ITE4053 - practice3-1 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. Whats different from previous experiments is that they used the cosine function for dataset generation.

2 Results

2.1 Estimated parameters

The best parameter that we got in these experiments is $W_1 = [-15.47286836]^\top$, $b_1 = -5.71985667$, $W_2 = [8.1135835]^\top$, and $b_2 = -0.67328457$. Figure 1 represents the output of the best model.

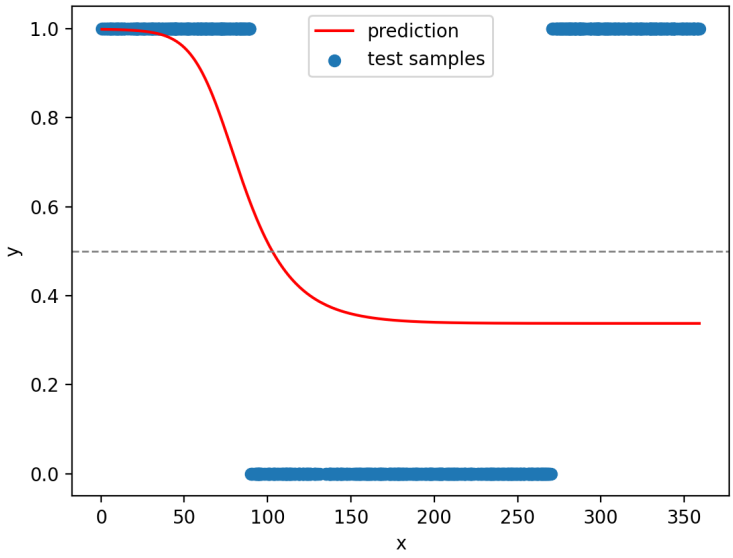

Figure 1: The best model for these experiments.

2.2 Best hyperparameter α

To get the best hyperparameter, we tests various candidates. In Fig. 2, the best hyperparameter α is 1.0 that fastest converges. And it makes the model best working at the test dataset.

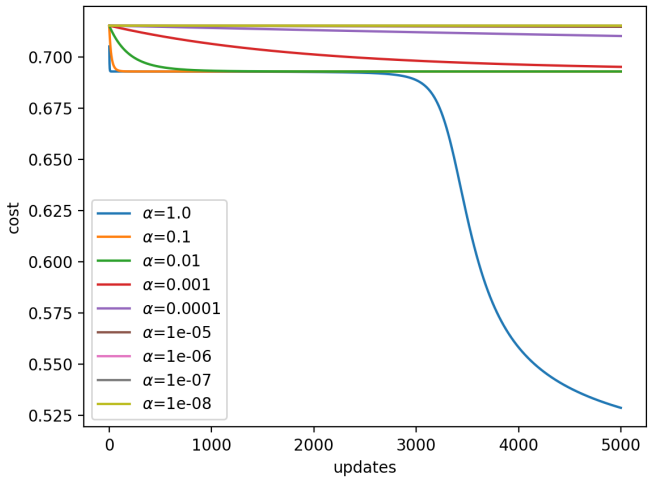

Figure 2: Model training performances for each hyperparameter α.

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)	80.0	75.0	71.91
Accuracy (test set)	49.9	74.1	70.6

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)	29.44	50.51	71.91
Accuracy (test set)	29.7	51.1	70.6

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

Figure 1 shows that The data is not separated in a single straight line. At least two lines are required to classify the entire dataset samples. In fact, the model with the best hyperparameter is not perfect not only the test dataset, but also training set. So, we are able to think that this is underfitting becase of the simplicity of the model. Therefore, to classify the whole dataset, we need to make more complex model.