Logistic Regression

TDung

October 2020

1 Introduction

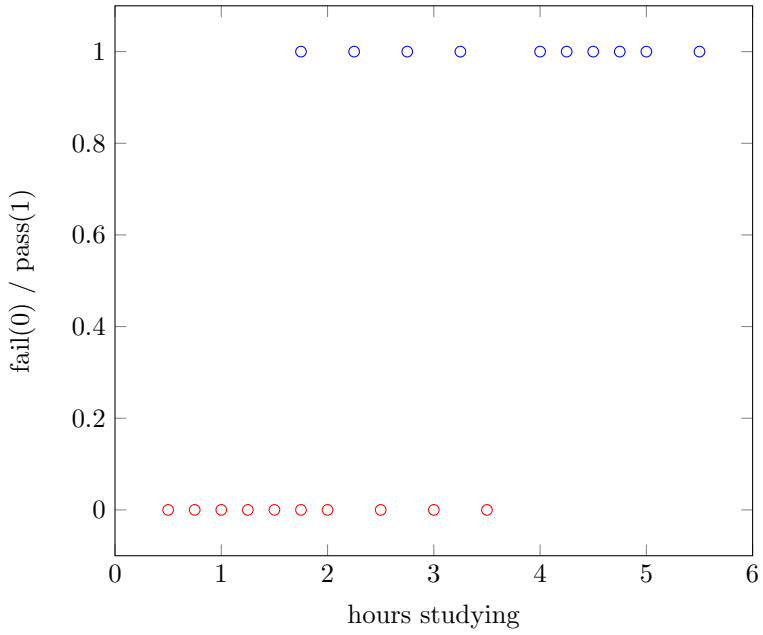
Regression methods have become popular in machine learning, data analysis, and statistics. We need to describe the relationship of the variable (response) Y and variables (explanatory variables) X which a data collected . Normally, we will have the output variables as discrete variables. Logistic Regression is the one.

2 Formula For Logistic Regression

First we have to know the difference of Linear Regression and Logistic Regression. That is, we have the outcome variable of the Logistic to be binary and dependent on the model's parameter choice and its assumptions. However, logistic regression will follow some basic rules used in linear regression. We will come with a small example as follows.
Here is a small data taken from `https://en.wikipedia.org/wiki/Logistic_regression`. It is a group of students who spend 0 to 6 hours reviewing for each student's test and test results after the test.

STT	X	Y	STT	X	Y
1	0.5	0	11	2.75	1
2	0.75	0	12	3	0
3	1	0	13	3.25	1
4	1.25	0	14	3.5	0
5	1.5	0	15	4	1
6	1.75	0	16	4.25	1
7	1.75	1	17	4.5	1
8	2	0	18	4.75	1
9	2.25	1	19	5	1
10	2.5	0	20	5.5	1

Figure 1.1: Example results for number of hours study



Here, if we pay close attention, we will see that if the number of study hours increases, the likelihood of that corresponding student passing the test will increase.

The first different point that we were concerned about the relationship between X and Y. One of the key to answering this question is the $E(Y|X)$, also called the expected value for each X In linear regression, we assume that E will be calculated by:

$$E(X|Y) = \beta_0 + \beta_1 x \tag{2.1}$$

The above expression is possible for all X of infinity for $E(X|Y)$.

$$\lim_{x \to +\infty} = +\infty \tag{2.2}$$

$$\lim_{x \to -\infty} = -\infty \tag{2.3}$$

For the data given above or the data related to logistic regression, the expected value must be greater than or equal to 0 and less than or equal to 1. However, a line $E(X|Y)$ is used above, the obtained values may be greater than 1 or less than 0 (eg with X = 20,...). With a logistic regression model of the data above, we need a curve where the lowest or highest value only progresses gradually towards 0 or 1. The curved is said to be "S-shape". This means that the model used has logistic distribution.
So we will replace the formula $E(Y|X)$ - the expected value of Y with X values when using logistic distribution. The alternative formula is:

$$p(x; w) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x)}} \tag{2.4}$$

The second difference is the conditional distribution of the logistic regression output value. In linear regression we have $Y = E + e$ (e is called error), e is the difference value of predicted values with the actual average value, usually e obey normal distribution with a mean value equal to 0. It is a constant variance after estimated based on given data. For logistic regression we have $y = p(x; w) + e$. When $y = 1$ we have $e = 1 - p(x; w)$ with probability $p(x; w)$ and $y = 0$ we have $e = -p(x; w)$ with probability $1 - p(x; w)$. Variance is equal to $p(x; w)(1 - p(x; w))$.

3 Calculating Model

To select a good model for logistic regression, we must estimate the beta parameters based on given data. For the linear regression model, we will estimate the beta parameter by using the derivative of the residual sum of square for each parameter to get the best model. However, using this way to approach the model of logistic regression is not possible. Therefore, we will use the likelihood function method to approach this model. The method of maximum likelihood of value for the unknown parameters that maximize the probability of the observed set of data. This function expresses the probability of the observed data as a function of the unknown parameters. The maximum likelihood estimators of the parameters are the values that maximize this function.

$$p(x_i; w)^{y_i}(1 - p(x_i; w))^{1-y_i} \tag{3.1}$$

It can be known that if the outcome variable equal 1, we will have $p(x_i; w)$ and $1 - p(x_i; w)$ for the outcome variable equal zero. Thus, likelihood function is product of the terms given 3.1:

$$l(\beta) = \prod_{i=1}^{M} p(x_i; w)^{y_i}(1 - p(x_i; w))^{1-y_i} \tag{3.2}$$

But it is easier mathematically to work with the log of equation 3.2 - called the log-likelihood:

$$L(\beta) = \ln(l(\beta)) = \sum_{i=1}^{M}(y_i \ln[p(x_i; w)] + (1 - y_i) \ln[1 - p(x_i; w)]) \tag{3.3}$$

To find the value of β that maximizes $L(\beta)$ we differentiate $L(\beta)$ with respect to β_0 and β_1 and set the resulting expressions equal to zero. These equations known as the likelihood function:

$$\sum [y_i - p(x_i; w)] = 0 \tag{3.4}$$

And

$$\sum x_i [y_i - p(x_i; w)] = 0 \tag{3.5}$$

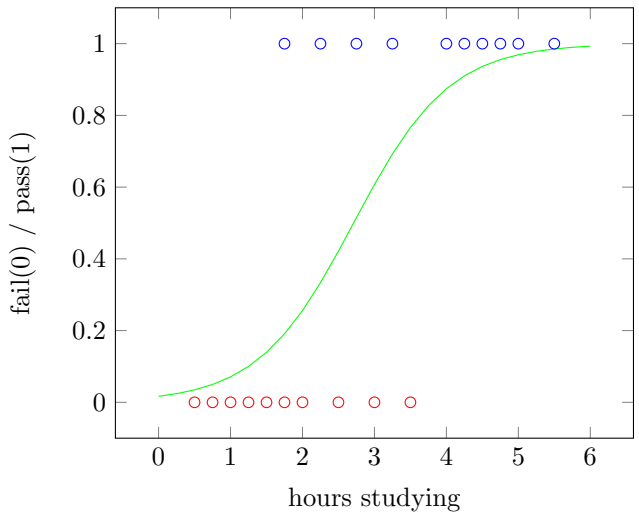
For logistic regression, the formula 3.4 and 3.5 is nonlinear with parameters of model. To estimate these parameters, we will some special method is mentioned in the last section.
With the method is calculated in last section, we have:

	Coeff	S.e	Z	Lower	Upper	p-value
β_0	-4.07723	-1.74428	-2.33748	-7.49596	-0.65850	0.019414
β_1	1.50445	0.62159	2.42031	0.28615	2.72274	0.015507

The maximum likelihood estimates of β_0 and β_1 are $\hat{\beta}_0 = -4.07723$ and $\hat{\beta}_1 = 1.50445$. The fitted values are given by the equation 2.4

$$p(x; b, w) = \frac{1}{1 + e^{-(-4.07723 + 1.50445\dot{x})}} \tag{3.6}$$

Figure 3.7: Simulate the curve of probability



4 Testing The Significance Of The Coefficients

The comparison of observed to predicted values using the likelihood function is based on the following expression:

$$D = -2\ln\left[\frac{\text{likelihood of the fitted model}}{\text{likelihood of the saturated model}}\right] \tag{4.1}$$

Using 3.3 and 4.1 becomes,

$$D = -2\sum_{i=1}^{M}\left[y_i \ln\left(\frac{\hat{p}_i}{y_i}\right) + (1-y_i)\ln\left(\frac{1-\hat{p}_i}{1-y_i}\right)\right] \tag{4.2}$$

The formula of D in 4.2 is called deviance for logistic regression (it has the same role of RSS - residual sum of square in linear regression).
If we try to calculate l(saturated model) bases on table described in section 2, we can see l(saturated model) $= 1$

$$l(\text{saturated model}) = \prod_{i=1}^{M} {y_i}^{y_i}(1-y_i)^{(1-y_i)} = 1.0 \tag{4.3}$$

Thus, we can set the formula for D is:

$$D = -2\ln(\text{likelihood of the fitted model}) \tag{4.4}$$

To calculate significance of an independent variable, we usually compare D with and without the independent variable in the equation.

$$G = D(\text{without the variable}) - D(\text{with the variable}) \tag{4.5}$$

Or it can be written:

$$G = -2\ln\left(\frac{\text{likelihood without the variable}}{\text{likelihood with the variable}}\right) \tag{4.6}$$

Let denote $M_1 = \sum_{i=1}^{M} y_i$ and $M_0 = \sum_{i=1}^{M}(1 - y_i)$, from 4.6, we have:

$$G = -2 \ln \left[\frac{\left(\frac{M_1}{M}\right)^{M_1} \left(\frac{M_0}{M}\right)^{M_0}}{\prod_{i=1}^{M} \hat{p}_i^{y_i} (1 - \hat{p}_i)^{1-\hat{p}_i}} \right] \tag{4.7}$$

For the data given, we can calculate the value of $G = 13.66613029$

5 Confidence Interval Estimation

The standard error of β_0 is SE(β_0). Hence a (1 - α) x 100 % confidence interval for β_0 is the set of points β_0 in the interval

$$\beta_0 - z_{1-\alpha/2} SE(\beta_0) \leq \beta_0 \leq \beta_0 + z_{1-\alpha/2} SE(\beta_0) \tag{5.1}$$

For β_1

$$\beta_1 - z_{1-\alpha/2} SE(\beta_1) \leq \beta_1 \leq \beta_1 + z_{1-\alpha/2} SE(\beta_1) \tag{5.2}$$

Where $z_{1-\alpha/2}$ is the upper $100\,(1 - \alpha/2)\,\%$ point from the standard normal distribution.
For 95% confident interval of β_0 we have z = 1.96 (Find it in normal distribution table):

$$\begin{aligned} -4.07723 - (1.96 * -1.74428) \leq & \beta_0 \leq -4.07723 + (1.96 * -1.74428) \\ -4.07723 - 3.4187888 \leq & \beta_0 \leq -4.07723 + 3.4187888 \\ -7.4960188 \leq & \beta_0 \leq -0.6584412 \end{aligned} \tag{5.3}$$

Similar for β_1 we have $0.28615 \leq \beta_1 \leq 2.722742$.

6 Estimate Parameters By Gradient Descent

To estimate parameters of model, we will use Stochastic Gradient Descent. As mentioned on section 3, the formula of log likelihood function is:

$$L(\beta) = \ln(l(\beta)) = \sum_{i=1}^{M} \left(y_i \ln[p(x_i; w)] + (1 - y_i) \ln[1 - p(x_i; w)] \right) \tag{6.1}$$

But here, we will put "-1" in front of the right equation and make it be loss function. It called negative log likelihood function.

$$L(\beta) = \ln(l(\beta)) = -\sum_{i=1}^{M} \left(y_i \ln[p(x_i; w)] + (1 - y_i) \ln[1 - p(x_i; w)] \right) \tag{6.2}$$

For any pair (x_i, y_i) in data set, we have:

$$J(w; x_i, y_i) = -\left(y_i \ln\left[p(x_i; w)\right] + (1 - y_i) \ln\left[1 - p(x_i; w)\right] \right) \tag{6.3}$$

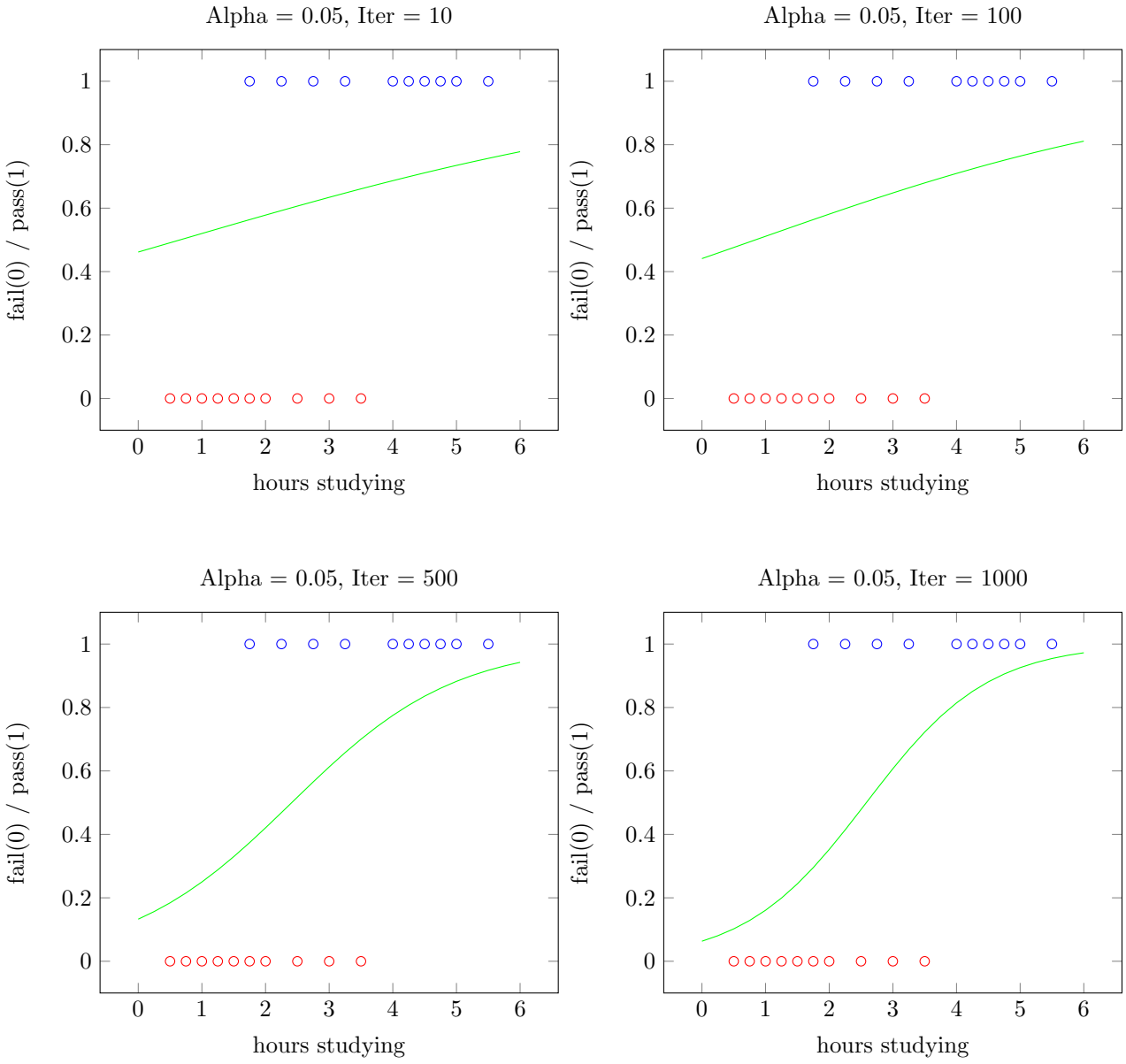
With derivatives

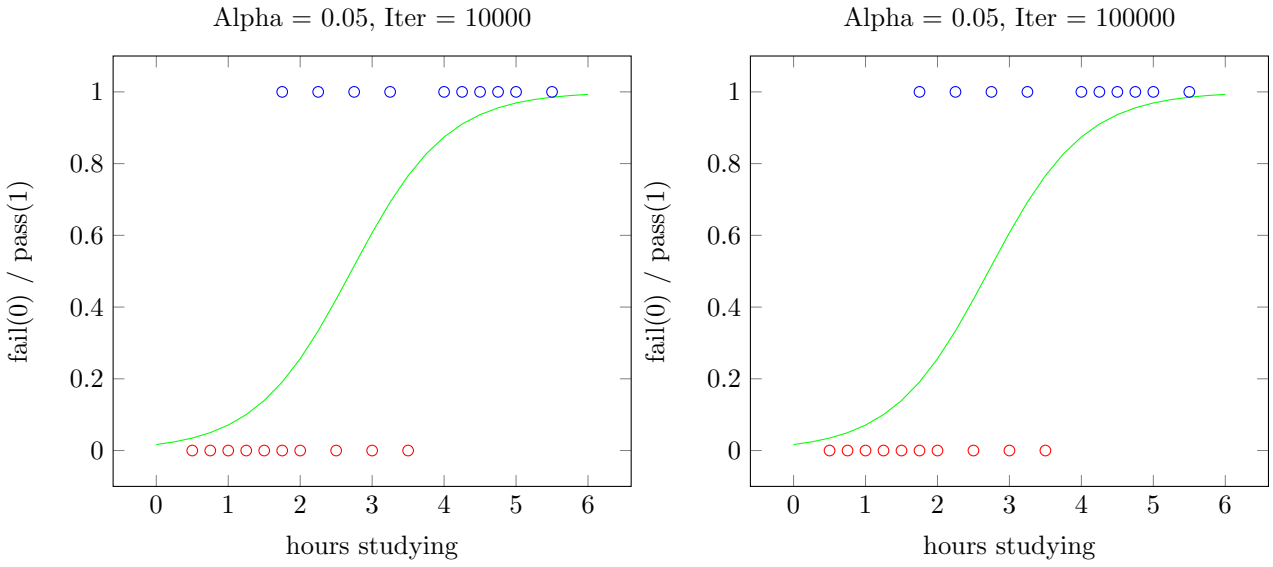
$$\begin{aligned} \frac{\partial J(w; x_i, y_i)}{\partial w} &= -\left(\frac{y_i}{p(x_i; w)} - \frac{1 - y_i}{1 - p(x_i; w)} \right) \frac{\partial p(x_i; w)}{\partial w} \\ &= \left(p(x_i; w) - y_i \right) x_i \end{aligned} \tag{6.4}$$

The formula to update W for logistic regression is:

$$w = w + \eta (y_i - p(x_i; w)) x_i \tag{6.5}$$

In here, we will set $\eta = 0.05$ (choosing learning rate η is very importance for a model).


Alpha = 0.05, Iter = 10
Alpha = 0.05, Iter = 100
Alpha = 0.05, Iter = 500
Alpha = 0.05, Iter = 1000
fail(0) / pass(1)
hours studying
0
0.2
0.4
0.6
0.8
1
0
1
2
3
4
5
6



7 Source Code For Estimating Parameters

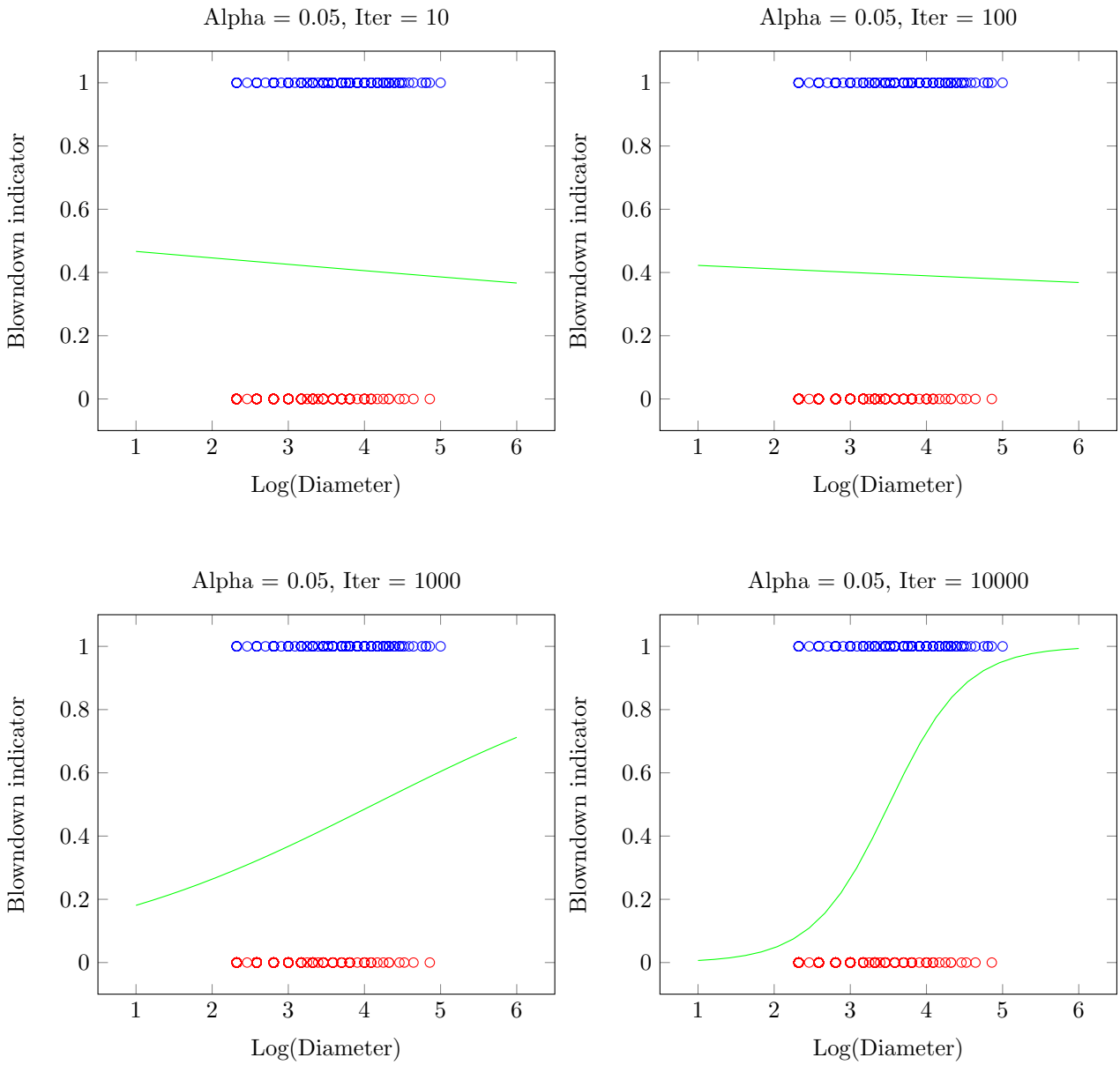
```
import numpy as np
import pandas as pd
import time


START_TIME = time.time()
# ----------------------------------------------------------------------

DATA_FILE_NAME = 'RESULT_EXAM_STUDY.csv'
ITERATION = 100000
ALPHA = 0.05


def load_data():
    df = pd.read_csv(DATA_FILE_NAME)
    X = df.values[:, 0: 1]
    y = df.values[:, 1]
    m = y.size

    X = np.concatenate((np.ones((m, 1)), X.reshape(-1, 1)), axis=1)
    return X, y


def sigmoid(z):
    return 1 / (1 + np.exp(-z))


def gradient_descent(X, y, theta, alpha, inter):
    m = y.size
    for _ in range(0, ITERATION):
        theta = theta - ALPHA * (np.dot(x.T, sigmoid(np.dot(x, theta)) - y)) / m
    return theta


x, y = load_data()
theta = gradient_descent(x, y, np.zeros(x.shape[1], ), ALPHA, ITERATION)
```

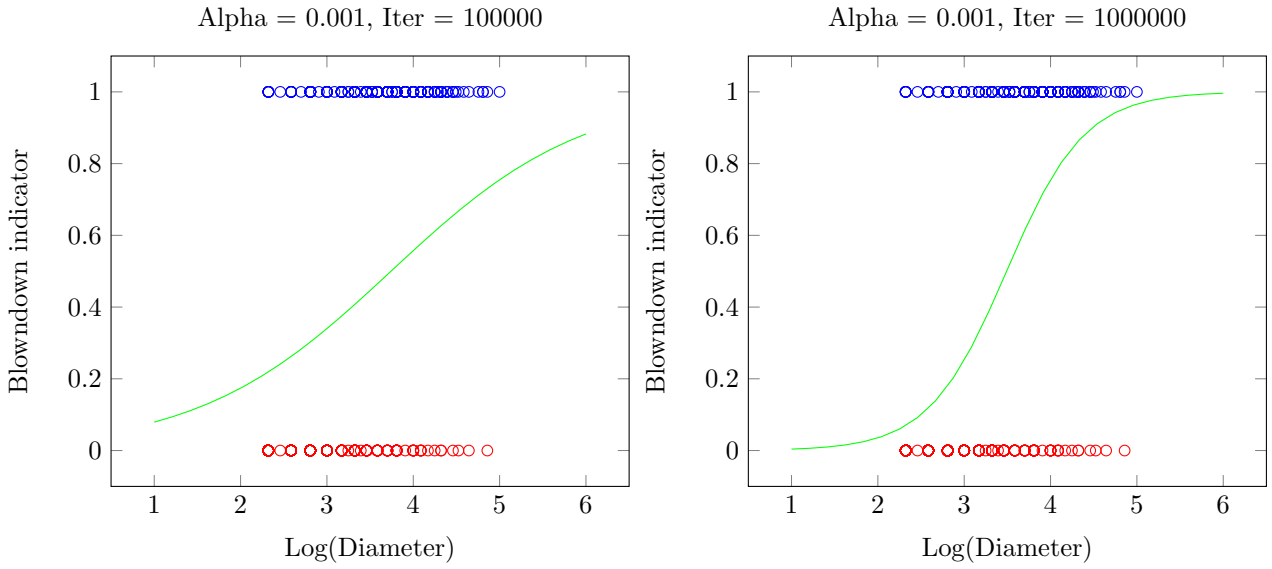
```
        print(theta)

        # ------------------------------------------------------------------

        END_TIME = time.time()
        print(f"Runtime of the program is {END_TIME - START_TIME}")

```

Listing 1: Source code for sample 1

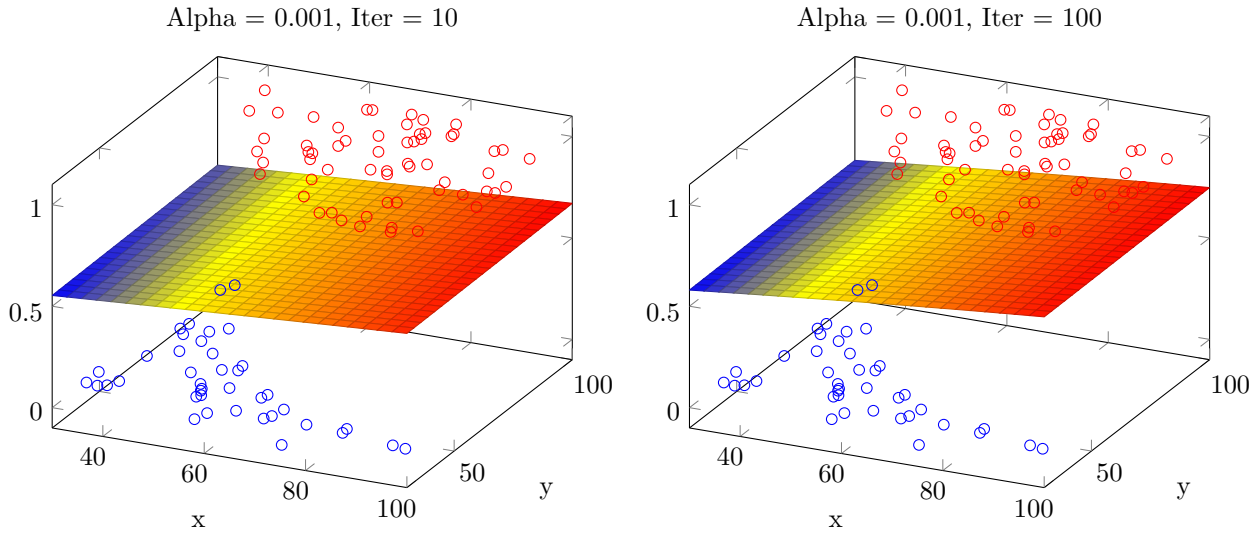
8 Practice Model In More Data Set

First, we will try model with new data set is blowBF.txt (M = 659) given.



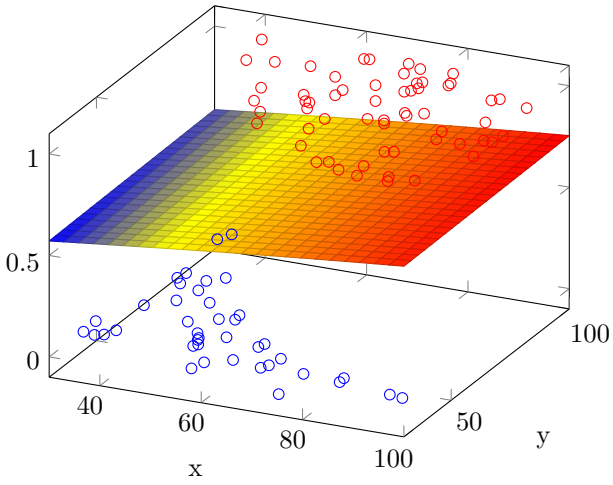


If we look at the last two graphs, we can see that with the coefficient alpha = 0.001, with the number of training times equal to 100000 (10 times the alpha of 0.05 to find the almost optimal solution), we almost no optimal solution yet. And to get a perfect curve as shown in graph 4^{th} (the second row on the right), we need 1000000 training times.
Second, we will come up with data that has two parameters need to estimate.

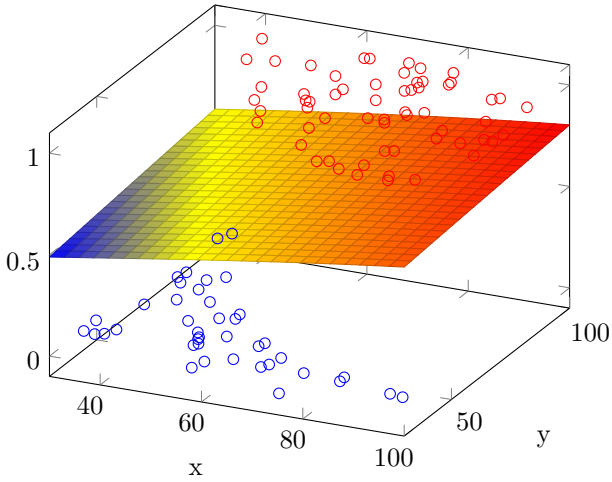


2

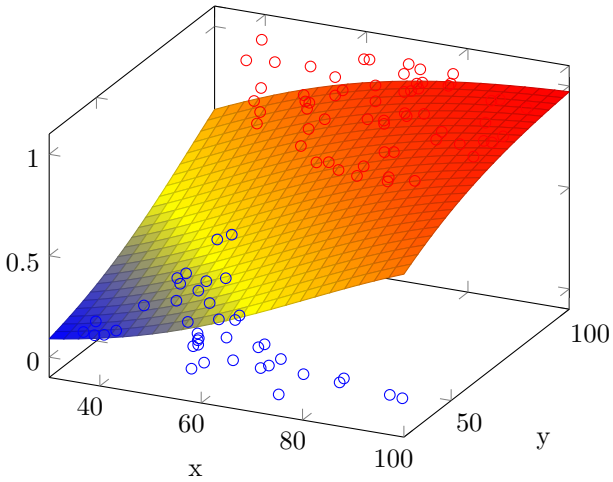
Alpha = 0.001, Iter = 1000



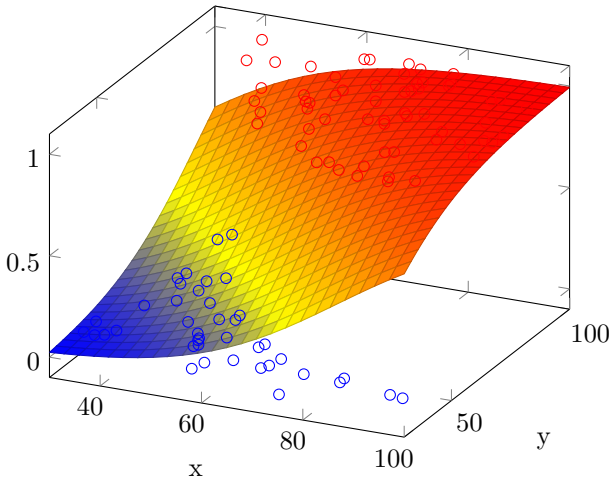
Alpha = 0.001, Iter = 10000

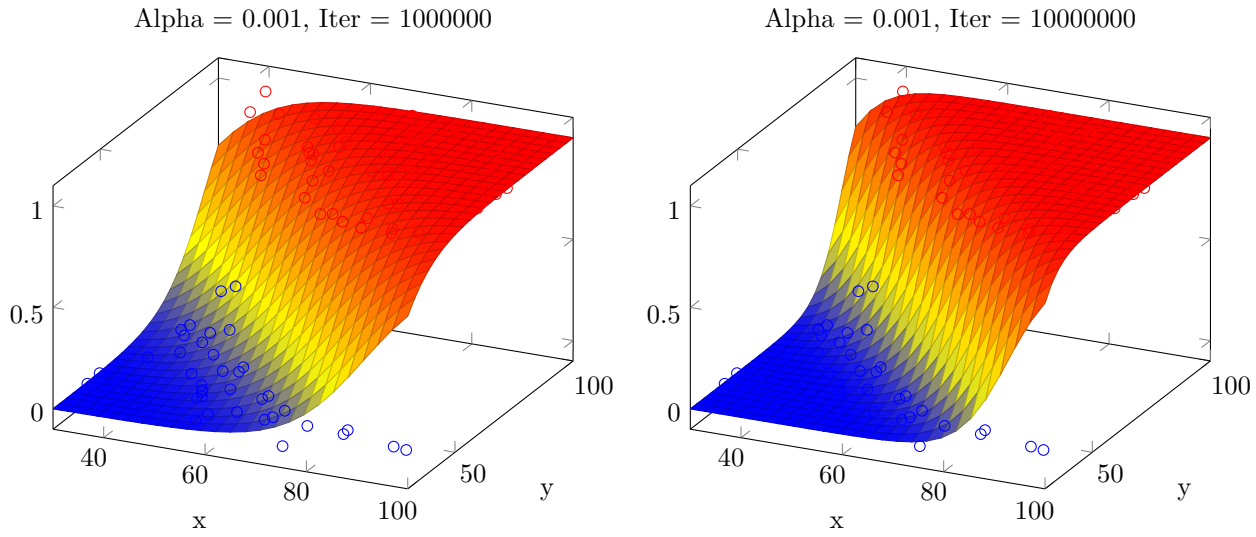


Alpha = 0.001, Iter = 100000



Alpha = 0.001, Iter = 200000





If we pay attention into the last two graphs, we can see the curvature is down very quickly. This means that the different probability between the two 0 and 1 becomes more clearly. This phenomenon is called over-fitting and can be minimized with the following formula:

$$w = w + \eta(y_i - p(x_i; w))x_i + \frac{\lambda}{m}w \tag{8.1}$$

Where lambda is a constant. We add the formula $\frac{\lambda}{m}w$ as a penalty, the bigger the error, the higher the penalty and the smaller the error, the lower the penalty.

9 References

https://en.wikipedia.org/wiki/Logistic_regression
https://machinelearningcoban.com/2017/01/12/gradientdescent/
https://machinelearningcoban.com/2017/01/27/logisticregression/
https://www.kaggle.com/msjaiclub/2classclassification/version/1
http://users.stat.umn.edu/~sandy/alr3ed/website/data/blowdown.txt
https://www.math.arizona.edu/~rsims/ma464/standardnormaltable.pdf