COMP4434 Homework 2

ZHANG Caiqi 18085481d

March 8, 2021

1 Question 1

1.1 a)

- L = | $h_\theta(x) - y$ | cannot fix the problem. It will still penalizes confident correct predictions. For example, if the $h_\theta(x)$ is under 0 and extremely small, it should be classify to 0 and the loss should be very small. However, using this loss function will lead to a large cost. Therefore, the absolute value still cannot fix the problem.

- The second version is acceptable and the missing part is: 1 - min(1, $h_\theta(x)$). If the $h_\theta(x)$ is under 0 and extremely small, it will be classified to 0 and the cost will be 0. When $h_\theta(x)$ is very large, it will be classified to 1 and the cost 1 - min(1, $h_\theta(x)$) = 0. This loss function can solve the problem.

1.2 b)

Conclusion: Model prediction rule will not change. The learnt model parameters will be the opposite to the origin one.
Explanation: First, we can find the following relationship.

$$\frac{1}{1+e^{-z}} + \frac{e^{-z}}{1+e^{-z}} = 1$$

If we still use the binary cross entropy loss function, as the activation function is changed, we need to use new method to update θ. Unlikely to the original sigmoid function, the new one $\frac{e^{-z}}{1+e^{-z}}$ is a decreasing function, which means the larger the z is, the smaller the function value will be. Therefore, we should sightly change the update function of θ to make it gradient ascent. Finally, the learnt model parameters θ is the opposite of the conventional logistic regression one.

$$\theta_j = \theta_j + \alpha \frac{\partial J(\theta_0, \theta_1, \cdots)}{\partial \theta_j}$$

About the model prediction rule, we should not change it. For example, for the data with label 1, the model is trying to make the cost as low as possible, in other words, to get a θ

to make the $h_\theta(x)$ closer to 1. Therefore, no matter which sigmoid function we use, we will still classify the data to 1 if $h_\theta(x) \geq 0.5$.
Experiment proof: We can use the the Question 2 in this assignment to prove the conclusion. The original θ after 50000 iteration is: [-10.748, 10.989, 50.422, -17.668]. If we change the sigmoid function, the final θ will be: [10.750, -10.993, -50.435, 17.672]. Meanwhile, using the same prediction rule can get the correct label.

2 Question 2

2.1 a)

We firstly change the label -1 to 0.

$$h_\theta(x) = \frac{1}{1 + e^{-\theta^T x}}$$

$(x)^1$ = [1, -0.12, 0.3, -0.01], $(x)^2$ = [1, 0.2, -0.03, -0.35], $(x)^3$ = [1, -0.37, 0.25, 0.07], $(x)^4$ = [1, -0.1, 0.14, -0.52]. $h_\theta(x)^1 = 0.464$, $h_\theta(x)^2 = 0.497$, $h_\theta(x)^3 = 0.462$, $h_\theta(x)^4 = 0.498$. They are all smaller than 0.5. The predicted label is [0, 0, 0, 0] under our assumption because we change label -1 to 0.

2.2 b)

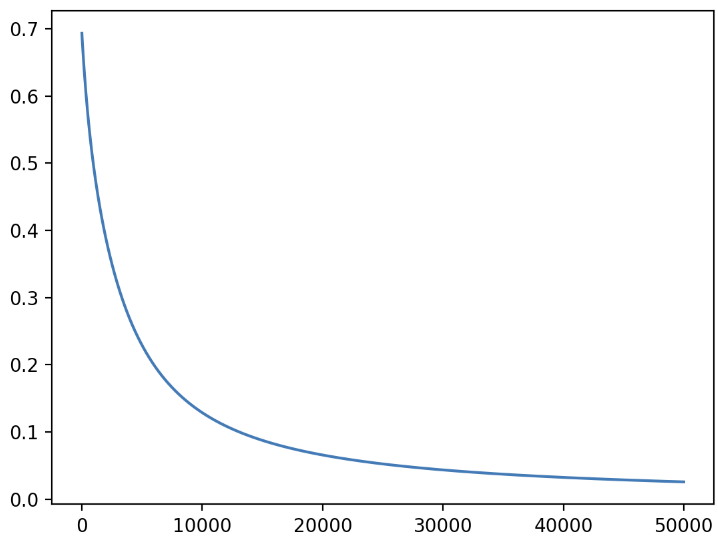
$$\theta_j = \theta_j - \alpha \frac{\partial J(\theta_0, \theta_1, \cdots)}{\partial \theta_j}$$

$$\frac{\partial J(\theta_0, \theta_1, \cdots)}{\partial \theta_j} = \frac{1}{m} \sum_{i=1}^{m} \left(h_\theta\left(x^{(i)}\right) - y^{(i)}\right) x_j^{(i)}$$

In this question, $\alpha = 0.2$.
-0.09 - 0.2 * (1 / 4) * -0.079 = -0.086
0 - 0.2 * (1 / 4) * 0.043 = -0.002
-0.19 - 0.2 * (1 / 4) * -0.131 = -0.183
-0.21 - 0.2 * (1 / 4) * 0.125 = -0.216
θ_1 = [-0.086, -0.002, -0.183, -0.216]
-0.086 - 0.2 * (1 / 4) * -0.072 = -0.082
-0.002 - 0.2 * (1 / 4) * 0.042 = -0.004
-0.183 - 0.2 * (1 / 4) * -0.129 = -0.177
-0.216 - 0.2 * (1 / 4) * 0.123 = -0.222
θ_2 = [-0.082, -0.004, -0.177, -0.222]

Therefore, θ_1 = [-0.086, -0.002, -0.183, -0.216], θ_2 = [-0.082, -0.004, -0.177, -0.222].

2.3 c)


0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0
10000
20000
30000
40000
50000