ITE4053 - practice2-1 report

Junyeong Park

1 Method

In this practice, we used 2-layer feed forward network. The input and output for each layer is $\mathbb{R}^1$, therefore, the shape of weight and bias for each layer is $\mathbb{R}^{1\times1}$. And we used sigmoid function as activation function. In summary, the formula of the model is as follows:

$$f(\mathbf{x}) = \sigma(W_2\sigma(W_1\mathbf{x} + b_1) + b_2) \tag{1}$$

where

$$\sigma(x) = \frac{1}{1+\exp(-x)} \tag{2}$$

Before training, we initialized model parameters randomly. The weight parameters are initialized with normal distribution $\mathcal{N}(0, \sqrt{2})$. The bias parameters are initialized to zero. We trained the model with simple gradient descent algorithm and cross entropy energy function. And during training and prediction, we normalize the input of the model to improve model performance. The range of x value in datasets is from 0 to 360.

2 Results

2.1 Estimated parameters

The best parameter that we got in these experiments is $W_1 = [-9.96751852]^\top$, $b_1 = -0.14833009$, $W_2 = [10.61349984]^\top$, and $b_2 = -4.89201807$. Figure **??** represents the output of the best model.

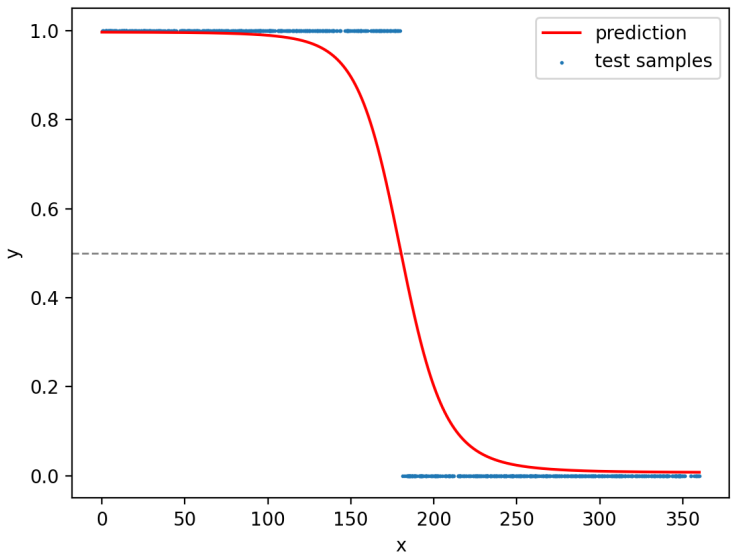

Figure 1: The best model for these experiments.

2.2 Best hyperparameter α

To get the best hyperparameter, we tests various candidates. In Fig. 2, the best hyperparameter α is 1.0 that fastest converges.

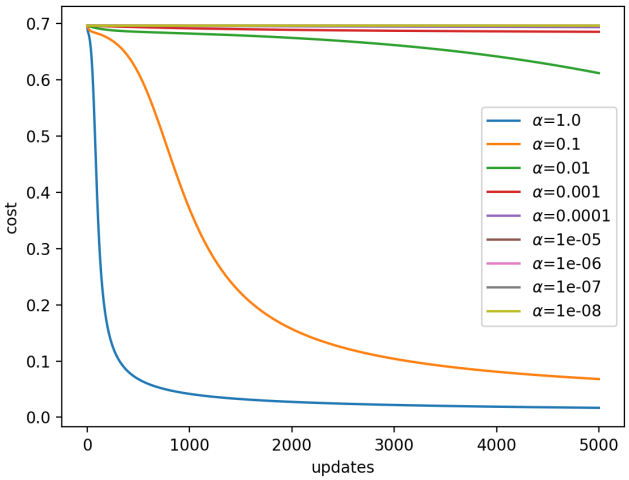

Figure 2: Model training performances for each hyperparameter α.

2.3 Accuracy

We frozen random seeds to reduce noise that is not an associated variable of the experiments. Below results are from the same datasets and the same initial parameters.

	(10, 1000, 5000)	(100, 1000, 5000)	(10000, 1000, 5000)
Accuracy (training set)	100.0	100.0	99.95
Accuracy (test set)	99.0	98.3	100.0

Table 1: Performances according to the number of training samples. (m, n, K) denotes that the model was updated K times and that m training samples, and n test samples were used.

	(10000, 1000, 10)	(10000, 1000, 100)	(10000, 1000, 5000)
Accuracy (training set)	50.22	98.49	99.95
Accuracy (test set)	55.60	98.4	100.0

Table 2: Performances according to the number of updates. (m, n, K) denotes that the model was updated K times and that m training samples, and n test samples were used.

3 Discussion

We checked that the norm of the parameters of the first and second layers of the second layers was greater. This is thought to be due to the nature of the differential value of a sigmoid function. The gradient value of the sigmoid function is up to 0.25. For layers except the last layer, the differential value of the sigmoid is multiplied when the gradient calculation is calculated. Therefore the higher the layer, the smaller the gradient, so the parameters wouldnt have changed much when updating.