Predicting Heart Attack using local L_P-norm estimators

Ladan Tazik

December 7, 2020

Abstract

In this report, I evaluated the performance of a newly developed method, local $L_p$norm in classification task. The dataset used in this project is Framingham heart that contains variables to predict chance of having cardiovascular disease in ten years. The evaluation metrics are compared with famous logisitic regression analysis. In order to use our method, I needed to select a feature from dataset which not only has a non-normal distribution, but also is significant enought to partially explain the target variable is our dataset. I used backward selection method based on p-value for feature selection and then chose 2 numerical variables to build two different univariate model. Then I applied both logisitic regression and local $L_p$norm method and compared recall, precision score and accuracy of these two models. Results indicate that our method perform poorly in predicting heart disease.

1 Introduction

Samples that are taken from physical phenomenon, are usually noisy. There are factors that can cause noise such omitted variables, nonlinearities, measurement errors, unpredictable effects, etc [6]. The data set [1] that I used for this project is from a research in cardiovascular disease which contains 13 variables including the variables that are from the medical sensors like blood pressure measurement devices. These kind of variables added to some demographic variables such age and sex are used to predict a 10 year risk of heart attack. Recently, I've been working on a method that can be used when the ordinary least squares method is not efficient, or more specifically when stochastic term of the model is not normally distributed, which most of the time, in practice, like this data set we are dealing with such data. My research question was to evaluate our method in prediciting heart disease and compare our method with logistic regression which is a well-known method in classification. The metric used for evaluation is accuracy, precision and recall score.

2 Materials and Methods

We are aiming to understand whether there are advantages to using local L_p norm instead of logistic in the non-parametric regression context. First, I want to briefly explain how our method work.

2.1 Local $L_p$norm Estimator

The family of L_p norm is a generalization of the least square's technique. Under the regular assumptions, the log-likelihood associated with the sample, when density function is not normal, Maximum Likelihood estimator is equivalent to L_p norm when p is specified . In other words, L_p norm produces lower variance than alternatives, so it's an efficient estimator. Like in ordinary least squares; L_p norm estimators can be obtained by minimizing the pth powers of absolute value of residuals [5]:

$$\sum_{i=0}^{n} |y_i - \hat{y}_i|^p = \sum_{i=0}^{n} e_i^p$$

Where y is the actual value and $\hat{y}$ is the predicted value of the model.

In local $L_p$norm, we combine the idea of local polynomial regression with L_p-norm estimations. In local regression, the polynomial is fit using weighted least squares, giving more weight to points near the point whose response is being estimated and less weight to points further away [3]. Our idea is to use weighted *lp*-norm instead of weighted least squares to fit the model. Suppose response variable is a function of independant variables such that : $Y_i = g(X_i) + \varepsilon$. Local l_p method minimize the equation below to make the predicted value ($\hat{y}$) as close as possible to the actual value (y).

$$\sum_{i=0}^{n} [Y_i - \sum_{j=0}^{n} (X_i^{(j)} - x^{(j)})]^p K_H(X_i - x)$$

K_H is a kernel function that is used to assign weight to observations near the current point.[4]

2.2 Feature Selection

The current version of local L_p norm estimators algorithm is only accept univariate dataset as input. So, first of all, to have a fair comparison between logistic regression and local $L_p$norm method, I need to select a variable that explains the variation in the target variable more than any other variables. Before that I need to make sure that variable is not normally distributed to meet the assumption of non-normality for local $L_p$norm. Using density plot, the distribution of numerical variables(*number of cigarettes per day, total cholesterol level, blood pressures, glucose level, BMI and heart rate*) is obtained. Among those variables, *number of cigarettes per day* and *systolic blood pressure* shows outliers and slight right skewness respectively.

Next question is, If I include only these variables in my model, does any one these variables could sufficiently explain the target variable? To answer this question, I used Backward Selection method [7]. In this method, all the variables are considered in the regression model and method is going to remove the least significant variables one at a time. The least significant variable is selected to remove if has the highest p-value in the model, or its elimination from the model causes the lowest drop in R^2, or the lowest increase in Residuals Sum of Squares compared to other features. [2].

By narrowing down to 6 most relevant features, both *number of cigarettes* *per day* and *systolic blood pressure* are among 6 top variables that have most effect on the target variable. Then, I applied both logistic regression and local $L_p$norm on two models, with one of these variable each time.

There are standard metrics that are used to evaluate a model. These metrics are:

- Precision: What proportion of positive prediction was actually correct? $(\frac{TruePositive}{TruePositive+FalseNegative})$
- Recall: What proportion of actual positives was identified correctly? $(\frac{TruePositive}{TruePositive+FalsePositive})$
- Accuracy: What proportion of prediction was identified correctly? $(\frac{TruePositive}{TotalNumberofPredictions})$

3 Result

For logistic, I split the dataset into two parts; 80% for training the logistic and 20% for test data.Tables 1 and 2 below summarize the performance of logistic regression and local $L_p$norm method.

4 Discussion

One obvious result from both tables is that none of those variables, alone is not sufficient to explain 10 year risk of heart attack, as recall and precision score of both model is pretty low, especially for the feature *number of*

Table 1: Evaluation result for the model with feature:*number of cigarettes per day*

metric	**logistic regression**	**local $L_p$norm**
accuracy	0.849	0.847
precision	0.150	0.092
recall	0	0

Table 2: Evaluation result for the model with feature:*systolic blood pressure*

metric	**logistic regression**	**local $L_p$norm**
accuracy	0.847	0.848
precision	0.155	0.075
recall	0.176	0.054

cigarettes per day. A very interesting result is high accuracy in both models, while other two metrics are not promising. By definition of accuracy ; we conclude that both model perform significantly better in finding true positives. On the other hand, recall and precision is very low, it means that both model produce high rate of false negative and false positive.

For the purpose of comparing our new method with logistic regression, results from table 1 and 2 would indicate that local $L_p$norm perform worse than logistic regression and won't be an option for classification task. The reason behind this result might be the nature of local regression which is going to consider the neighbors of a datapoints and if the target value is bouncing between two variables with large gap in the values, local L_p norm could not efficiently smooth the curve that is responsible for creating the distribution of the observation.

5 Conclusion

Although those two variable alone does not build a good model, but for the purpose of comaring our method to logisitic regression, I needed to include only continous non-normal variables. I set logisitic regression as a benchmark to compare the evaluation of local L_p norm method. Results indicate that our method produce higher rate of false prediction compare to logistic regression as precision and recall score is lower than logistic regression. Accuracy of local l_p norm was similar to logistic regression, so our model perform significantly better in finding true positives. Overal, local $L_p$norm perform poorly in predicting risk of 10-year heart disease as its misclassification rate (false negative and positive rate) is high.

References

[1] Aman Ajemra. Framingham heart study dataset, 2017, (accessed Oct,2020).

[2] George Choueiry. Understand forward and backward stepwise regression, 2020 (accessed 29 November 2020).

[3] Jianqing Fan, Irène Gijbels, Tien-Chung Hu, and Li-Shan Huang. A study of variable bandwidth selection for local polynomial regression. *Statistica Sinica*, pages 113–127, 1996.

[4] Hugh Miller, Peter Hall, et al. Local polynomial regression and variable selection. In *Borrowing Strength: Theory Powering Applications–A Festschrift for Lawrence D. Brown*, pages 216–233. Institute of Mathematical Statistics, 2010.

[5] AH Money, JF Affleck-Graves, ML Hart, and GDI Barr. The linear regression model: Lp norm estimation and the choice of p. *Communications in Statistics-Simulation and Computation*, 11(1):89–109, 1982.

[6] Daniele Signori. Chapter 1: An overview of regression statistical analysis of economic data, Sept 2013.

[7] S. Srinidhi. Backward elimination for feature selection in machine learning, 2019, November 19 (accessed Dec 01,2020).