Binary Classification with Logistic Regression

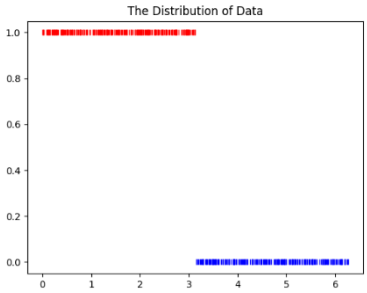
Department of Computer Software Engineering
College of Engineering, Hanyang University
Gaon Choi(최가온), Student ID: 2019009261

Introduction

The purpose of this experiment is to solve the problem based on the newly introduced dataset through the binary classification method designed last time. The dataset generation method is as follows.

- Random value between 0 and 360 is extracted from Uniform Distribution and determined as an x value.
- The value extracted from No.1 is passed into radian conversion function and then put into the sine function.
- If the sine value is positive, y is 1, otherwise 0(if it is negative).

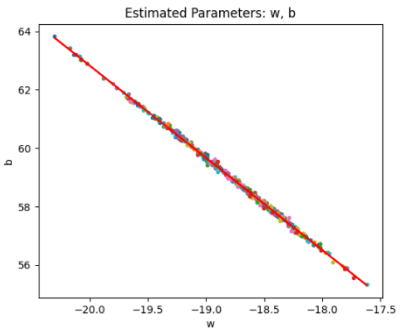
After creating datasets in the manner presented above, we visually examined their distribution. In the figure below, the data are expressed in different colors (+(1): red, -(0): blue) according to the y value. Judging from the distribution of data, it can be expected that the ideal step function will be the fitting curve that best describes the distribution of data.



Experiment

1. Estimated unknown function parameters W & b

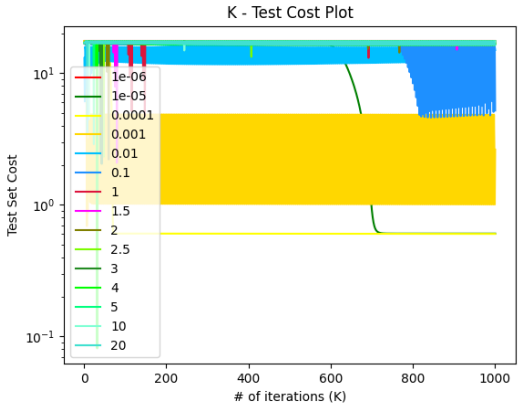
Both x and y of the datasets created in this experiment are one-dimensional. Thus, the dimension of W is also one dimension. Although described in Section 2, it was found that the higher the value of the learning rate, the better the performance gradually.(it is restricted to this experiment.) A total of 400 independent experiments were performed with m=10,000, n=1,000, k=1,000, and α=20 as the basic environment. When α=20, there was little change in validation cost when k=1,000 or higher, and the accuracy is nearly 100 in every experiment.

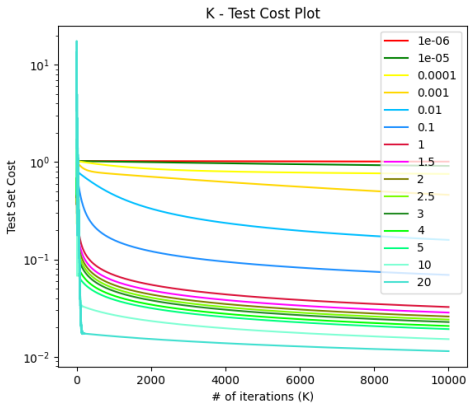


It was found that w and b had a negative correlation with each other. The red line is a trend line based on them.

2. Empirically determined (best) hyper parameter, α

When the range of x is [0, 360], the accuracy was found to be under 50% for all alpha values. It is assumed that this is a phenomenon in which the range of x is substantially wide compared to the range of y (range: $y \in \{0, 1\}$). When the range of x is normalized, converting to a radian value, this phenomenon was significantly improved, and almost all of the accuracy was ≥ 99%. If K = 100,000, the accuracy was 100% except when α=1e-6 and α=1e-5.





3. Accuracy

3-1. Accuracy comparison by the value of m (the size of train set) (α=0.01)

While keeping the value of n and k constant, the trend of accuracy when m is gradually increased is as follows.

	m=10, n=1000, K=5000	m=100, n=1000, K=5000	m=10000, n=1000, K=5000
Accuracy (with 'm' train samples)	100.00% (80.0%)	95.00% (47.0%)	96.16% (55.53%)
Accuracy (with 'n' test samples)	95.80% (50.5%)	95.89% (52.6%)	96.39% (53.2%)

※ ( ) : when x ∈ [0, 360]

3-2. Accuracy comparison by the value of K (the number of iterations) (α=1.0)

One hundred independent experiments were performed on three cases where the value of K was 20, 200, and 2000, respectively. For each experiment, the dataset was newly created and proceeded every time. The accuracy below is the average of the results derived from 100 experiments.

	m=10000, n=1000, K=10	m=10000, n=1000, K=100	m=10000, n=1000, K=5000
Accuracy (with 'm' train samples)	92.75% (49.70%)	97.33% (50.23%)	99.97% (53.84%)
Accuracy (with 'n' test samples)	92.20% (49.40%)	98.00% (50.60%)	100.00% (49.70%)

※ ( ) : when x ∈ [0, 360]

Discussion

Contrary to what was expected, the accuracy was lower, when x was not normalized(x ∈ [0, 360]). The first thing I did at this time was to visually display all the generated data. Once again, I realized that it is important to properly grasp the characteristics of data in advance as much as designing a model. The model itself is the same as the last time, but through the two experiments, I was able to understand the inherent principles of Logistic Regression.