Scientific Research Publishing

Open Journal of Social Sciences, 2019, 7, 126-132
https://www.scirp.org/journal/jss
ISSN Online: 2327-5960
ISSN Print: 2327-5952

A Preliminary Study on Parkinson's Disease with Regularized Logistic Regression Method

Jophy Lin

Arcadia, CA, USA
Email: jophylinyi0429@gmial.com

How to cite this paper: Lin, J. (2019) A Preliminary Study on Parkinson's Disease with Regularized Logistic Regression Method. *Open Journal of Social Sciences*, **7**, 126-132.
https://doi.org/10.4236/jss.2019.711010

Received: October 18, 2019
Accepted: November 17, 2019
Published: November 20, 2019

Abstract

Parkinson's disease is a long-term degenerative disorder of the central nervous system. In this paper, we have used gradient descent method with regularized logistic regression to give suggestions on the possible reason for the rise of this disease. Results with different regularization are presented and compared. We hope this could open a new way for studying complex diseases using data methods.

Keywords

Mental Depression, Parkinson's Disease, Regularized Logistic Regression, Gradient Descent

1. Introduction

Parkinson is a very common disease around the world. It has been known from ancient times. Parkinson has another name "Kampavata" in the ancient Indian medical system. But in a Western country, we call it "Parkinson". Parkinson formed by nerve cell damage in the brain causes dopamine levels to drop, after dopamine decreasing in our body system, leading to tremor. Tremor often occurs in one hand; it looks like shaking hardly and frequently especially in finger, hand or foot. Other early signs of Parkinson's disease are voice changes, uncontrollable movements during sleep, unbalance of handwriting, and limb stiffness or slow movement etc.

Here, we try a new method to investigate the reason that causes Parkinson by data analysis with regularized logistic regression method (this method will be mentioned later). There are certain factors which cause Parkinson. The first is environment. Environmental factors include the relationships with a person who you contact with or living condition. Second, genetics plays a role in Parkinson's

diseases with some of the patients' family history. Third, Parkinson's diseases are also affected by individual personality. A very common personality characteristic is being ambitious and rigid. Many experts studying on Parkinson's disease discovered that Parkinson patient's wife or husband also develop a similar symptom. The person who has Parkinson, showed some personality traits long before they develop the disease.

2. Background

In the advent of global aging, the rapid growing of elderly populations results in a rising number of PD patients, especially in China. According to the statistical figure "Prevalence of Parkinson's disease in China". The rates for the whole age group were 8 - 18/100,000 years, 50/100,000 years for over 65 years, 150/100,000 years for over 75 years, 400/100,000 years for over 85 years [1]. According to the cumulative incidence of age, the risk of Parkinson's disease among 60-year-olds at age 80 is about 2.5%. These data were derived from other studies, and we found that the population with Parkinson's disease was predominantly elderly. However, we still need to collect data from different countries to prove that Parkinson's disease mostly present in elderly patients. Another data shows that approximately 1 million Americans currently have the disease. The incidence of PD in the US is approximately 20 cases per 100,000 people per year (60,000 per year), with the mean age of onset close to 60 years. The prevalence of PD is reported to be approximately 1% in people 60 years of age and older and increases to 1% to 3% in the 80-plus age group [2] [3] [4]. With the serious situation of Parkinson's disease, methods to cure or control the disease are highly desired. Here, we employ the logistic regression with regularization to analyze the potential cause of Parkinson's disease. A decision boundary is given based on provided data. More complicated regression can be generalized.

3. Logistic Regression Analysis of Parkinson's Disease

A process that preceding Parkinson's disease in order to distinguish healthy people and patient. The purpose to utilize each data point from the collection document was trying to show the division of two types of situation. How does this data help us analyze Parkinson—people in Parkinson are hard to segregate while only appears on its superficial if thousand people mixing in a circle then we need to define their separation with a standard. Here are some steps that explain how to plot a decision boundary for Parkinson's disease.

Regression is a statistical modeling method. Regression analysis is a process estimating the relationships among variables [5] [6] [7] [8]. A measure of the relation is between the mean value of one variable and the corresponding values of other variables. For the initial steps, it is necessary to set up two values for helping load data from the Parkinson's data table such as x and y. Both represent the data in direction order likes from where to the end. Also label values vertically and horizontally such as x label ("Measure 1") and y label ("Measure 2"). In the

last step of the initial part, the last one is specified in plot order which means given specific and different standards like positive or negative. The reason to set up the direction to discover the overall Parkinson group being easy to recognized the patients by the plot that we can use in the actual psychological experiment. For example, $x_1 + x_2 = 10$, predict $y = 1$ if $x_1 + x_2 > 10$, and $y = 0$ if $x_1 + x_2 < 10$. Legend 1 = legend("y = 1", "y = 0") here using 1 and 0 to represent the healthy person and Parkinson patients (see Figure 1). The steps here are to establish the foundation of a map a feature. To develop the idea of Parkinson's diseases in many complicated factors. In logistic regression utilized sigmoid function because sigmoid can create a perfect curve line from negative infinity to positive infinity such as (0,1). Notice Equation (1) is sigmoid function or logistic function.

$$h(x) = 1/\left[1+\exp(-x\theta)\right] = g(x\theta) = g(Z) = 1/\left[1+\exp(-Z)\right] \quad (1)$$

Logistic regression has benefits that come out appropriate S-shaped curve. Sigmoid function to return a probability value that makes easier to map the discrete variables. Basically what the logistic function is doing is taking the log of the odds of an event. The odds of an event are the probability divided by the probability of the event not occurring. The purpose right here is modeling people with Parkinson or non-Parkinson as a linear combination of various features.

4. Predict the Size and Accuracy of the Feature

The last steps provided a basic structure of the graph. The next part(Part one) was matching the size of the feature. So set the value of x by seeking its size

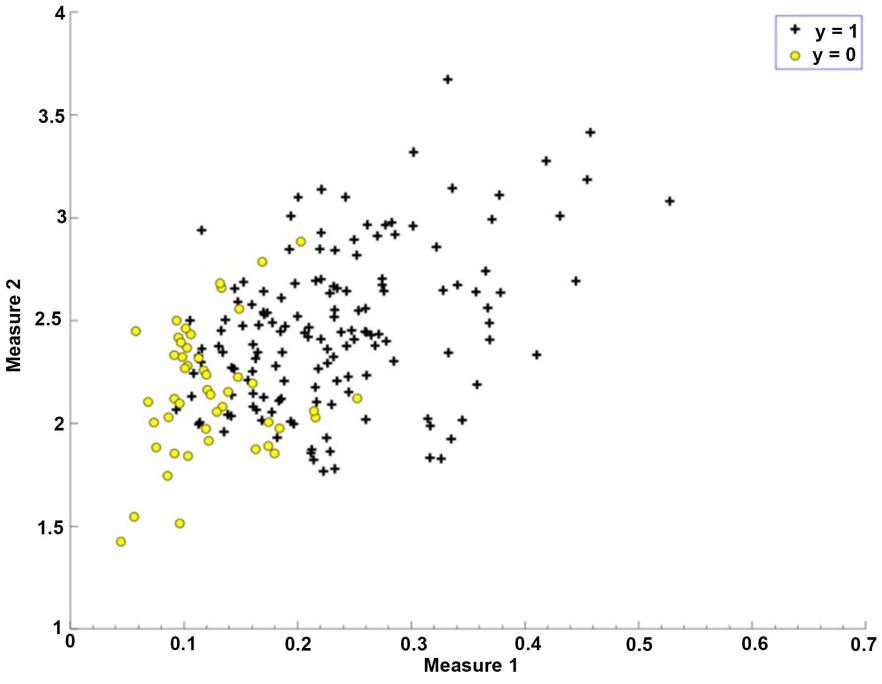

Figure 1. Patients with (y = 1) and without (y = 0) Parkinson's disease. Measure 1 and Measure 2 are two factors that potentially cause Parkinson's disease.

which turns to size (x). Then initialize fitting parameters. First, found the size of initial_theta = zeros (size(x, 2), 1). Because of given some specific area on which column and row. Second, after we found the initial_theta then set regularization parameter lambda, lambda = 0.1. By using lambda to create an exact estimation, when lambda smaller or more appropriate and it will suitable for plotting features, in other words, it won't cause overfitting or failed to the training data. For example, if lambda extremely large value, then the line—decision boundary,it would be a flat line. For plotting need to use Equation (2) is gradient descent,

$$\theta_j := \theta_j - \alpha\left(\partial/\partial\theta_j\right) J\left(\theta_0, \theta_1\right), \quad \text{with} \quad j = 0,1 \tag{2}$$

The purpose of using gradient descent is finding the minimum of a function of multiple variables in order to figure out the data carry Parkinson's disease. Then, compute and display initial cost and gradient for regularized logistic regression. Also, Equation (3) is cost function with regularization, when theta or coefficient small and reduce its effect of hypothesis function in order to get simpler results.

$$J\left(\theta_0, \theta_1\right) = \frac{1}{2m}\sum_{i=1}^{m}\left[h_\theta\left(x^{(i)}\right) - y^{(i)}\right]^2 \tag{3}$$

theta needs to be small that can minimizing the function and close to 0 which get simple hypothesis. Cost function including [cost, grad] = cost FunctionReg (initial_theta, x, y, lambda). The reason we need to use cost function, lambda, and matrix in this part to label the points in the graph but giving a range. Then classify Parkinson patient through the points (see Figure 2).

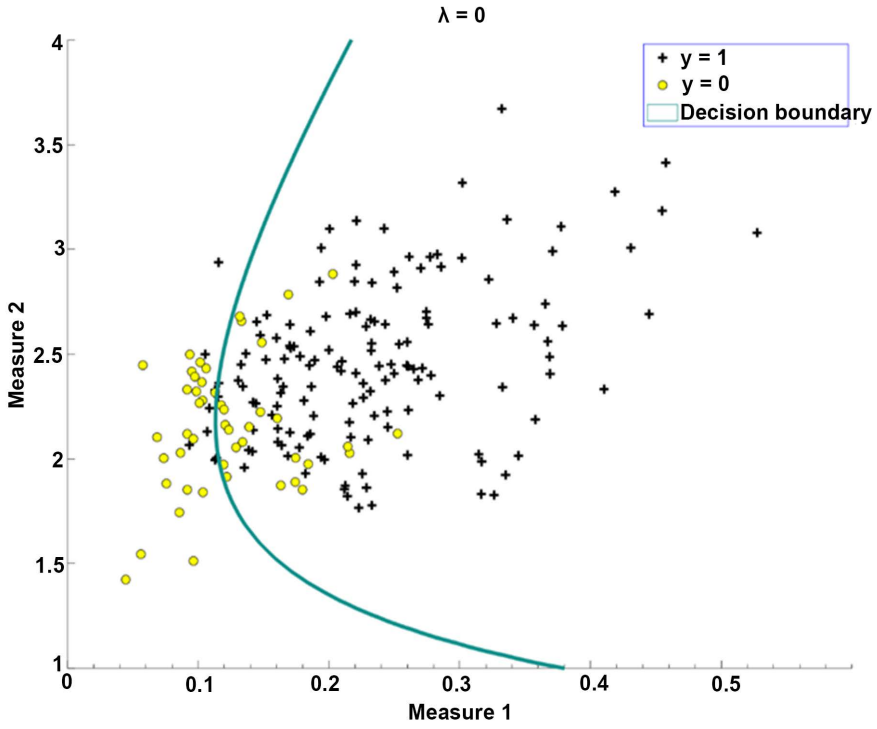

Figure 2. Decision boundary with regularization parameter $\lambda = 0$.

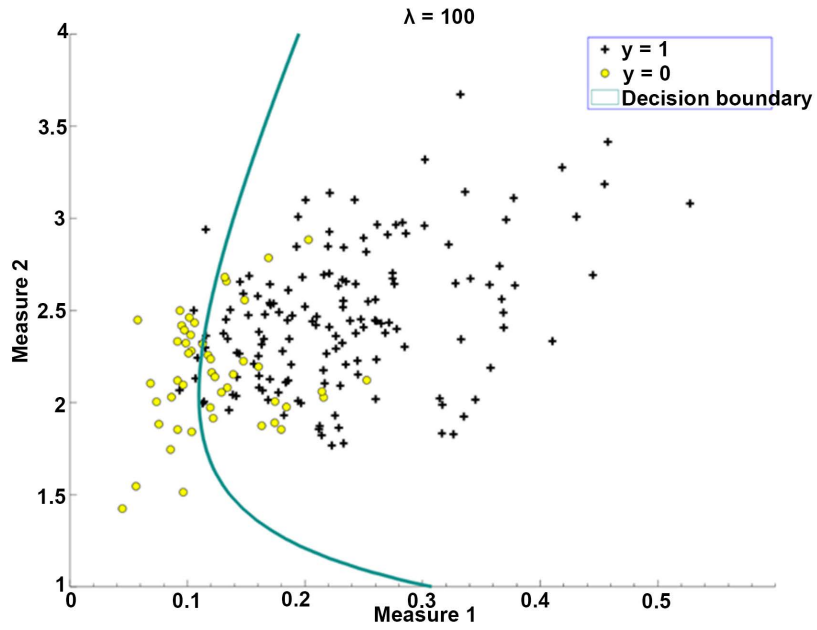

Figure 3. Decision boundary with regularization parameter λ =100.

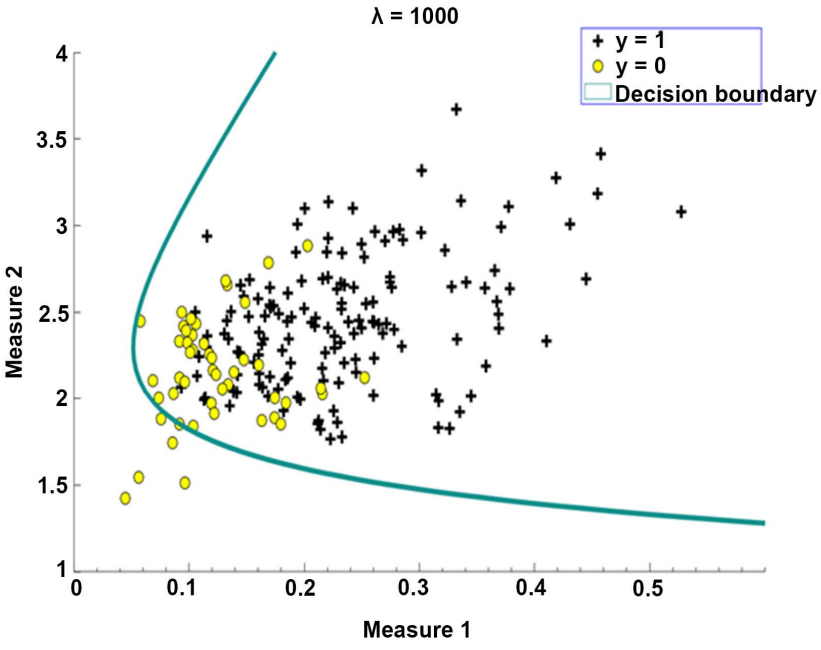

Figure 4. Decision boundary with regularization parameter λ = 1000.

Since already loading a specific range of theta and lambda, the final part is regularization and accuracies. Here need to repeat the steps of finding initial_theta which is the same as the last part—regularized logistic regression. Then the next step is set regularization parameter lambda to 1, lambda = 0. Equation (4) is formula Cost function with lambda:

$$J(\Theta)=\frac{1}{2m}\left\{\sum_{i=1}^{m}\left[h_\theta\left(x^{(i)}\right)-y^{(i)}\right]^2+\lambda\sum_{j=1}^{n}\theta_j^2\right\} \tag{4}$$

If we look at the part before lambda as an "objective", then lambda is keep the training features small. On the one hand, when lambda estimates towards zero reduced the risk overfitting. On the another hand, Lambda set on zero because when lambda bigger and bigger then the decision boundary will not be accurate. After set lambda, it can directly go to set options and optimized (meliorate) FUNC such as *t*, *x*, *y*, lambda, initial_theta, options. The last step in this part was the most important. For this step need to plot decision boundary. First, labels and legends are given measure as *x* and *y* on the coordinate axis. The form as *x* label ("Measure 1") and *y* label ("Measure 2"). Second, then repeat the same rule as Part one. Set "$y = 1$" for healthy people, "$y = 0$" for patients with Parkinson's disease. Third, the last term is training accuracy and predicts variables in **Figure 3** and **Figure 4**.

Parkinson's has such a large population. It's impossible to give accurate results in a single table. Therefore, logistic regression can help us collect information for psychological application, which can effectively distinguish Parkinson's patients from non-Parkinson's patients in the population.

5. Conclusion

In this paper, we have proposed a method with regularized logistic regression to study the origin of Parkinson's disease. Although this method is preliminary, we have given a decision boundary for deciding whether a patient has Parkinson's disease or not by including two measures. More measures could be included in the future with high performance computers. With the number of patients with Parkinson's disease increasing every year, the investigation into the disease is important. Since psychology is a discipline that needs to cooperate with statistics and conduct experiments, it is very important to analyze the independent variables in order to determine the outcome. Through calculation, the data came directly reflect the changes caused by the influence of some factors on the real population. We hope the method with preliminary results presented here could help open a new way for further investigation.

Conflicts of Interest

The author declares no conflicts of interest regarding the publication of this paper.

References

[1] Leroy, E., Boyer, R., Auburger, G., *et al.* (1998) The Ubiquitin Pathway in Parkinson's Disease. *Nature*, **395**, 451-452. https://doi.org/10.1038/26652

[2] De Lau, L.M.L. and Breteler, M.M.B. (2006) Epidemiology of Parkinson's Disease. *The Lancet Neurology*, **5**, 525-535. https://doi.org/10.1016/S1474-4422(06)70471-9

[3] Lang, A.E. and Lozano, A.M. (1998) Parkinson's Disease. *New England Journal of Medicine*, **339**, 1130-1143. https://doi.org/10.1056/NEJM199810153391607

[4] Qi, H. and Li, S.X. (2014) Dose-Response Meta-Analysis on Coffee, Tea and Caffeine Consumption with Risk of Parkinson's Disease. *Geriatrics & Gerontology*

International, **14**, 430-439. https://doi.org/10.1111/ggi.12123

[5] Seber, G.A.F. and Lee, A.J. (2012) Linear Regression Analysis. John Wiley & Sons, Hoboken.

[6] Currier, J.S., Taylor, A., Boyd, F., *et al.* (2003) Coronary Heart Disease in HIV-Infected Individuals [J]. *Journal of Acquired Immune Deficiency Syndromes*, **33**, 506-512. https://doi.org/10.1097/00126334-200308010-00012

[7] Neter, J., Kutner, M.H., Nachtsheim, C.J., *et al.* (1996) Applied Linear Statistical Models. Irwin, Chicago.

[8] Tanaka, H., Hayashi, I. and Watada, J. (1989) Possibilistic Linear Regression Analysis for Fuzzy Data. *European Journal of Operational Research*, **40**, 389-396. https://doi.org/10.1016/0377-2217(89)90431-1