A L1 Regularized Logistic Regression Model for High-dimensional Questionnaire Data Analysis

Jiasheng Wang1*

1School of Sociology, Shanghai University, Shanghai, 200072, China
*Corresponding author’s e-mail: wangjiasheng@shu.edu.cn

Abstract. The L1 regularization method, or Lasso, is a technique for feature selection in high-dimensional statistical analysis. This method compresses the coefficients of the model by using the absolute value of the coefficient function as a penalty term. By adding L1 regularization to log-likelihood function of Logistic model, variable screening method based on the logistic regression model can be realized. The process of variable selection via Lasso is illustrated in Figure 1. The purpose of the experiment is to figure out the important factors that influence interviewees’ subjective well-being using L1 regularized logistic regression. Experiments have been performed on CGSS 2017 data. Important features have been successfully selected by using the L1 regularization method.

1. Introduction

A questionnaire is a research instrument that consists of a set of questions or other types of prompts that aims to collect information from a respondent. The questionnaire will be filled out by mail, face-to-face answers or follow-up interviews. By doing that, the researchers can understand the interviewees’ opinion on a certain phenomenon or problem.

However, many large-scale questionnaires contain hundreds of questions. Due to the large number of questions, it is difficult for researchers to determine which questions are critical. In fact, only the key points in the analysis of the questionnaire need to be taken into account and those less influential questions. Furthermore, because the questions in the questionnaire survey are more ordered than the scalar data used in traditional data mining, the traditional regression model is also difficult to handle.

The L1 regularization method, or Lasso, is a novel statistical technique for selecting significant variables in high-dimensional statistical analysis. Robert Tibshirani proposed the Lasso for the first time in 1996 [1]. The Lasso method compresses the coefficients of the model by using the absolute value of the coefficient function as a penalty term, so that the coefficient with a smaller absolute value is more likely to be zero. In recent years, as information technology has advanced, the amount of data has increased. Numerous lasso-based methodologies and empirical studies have been conducted in the field of social research in order to help improve the statistical analysis of data which has high-dimensional variables [2, 3, 4]. This paper applies the L1 regularized logistic regression model on the CGSS data, a large-scale nation-wide questionnaire, to select critical questions, laying the groundwork for future research. Logistic regression can be integrated with the L1 regularization method.
2. Method

Let $Y$ be a binary dependent variable with the value $\{0, 1\}$, $X = (x_1, ..., x_p)$ is a $p$-dimensional variable, and the $i$-th sample is $(x_i, y_i)$, $i = 1, ..., n$. Logistic regression model is as follows:

$$
\sum_{i=0}^{n} \frac{P(y_i=1|x_i)}{1-P(y_i=1|x_i)} = \beta_0 + \sum_{j=1}^{p} x_{ij} \beta_j
$$

(1)

Where $P(y_i=1|x_i)$ represents the probability of $Y$. $\beta_0$ is the intercept term. $\beta_j$ is the regression coefficient, $\beta = (\beta_1, \ldots, \beta_p)^T$ is the regression coefficient vector, then the logarithmic likelihood function is as follows:

$$
l(\beta_0, \beta) = \sum_{i=1}^{n} \left( y_i \left( \beta_0 + \sum_{j=1}^{p} x_{ij} \beta_j \right) - \log \left( 1 + \exp \left( \beta_0 + \sum_{j=1}^{p} x_{ij} \beta_j \right) \right) \right)
$$

(2)

By adding L1 regularization to the log-likelihood function of Logistic model, variable screening method based on Logistic regression model can be realized. That is:

$$
(\beta_{\text{L1}}, \beta) = \arg\min_{\beta_0, \beta} \left( -\frac{l(\beta_0, \beta)}{n} + \lambda \sum_{j=1}^{p} |\beta_j| \right)
$$

(3)

The process of variable selection via Lasso is illustrated in Figure 1. Consider a two-dimensional linear regression model, where $\beta$ denotes the point of minimum mean square error in the absence of constraints. After adding the L1 regularization, the feasible region of will fall into the square region, and the point tangent to the blue square is also the point that conforms to the constraint and minimizes the sum of squares of error. Due to the square constraint, the point of contact between the square and the concentric ellipse is typically at the square vertex. Because the vertices are on the coordinate axes, one of the coefficients of the independent variables that satisfy the constraint has a value of zero. When the coefficient is zero, it acts as the selector of variables. $\lambda$ is a hyperparameter that is used to control the penalty strength. The bigger the $\lambda$ is, the more likely the coefficients go to zero.

To sum up, the Logistic model with L1 regularization will have the absolute value of the compression coefficient in the regression process, making the variable coefficient approach to zero faster, and the variable coefficient zero plays the role of screening variables. Variables whose coefficients are not zero can be considered more important. Then the ordinary Logistic model is used to carry out regression on these important variables and obtain the unbiased estimation results of coefficients for in-depth analysis.
3. Experiment
The experiment is conducted on CGSS 2017. The purpose of the experiment is to figure out the important factors that influence interviewees’ subjective well-being using L1 regularized logistic regression. Let the subjective well-being of the interviewees as the binary scalar output $Y = \{y_1, \ldots, y_n\}$, where $n$ is the total amount of the interviewees. The questionnaire data act as the input values $X_{n \times p}$, $p$ is the amount of the questions. We train a model $f(X, \lambda) = Y$, where $\lambda$ is the hyperparameter of the L1 regularized logistic model to control the penalty strength of the coefficients.

3.1 Data Description
CGSS 2017 (China General Social Survey 2017) is a comprehensive academic survey project in China. In the form of questionnaires, CGSS collects data of social, community, family, and individual, summarizing the trend of Chinese social change. CGSS is widely used in scientific research and contributes significantly to the studying of Chinese society.

A15. How do you feel about your current physical health?

- Very unhealthy .......................................................... 1
- Unhealthy ................................................................. 2
- Normal........................................................................ 3
- Healthy ....................................................................... 4
- Very Healthy .............................................................. 5
- I don’t know .............................................................. 98
- I’d prefer not to answer ............................................... 99

Some questions in CGSS are completion questions resulting to a scalar answer, such as “My age is ___ years old” which is easy to handle during the analysis process. However, other questions are selection question with ordinal options as shown in Figure 2:

When applying the regression model, the coefficients of these questions cannot be interpreted directly. The optimal approach is to implement a feature selection method to identify the critical questions and analyze them case by case. This is another reason for utilizing L1 regularized logistic regression in the questionnaire data analysis process.

![Figure 3. Variations in coefficients.](image-url)
3.2 Experiment Results
In this experiment, in data matrix $X$ there are 155 candidate variables which has the potential to affect the subjective well-being. Code means the codes of the questions in CGSS 2017. There is a penalty factor $\lambda$ in L1 regularized logistic model. We set $\lambda$ as 1, 0.1, 0.01. The coefficients of the regression are shown in Table 1. According to the Table 1, we can see that a15, a17, a27k1, a283, a301, a304, a306, a308, a309 are important variables. Figure 3 shows the change in coefficients while $\lambda$ varies.

4. Conclusion
In this paper we describe the framework of L1 regularization and demonstrate an application on high-dimensional CGSS data. Feature selection on the high-dimensional data has been easily completed. In today's era of increasing data volumes, feature selection in high-dimensional data is a critical part of data analysis.

References
[1] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society: Series B (Methodological), 58(1), 267-288.
[2] Belloni, A., Chen, D., Chernozhukov, V., & Hansen, C. (2012). Sparse models and methods for optimal instruments with an application to eminent domain. Econometrica, 80(6), 2369-2429.
[3] Finkelstein, A., Gentzkow, M., & Williams, H. (2016). Sources of geographic variation in health care: Evidence from patient migration. The quarterly journal of economics, 131(4), 1681-1726.
[4] Ashraf, N., Bau, N., Low, C., & McGinn, K. (2020). Negotiating a better future: How interpersonal skills facilitate intergenerational investment. The Quarterly Journal of Economics, 135(2), 1095-1151.

Table 1. Coefficients of variables.

| Code | $\lambda = 1$ | $\lambda = 0.1$ | $\lambda = 0.01$ | Code | $\lambda = 1$ | $\lambda = 0.1$ | $\lambda = 0.01$ |
|------|---------------|----------------|----------------|------|---------------|----------------|----------------|
| a15  | 0.178         | 0.268          | 0.287          | a285 | -0.028        | -0.041          |
| a15b | 0.013         | 0.017          | a286           | 0.044 | 0.053        |
| a16  | -0.020        | -0.042         | a29            |      |              | 0.018          |
| a17  | 0.446         | 0.528          | 0.542          | a301 | -0.083        | -0.104          |
| a18  | 0.017         | 0.027          | a302           |      |              | -0.008         |
| a21  | -0.168        | -0.187         | a303           |      |              | -0.003         |
| a27b | -0.024        | -0.029         | a304           | -0.087 | -0.102    | -0.106         |
| a27d | -0.024        | -0.039         | a305           |      |              | -0.107          |
| a27f | -0.004        | a306           | -0.013         |      |              | -0.175         |
| a27h | 0.013         | a307           | 0.095          |      |              | 0.111          |
| a27k1| -0.051        | -0.090         | a308           | -0.033 | -0.078    | -0.085         |
| a27k2| -0.029        | -0.035         | a309           | -0.033 | -0.042    | -0.045         |
| a281 | 0.005         | 0.005          | a3010          | 0.156 | 0.198        |
| a282 | 0.010         | 0.014          | a3011          |      |              | 0.011          |
| a283 | 0.023         | 0.082          | 0.087          | a3012 | 0.013        | 0.012          |