CS589 Homework #5

Trung Dang

October 19, 2023

FOR TYPESET SIMPLICITY, WE WILL WRITE x_p INSTEAD OF $\dot{x}_p$ IN THE FOLLOWING SOLUTION

1 Task 1

Show that the multi-class Softmax cost reduces to the two-class softmax cost when $C = 2$ and $y_p \in \{1, -1\}$

1.1 Proof

Indeed, the formula for the multiclass Softmax cost function is:

$$g(w_0, \cdots, w_{C-1}) = \frac{1}{P} \sum_{p=1}^{P} \left[\log(\sum_{j=0}^{C-1} e^{x_p^T w_j}) - x_p^T w_{y_p} \right]$$

substituting $C = 2$, we have:

$$\begin{aligned} g(w_0, w_1) &= \frac{1}{P} \sum_{p=1}^{P} \left[\log(e^{x_p^T w_0} + e^{x_p^T w_1}) - x_p^T w_{y_p} \right] \\ &= \frac{1}{P} \sum_{p=1}^{P} \left[\log(e^{x_p^T w_0} + e^{x_p^T w_1}) - \log(e^{x_p^T w_{y_p}}) \right] \\ &= \frac{1}{P} \sum_{p=1}^{P} \left[\log(\frac{e^{x_p^T w_0} + e^{x_p^T w_1}}{e^{x_p^T w_{y_p}}}) \right] \end{aligned}$$

Let $w_{y_p} = w_1$ if $y_p = 1$ and $w_{y_p} = w_0$ if $y_p = -1$. Then

$$g(w_0, w_1) = \begin{cases} \frac{1}{P} \sum_{p=1}^{P} \left[\log(1 + \frac{e^{x_p^T w_0}}{e^{x_p^T w_1}}) \right] & \text{if } y_p = 1 \\ \frac{1}{P} \sum_{p=1}^{P} \left[\log(1 + \frac{e^{x_p^T w_1}}{e^{x_p^T w_0}}) \right] & \text{if } y_p = -1 \end{cases}$$
$$= \begin{cases} \frac{1}{P} \sum_{p=1}^{P} \log(1 + e^{x_p^T (w_0 - w_1)}) & \text{if } y_p = 1 \\ \frac{1}{P} \sum_{p=1}^{P} \log(1 + e^{x_p^T (w_1 - w_0)}) & \text{if } y_p = -1 \end{cases}$$

Let $w = w_1 - w_0$, then

$$g(w) = \frac{1}{P} \sum_{p=1}^{P} \left[\log(1 + e^{-y_p x_p^T w}) \right]$$

■

1.2 Proof explanation

1. First we substitute $C = 2$ into the cost function
2. Then we convert $x_p^T w_{y_p}$ to $\log(e^{x_p w_{y_p}})$
3. we know that $\log(a) - \log(b) = \log(\frac{a}{b})$, so we get the division
4. By substituting w_{y_p} for the corresponding values as explained in the proof, and extracting 1 from the fractions, we get the desired form
5. By substituting $w_1 - w_0$ for w, we have Q.E.D

FOR TYPESET SIMPLICITY, WE WILL WRITE x_p INSTEAD OF $\dot{x_p}$ IN THE FOLLOWING SOLUTION.

2 Task 2

Demonstrate that when $C = 2$ and, $y_p \in \{0, 1\}$ the multi-class Softmax is equivalent to the two-class Cross Entropy cost.

2.1 Proof

Substituting $C = 2$, and By the same transformation above we have:

$$\begin{aligned} g(w_0, w_1) &= \frac{1}{P}\sum_{p=1}^{P}\left[\log(e^{x_p^T w_0} + e^{x_p^T w_1}) - x_p^T w_{y_p}\right] \\ &= \frac{1}{P}\sum_{p=1}^{P}\left[\log(e^{x_p^T w_0} + e^{x_p^T w_1}) - \log(e^{x_p^T w_{y_p}})\right] \\ &= \frac{1}{P}\sum_{p=1}^{P}\left[\log(\frac{e^{x_p^T w_0} + e^{x_p^T w_1}}{e^{x_p^T w_{y_p}}})\right] \\ &= -\frac{1}{P}\sum_{p=1}^{P}\left[\log(\frac{e^{x_p^T w_{y_p}}}{e^{x_p^T w_0} + e^{x_p^T w_1}})\right] \end{aligned}$$

Let $w_{y_p} = w_1$ if $y_p = 1$ and $w_{y_p} = w_0$ if $y_p = 0$.
Then

$$\begin{aligned} g(w_0, w_1) &= \begin{cases} -\frac{1}{P}\sum_{p=1}^{P}\left[\log\left(1 - \frac{1}{1+\frac{e^{x_p^T w_0}}{e^{x_p^T w_1}}}\right)\right] \text{ if } y_p = 0 \\ -\frac{1}{P}\sum_{p=1}^{P}\left[\log\left(\frac{1}{1+\frac{e^{x_p^T w_0}}{e^{x_p^T w_1}}}\right)\right] \text{ if } y_p = 1 \end{cases} \\ &= \begin{cases} -\frac{1}{P}\sum_{p=1}^{P}\log\left(1 - \frac{1}{1+e^{x_p^T(w_0 - w_1)}}\right) \text{ if } y_p = 0 \\ -\frac{1}{P}\sum_{p=1}^{P}\log\left(\frac{1}{1+e^{x_p^T(w_0 - w_1)}}\right) \text{ if } y_p = 1 \end{cases} \end{aligned}$$

Let $w = w_1 - w_0$, we have:

$$g(w) = \begin{cases} -\frac{1}{P}\sum_{p=1}^{P}\log(1 - \sigma(x_p^T w)) \text{ if } y_p = 0 \\ -\frac{1}{P}\sum_{p=1}^{P}\log(\sigma(x_p^T w)) \text{ if } y_p = 1 \end{cases}$$

$$g(w) = -\frac{1}{P}\sum_{p=1}^{P}\left(y_p \log(\sigma(x_p^T w)) + (1-y_p)\log(1-\sigma(x_p^T w))\right)$$

■

2.2 Proof explanation

1. First we substitute $C = 2$ into the cost function
2. Then we convert $x_p^T w_{y_p}$ to $\log(e^{x_p w_{y_p}})$
3. we know that $\log(a) - \log(b) = \log(\frac{a}{b})$, so we get the division
4. By substituting w_{y_p} for the corresponding values as explained in the proof, and extracting 1 from the fractions, we get the desired form
5. Notice that:

$$\begin{aligned}\frac{e^{x_p^T w_0}}{e^{x_p^T w_0} + e^{x_p^T w_1}} &= 1 - \frac{e^{x_p^T w_1}}{e^{x_p^T w_0} + e^{x_p^T w_1}} \\ &= 1 - \frac{1}{1 + \frac{e^{x_p^T w_0}}{e^{x_p^T w_1}}}\end{aligned}$$

   by dividing both the denominator and numerator by $e^{x_p^T w_1}$
6. By substituting $w_1 - w_0$ for w, we have Q.E.D

3 Task 3

In this task, we are using the multiclass Softmax Optimization method, with a soft-margin weighted addition of the sum of squares of the 2-norm of feature-touching weights. The full formula is:

$$g(w_0, \cdots, w_3) = \frac{1}{P} \sum_{p=1}^{P} \left[\log(\sum_{j=0}^{3} e^{x_p^T w_j}) - x_p^T w_{y_p} \right] + \lambda \sum_{c=0}^{3} \|w_c\|_2^2$$

In the model, we chose the parameters as follows:

$$\begin{cases} \alpha = 1.1e^{-1} \\ \lambda = 1e^{-5} \\ \text{iterations } = 5000 \end{cases}$$

For the learning rate α, I have attempted to run the model at $\alpha = 1, 1e^{-1}$, and $1e^{-2}$. The model at $1e^{-2}$ was largely underfitting, as shown in the appendix.

For the λ, I attempted running with $\lambda = 1e^{-i}$ for $i \in \{0, \cdots, 5\}$, and found that the loss increases as λ gets larger. Thus, I chose the optimal lambda of $1e^{-5}$

For the number of iterations, I have run the model with as few as 500 iterations and as many as $100,000$ iterations. While any run with larger than 1000 iterations does not offer better accuracy, it does reduce the cost of the model. To avoid overfitting, I used 5000 iterations. Still, I demonstrate the results of $1,000$ and $100,000$ iterations in the appendix for reference.

The final accuracy is: $\boxed{0.75}$
The final cost is: $\boxed{0.5298495292663574}$

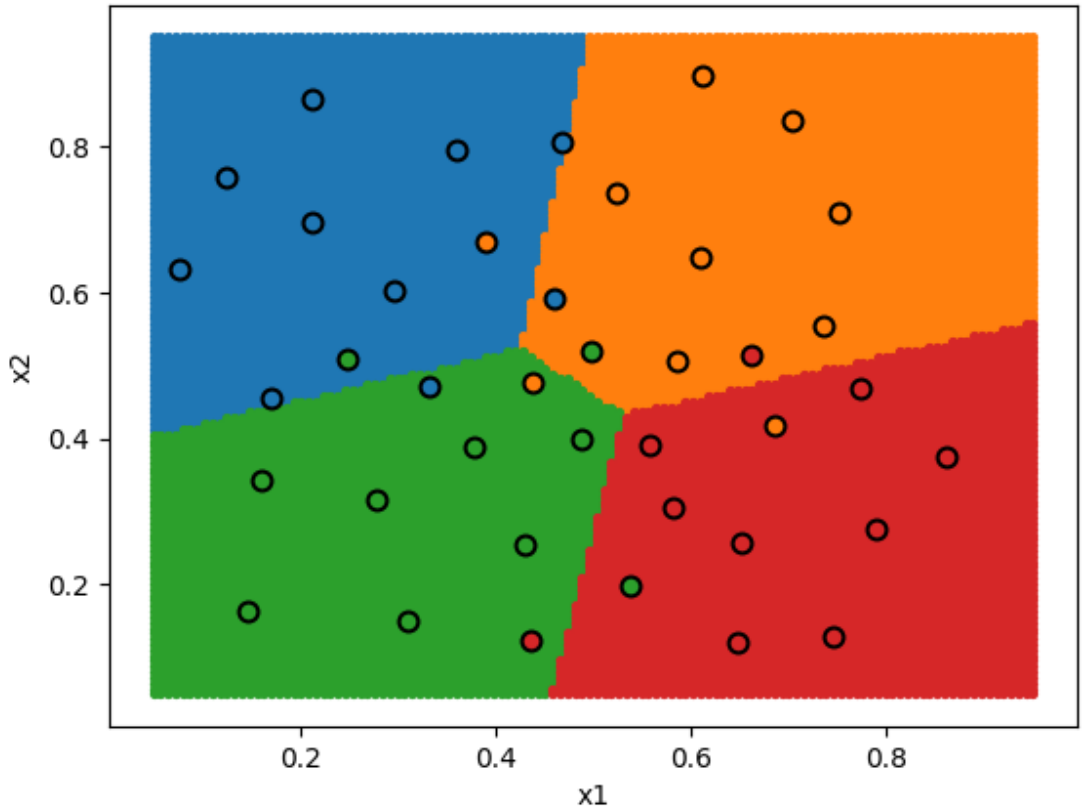

Figure 1: The data and regions of the solution model, wherein $\alpha = 1.1e^{-1}, \lambda = 1e^{-5}$, iterations $= 5000$

Appendix

As illustrated below are a few graphs, accuracies, and costs of different runs with different hyperparameters

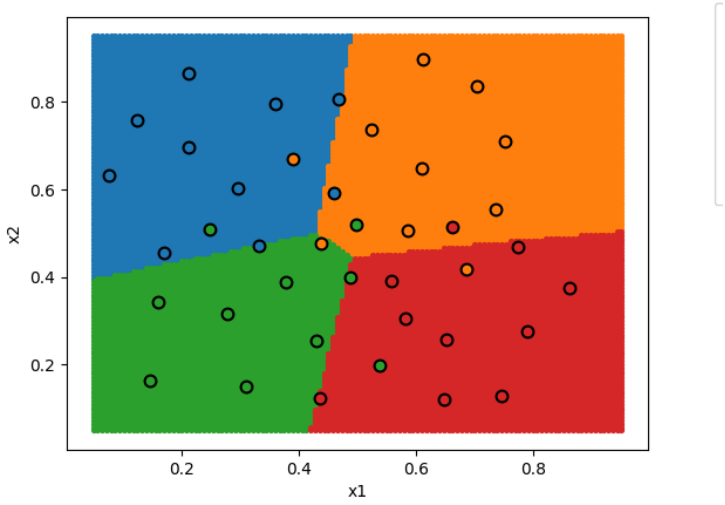

Figure 2: Model run with 1000 iterations

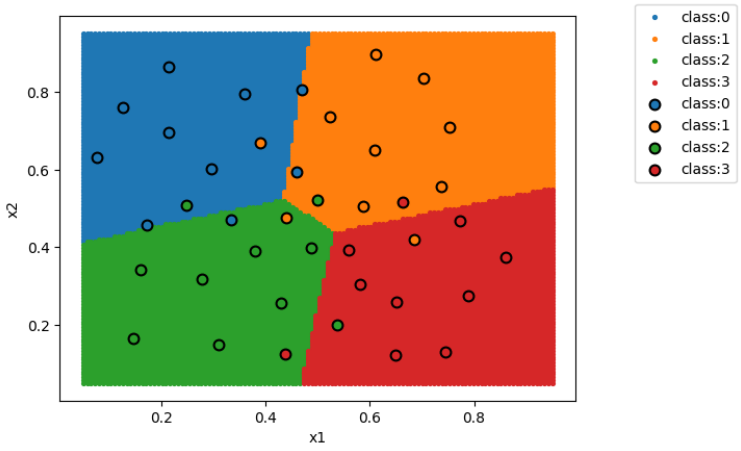

Figure 3: Model run with 100,000 iterations

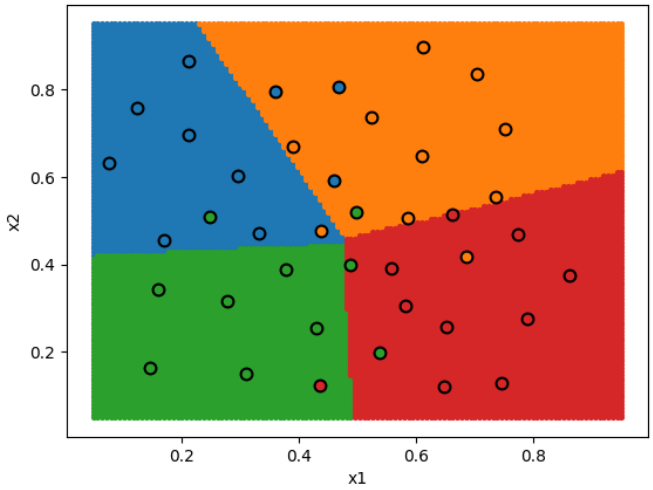

Figure 4: Model run with $\alpha = 1e^{-2}$

λ	Accuracy	Cost
1	0.775	1.269
$1e^{-1}$	0.725	0.917
$1e^{-2}$	0.725	0.923
$1e^{-3}$	0.75	0.711
$1e^{-4}$	0.725	0.674
$1e^{-5}$	0.75	0.530

Table 1: Accuracy and Cost of Model with different λ