \\ 2x^T y &= 2w x^T x \\ w^* &= (x^T x)^{-1} x^T y\end{aligned}$$

Now we will find σ^*

$$\begin{aligned}\frac{\partial}{\partial \sigma} \log \mathcal{L}(w|x) &= -\frac{1}{2}\frac{\partial}{\partial \sigma}(\frac{1}{\sigma^2} RSS(w) + n \log \sigma^2) + const \\ &= \sigma^{-3} RSS(w) - \frac{n}{\sigma} \\ &= \frac{RSS(w)}{\sigma^3} - \frac{n}{\sigma} \\ &= \frac{RSS(w) - n}{\sigma^3}\end{aligned}$$

Now we will find optimalσ

$$\begin{aligned}RSS(w) &= n\sigma^2 \\ \sigma^2 &= \frac{RSS(w)}{n} \\ \sigma^* &= \sqrt{\frac{RSS(w)}{n}}\end{aligned}$$

B) We know that the RSS measures the discrepancy between the actual label and predicted ones. From the equation of the σ^*, we can see that it is equal to the square root of average RSS and it resembles to the formula of standard deviation. This tells us that the variance occurs when the model we have is highly sensitive to fluctuations.

Question 5

We know that

$$sigmoid(y) = \frac{1}{1 + e^- y}$$

and

$$softmax(y) = \frac{e^{y_j}}{\sum_{i=1}^{c} e^{y_i}}$$

But for binary classification, we know that $c = 2$. Therefore, the softmax function with $c = 2$ will look like

$$\begin{aligned} softmax(y) &= \frac{e^{y_j}}{\sum_{i=1}^{2} e^{y_i}} \\ &= \frac{e^{y_j}}{e^{y_1} + e^{y_2}} \end{aligned}$$

For a binary logistic regression

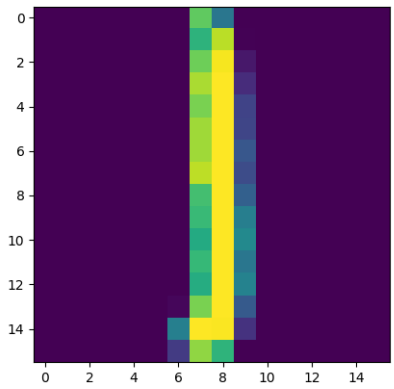
$$\begin{aligned} softmax(y_0) &= \frac{e^{y_0}}{e^{y_0} + e^{y_1}} \\ &= \frac{e^{w_0^T x_i}}{e^{w_0^T x_i} + e^{w_1^T x_i}} \\ &= \frac{e^{w_0^T x_i}}{e^{w_1^T x_i}(1 + e^{(w_0^T - w_1^T)x_i})} \\ &= \frac{e^{(w_0^T - w_1^T)x_i}}{1 + e^{(w_0^T - w_1^T)x_i}} \\ &= \frac{e^z}{1 + e^z} \end{aligned}$$

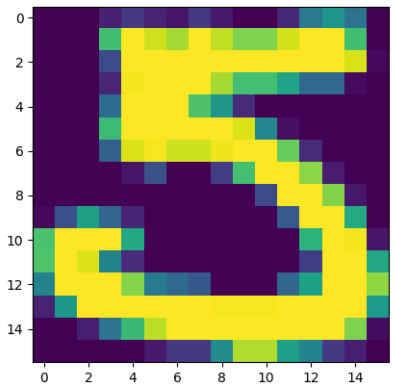
$$\begin{aligned} softmax(y_1) &= \frac{e^{y_1}}{e^{y_0} + e^{y_1}} \\ &= \frac{e^{w_1^T x_i}}{e^{w_0^T x_i} + e^{w_1^T x_i}} \\ &= \frac{1}{1 + e^{(w_0^T - w_1^T)x_i}} \\ &= \frac{1}{1 + e^z} \end{aligned}$$

From the above derivations, we can see that softmax for binary logistic regression is the same as sigmoid.

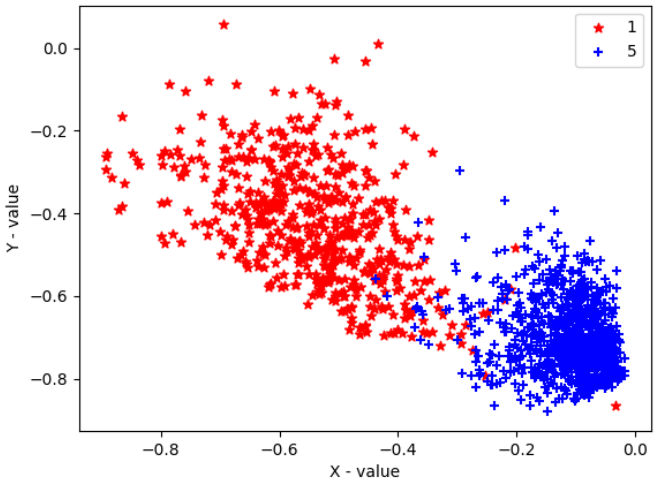
Question 6

A) Below are the images plotted by completing *show_images* function.





B) Using the two features extracted (Symmetry and average intensity), I have completed the **show_features** method and plotted the features on a scatter plot.



C) Training and Test Accuracy for all 5 cases after completing the Perceptron class.

```
Case 1: max iteration:10  train accuracy:0.980782  test accuracy: 0.957547.
Case 2: max iteration:30  train accuracy:0.983985  test accuracy: 0.959906.
Case 3: max iteration:50  train accuracy:0.973735  test accuracy: 0.938679.
Case 4: max iteration:100  train accuracy:0.974375  test accuracy: 0.938679.
Case 5: max iteration:200  train accuracy:0.978860  test accuracy: 0.943396.
```

D) Plot for show_result method