ESPCI PARIS | PSL

http://espci.psl.eu

Travaux Pratiques Statistique Appliquée

Logistic Regression

Jinxin HU* Aline Brez RAMIREZ* Aline Kneubl ANDREUSSI*

Abstract

Due to the growing volume of computational data that is produced and stored, data techniques become increasingly necessary. There are countless examples of the application of logistic regression in the financial, environmental and epidemiological areas because it's a strong tool for analyzing categorical response data, which makes it possible to estimate the probability of the occurrence of events.The aim of this work is to implement an automatic classification chain based on logistic regression: inference of the model, learning of its parameters, experimentation and extension.

Keywords: Logistic Regression; Probability; Binary Classification

Supplementary Material: Downloadable Code for data set 1 and 2 — Code for data set 3

*jinxin.hu@espci.fr, alinebraz09.ab@gmail.com, alineandreussi@gmail.com, *Élève ingénieur Promotion 138, ESPCI Paris*

1. Introduction

Machine learning is a method of data analysis that automates the construction of analytical models. It is a branch of artificial intelligence based on the idea that systems can learn from data, identify patterns and make decisions with minimal human intervention. Most industries working with large amounts of data have recognized the value of machine learning technology. Industries working with financial services, governments, health care operators, marketing and sales, oil and gas and transport are some examples.[3]

There are different methods of learning. Supervised learning algorithms are trained by using labeled examples, such as an entry where the desired exit is known. Supervised learning is commonly used in applications where historical data predict likely future events. For example, it can anticipate when credit card transactions are susceptible to be a fraud.

Unsupervised learning is used in data that has no historical labels. The "right answer" is not informed to the system. For example, it can identify customer segments with similar attributes that can then be treated similarly in marketing campaigns. Popular techniques include self-organizable maps, proximity mapping, k-means grouping, and decomposition into unique values. These algorithms are also used to segment text topics, recommend items and identify discrepant points in the dataset.

Semi-supervised learning is used for the same applications as supervised learning. But this one handles both labelled and unlabelled data for training. This type of learning can be employed with methods such as classification, regression and prediction. Basic examples include the identification of a person's face on a webcam.

Finally, booster learning is commonly used in robotics, gaming and navigation. The goal of this type of learning is to determine which actions yield the greatest rewards through 'trial and error' tests.

In this work, supervised learning algorithms will be used. The logistic regression is a type of regression that predicts whether something is true or false (binary data classification), instead of predicting something continuous, as "size". It is based on a probabilistic

criterion.

Given a dataset D, since the decision rule is to define $f_\theta : x \to y$ where θ are the parameters of f, the main task for machine learning is to learn θ and the learning finds the good value for θ by minimizing a Loss Function $L(\theta, D)$.

$$Let \quad D = (x_{(i)}, c_{(i)})_{i=1}^{n} = (X, \tilde{C}) \tag{1}$$

and a logistic regression model, $\theta = (w_0, W)$.

$$\begin{aligned} L(\theta, D) &= -log(P(\tilde{C}|X, \theta)) \\ &= -(\sum_{i=1}^{n} (c_{(i)} log(y_{(i)}) + (1 - c_{(i)}) log(1 - y_{(i)}) \end{aligned} \tag{2}$$

with: $y_{(i)} = \sigma(w_0 + \mathbf{W}^t \mathbf{x}_{(i)}) = \frac{1}{1+e^{-(w_0 + \mathbf{W}^t \mathbf{x}_{(i)})}}$.

Since the function is convex, it can be easily minimized by Stochastic gradient descent (SDG).

The aim of this work is to implement an automatic classification chain based on logistic regression: inference of the model, learning to adjust its parameters, experimentation and extension.

2. Description of the statistical tools

Logistic regression is a statistical technique that aims to produce a model, based on a set of observations, which allows the prediction of values taken by a dependent variable as a function of other independent variables.

For this study we consider a dependent variable C, which follows a Bernoulli distribution, assuming a value of 0 or 1, which are responsible for determining the class that the points belong to. It's important to remember that when the regression follows a distribution, it is necessary to connect all independent variables to the same type of distribution as well.

The function responsible for making this connection is called the sigmoid function or the logistic function. It's a Bernoulli distribution that describes the probability that the random variable belongs to class 1, or event of interest.

With the data set (X, C), where X represents the model inputs and C the class that is equal to 0 or 1, and the parameters w and w_0 it was possible to calculate the output of the logistic regression model in the training data set (X). The result is a vector, "a", of size N, with N probabilities of belonging to class 1, and N has the same size of vector C, that can be described as a = $(w_0 + \mathbf{W}^t \mathbf{x}_{(i)})$.

From these data it was possible to describe the sigmoid function, represented by sigma in the expression, which generates the probability that the points belong to class 1, as demonstrated in the introduction.

Considering the loss function as the function that takes as input the predicted value corresponding to the real value of the data. This function will be used mainly to check the training progress. Therefore, we can describe it as represented in equation 1, whose first term in the sum is the probabilities of belonging to class 1 and the second term the probabilities of belonging to class 0. To calculate the best parameters of the vector "a", we try to minimize the loss function by calculating its gradient in relation to the model parameters such as equation 3:

$$w_i = w_i - \eta \partial L / \partial w_i \tag{3}$$

where η represents the learning rate and $\partial L / \partial w$ is the partial derivative of loss function.

In this part we are going to calculate the derivative of L as a function of k values ranging from 0 to 2. The values of w refer to the linear and angular coefficients that will be used to create an adjustment in the values of the final parameters of the equation. To calculate this derivative we introduce the following function:

$$\frac{\partial L}{\partial wk} = \frac{1}{n} \sum (y_i - c_i) x_{ki} \tag{4}$$

with, $a_i = w_0 + w x_i$.

This function will be of great help later to classify the different values of x_i and to interpret the results.

For this last part of the regression, it's important to optimize the result in order to find an ideal value between calculation time and precision. For example, if we choose a very small learning rate the value will be accurate but the calculation time is not viable (small steps), and also if we choose a very large rate the calculation time will decrease but the problem will have a high probability of divergence due to large oscillation between values (large steps).

Performing a good optimization of the problem, we can say in general that it's a simple model, there is little risk of over-learning and the results tend to have a good generalization power.

3. Analysis of the results

3.1 Data set 1

As shown in Figure 1, a) shows the distribution of the data set given, in which, red stands for C=0 and green stands for C=1. In the figure b, the blue line stands for the initial random value of w and w_0. Using gradient decent method, an optimized value of w and w_0 can be generated from enough times of iteration. As a result,

a small displacement can be found between the orange line (new) and blue line, which to show schematic effect of one iteration.

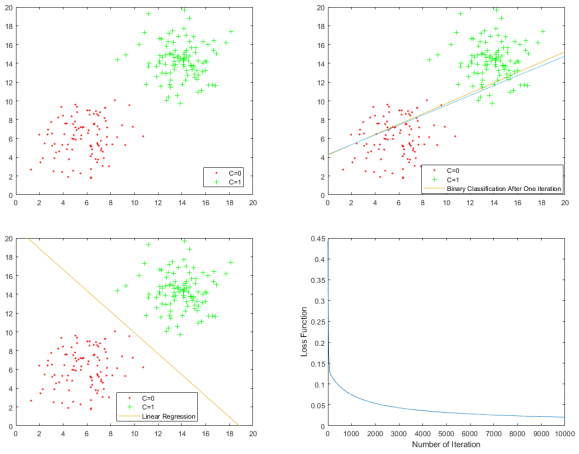

Figure 1. a)Distribution of the data set 1 given. b) Schematic diagram of linear regression after one iteration. c)Linear regression diagram after 10000 iterations. d)Loss function over number of iteration

In figure c), after 10000 times of iteration of linear regression, a distinct and robust binary classification can be observed. Figure d shows how loss function varies with each time of iteration. Mathematically, it is also accessible to reach the bottom through the other bunch of this curve by gradient decent method.

3.2 Data set 2

Similar with data set 1, as shown in Figure 2, however, there is no linear boundary between C=0 and C=1. Using linear regression method same as in data set 1, a binary classification line can be obtained in Figure b). However, as the figure shows, this linear regression is not robust at all, and the error rate at last would not become zero as there are red dots and green cross in both sides of this line. This just shows the limitation of linear regression for solving such problems.

As Figure 2 c) shows, by zooming it, an oscillation in loss function can be observed. It is just the case when the learning rate is too large (lr=0.08) for gradient decent method, the value of loss function will oscillate between the bi-branch of parabola each time of iteration, and reach its bottom at last, as shown in Figure3 a).

However, when a smaller learning rate value is chosen (lr=0.01), similar regression line can be drawn in Figure 3 b), while there is no oscillation in the loss function and error rate over each iteration.

In a conclusion, linear regression method is not appropriate for those data set without linear boundary. Multiple layer of neural networks would be more suitable for this kind of problems.

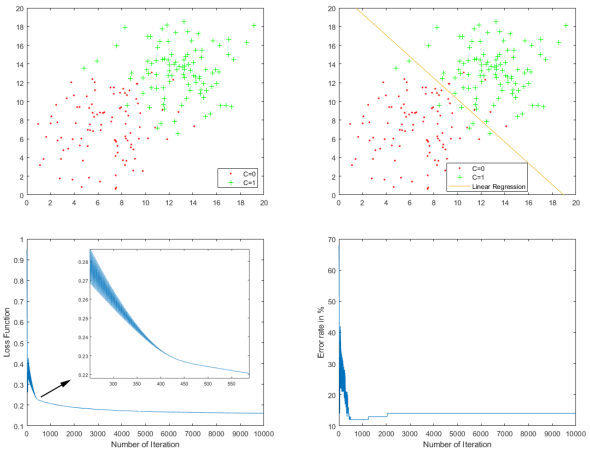

Figure 2. a)Distribution of the data set 2. b)Linear regression diagram after 10000 iterations. c)Loss function changes over number of iteration. d)Error rate changes over number of iteration.

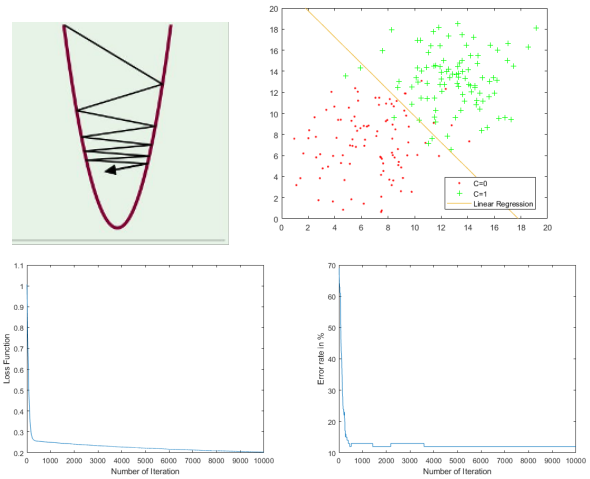

Figure 3. a)Loss function in parabola.[1] b)Linear regression result in a low learning rate(lr=0.01). c)Loss function changes over number of iteration in a low learning rate. d)Error rate changes over number of iteration in a low learning rate.

3.3 Data set 3

Completely different with data set 1 and 2, red dots are surrounded by green cross in data set 3 as shown in Figure 4 a). It is obvious to find the boundary is an oval. Therefore, in this case, w can be written as $\mathbf{w} = (w_1\ w_2\ w_3\ w_4\ w_5)^T$; and X can be written as $\mathbf{x} = (x_1\ x_2\ x_1x_2\ x_1^2\ x_2^2)^T$ Therefore, we got equation 5:

$$w_0 + w_1x_1 + w_2x_2 + w_3x_1x_2 + w_4x_1^2 + w_5x_2^2 = 0 \quad (5)$$

With the same principle, however in 5 dimensions, a logistic regression can be acquired as is shown in Figure 4 b). As the orange line shows a clear boundary for this binary classification problem. In Figure 4 c), it shows the loss function reach the minimum as in the previous 2 data sets. In figure d), the error rate reaches to 0 at last.

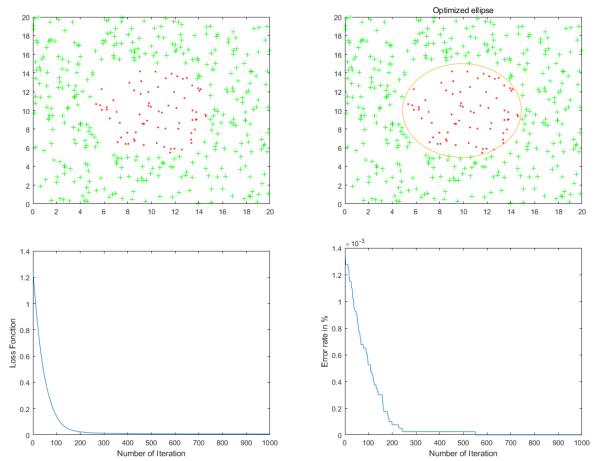

Figure 4. a)Distribution of the data set 3. b)Linear regression diagram after 1000 iterations. c)Loss function changes over number of iteration. d)Error rate changes over number of iteration.

3.4 Norm of w+w$_0$

Additionally, a further explanation on the norm of $w_0 + w$ can be explored. As shown in Figure 5 a) and b), the norm of $w_0 + w$ in data set 2 and 3 are displayed separately. When w is in 2 dimension, the norm of w continue to grow with the number of iteration going, while the norm of w in 5 dimension reaches a certain limit and maintain that value for the rest of iteration.

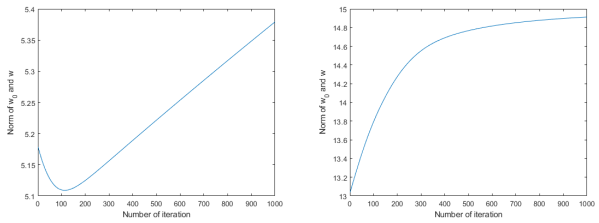

Figure 5. a)Norm of $w_0 + w$ in data set 2 over 1000 time of iteration. b)a)Norm of $w_0 + w$ in data set 3 over 1000 time of iteration.

Mathematically speaking, $||w_0 + w||$ in someway stands for the distance between origin and line $w_0 + w^T x = 0$ for linear regression as shown in Figure 6.

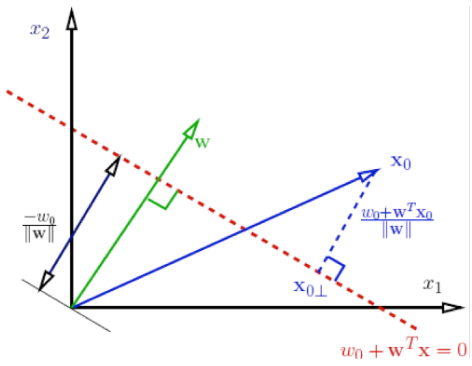

Figure 6. Schematic diagram of Mathematical meaning of w_0 and $||w||$ [1]

4. Conclusion

This practice could contribute to our statistical knowledge by using logistic regression algorithms to do binary classification and to estimate the probability of the occurrence of dual consequence events. What's more, mastering the mathematical principle makes us able to start a new project even from scratch, no matter it is a linear of circular problem.

Additionally, through this practice we gradually logistic regression is a method widely used, for example, in the development of credit analysis models and autopilot cars in a single neural network in machine learning. It is interesting to note that logistic regression, just as any other computational method, has its peculiarities and therefore, there is no method that is for any and all purposes.[2]

Acknowledgements

We would like to thank professor Isabelle Rivals for her wonderful lectures in statics. And special thanks to professor Alexandre Allauzen for his patient guidance in the TP of logistic regression. This report would not have been finished without their support.

References

[1] ALLAUZEN, A. *STAP Slides on Régression logistique*. Moodle, 2020. Available at: https://moodle.espci.fr/pluginfile.php/7507/mod_resource/content/1/stap_reglog_espci.pdf.

[2] DUCHI, J. *CS229 Supplemental Lecture notes*. September 2018. Available at: http://cs229.stanford.edu/extra-notes/loss-functions.pdf.

[3] NG, A. *Supervised Learning cheatsheet*. September 2018. Available at: https://stanford.edu/~shervine/teaching/cs-229/cheatsheet-supervised-learning#introduction.