ITE4053 - practice3-2 report

Junyeong Park

1 Method

All of experiment settings are almost same with the last practice (practice 3-1). There are two different settings; the model architecture and the optimizer. In the last practice, we used 2-layered feed forward network that each layers have just one unit. However, for now, we use 2-layered feed forward network that 1st layer has two units and 2nd layer has one unit. And in the last practice, we used vanilla stochastic gradient descent (SGD) optimizer. But, for this practice, we use Momentum optimizer with momentum 0.9.

2 Results

2.1 Estimated parameters

The best parameter that we got in these experiments is as follows:

$$W^{[1]} = \begin{bmatrix} -29.23949244 \\ 27.13686152 \end{bmatrix}, \; b^{[1]} = \begin{bmatrix} 7.13147819 \\ 6.61166649 \end{bmatrix}, \; W^{[2]} = \begin{bmatrix} -20.82970952 & -20.6492181 \end{bmatrix}, \; b^{[2]} = 30.24641471$$

Figure 1 represents the output of the best model.

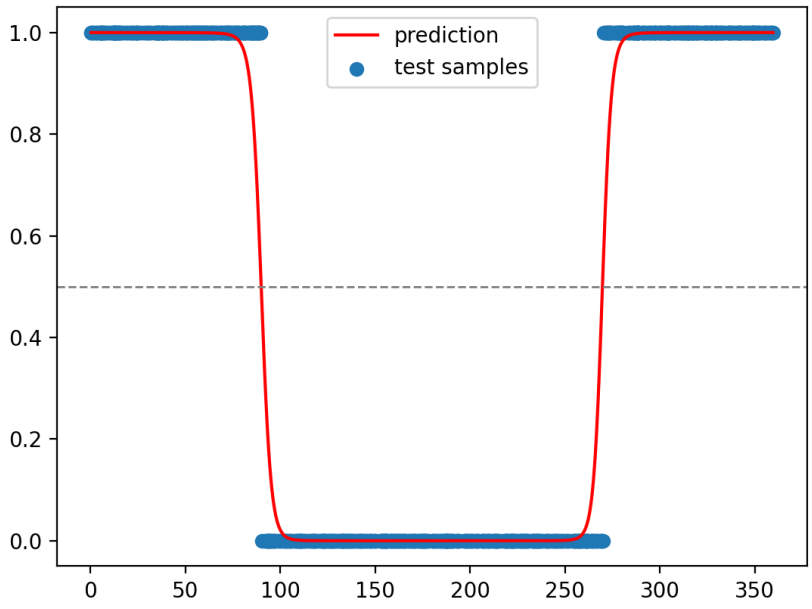

Figure 1: The best model for these experiments.

2.2 Best hyperparameter α

To get the best hyperparameter, we tests various candidates. In Fig. 2, the best hyperparameter α is 0.001 that fastest converges.

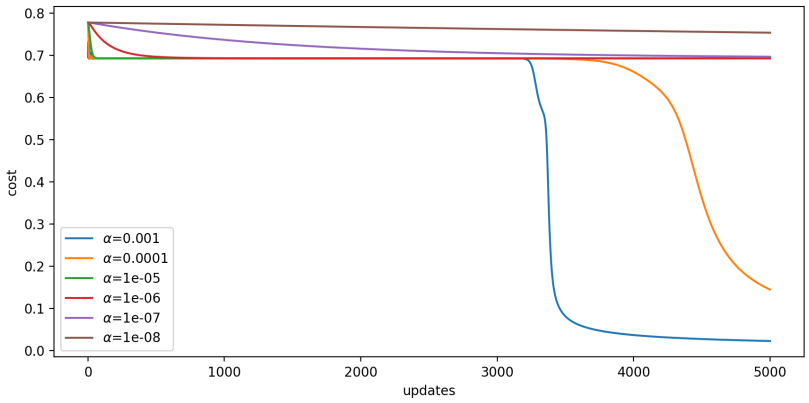

Figure 2: Model training performances for each hyperparameter α.

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)	60.0	60.0	99.94
Accuracy (test set)	50.5	57.8	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)	49.3	50.22	99.94
Accuracy (test set)	50.3	48.4	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

A minimum of two straight lines are required to classify this data. From this practice, we came to realize that a node in a layer represents a line. Because what changed from the previous exercise is that the number of nodes in the first layer has been increased to two. If we use just two lines, we will be able to use two straight lines symmetrical to $x = 180$. Actually, the best model parameter was able to confirm that it came out according to our intuition. In detail, $W_1^{[1]} \approx -W_2^{[1]}$.