The 33rd Annual Conference of the Japanese Society for Artificial Intelligence, 2019

A Matrix-Operation Fast Approximated Solution for Logistic Regression with Strong L2 Regularization

Zeke Xie[1], Jinze Yu*[1], and Yanping Deng[2]

[1]The University of Tokyo, Japan
[2]Waseda University

We propose a second-order approximated solution for Logistic Regression with strong L2 regularization based on matrix operations. As training a Logistic Regression model is a convex optimization, researchers have efficient techniques solving it, such as Gradient Descent. But, to our best knowledge, a solution in the form of matrix operations has not been revealed. Generally speaking, matrix operation is faster and more convenient than solving optimization problems. In principle, the matrix-operation approximated solution is only applicable to Logistic Regression with very strong L2 regularization, however, it also works as a pretty good approximated solution in our empirical analysis even when the L2 regularization strength is set in a practical range. This method can also generate good parameter initialization efficiently. The mathematical proof is presented in this paper.

1. Introduction

Logistic regression is the appropriate regression analysis to conduct when dealing with dichotomous (binary) problem. Like all regression analyses, the logistic regression is a predictive analysis. Logistic regression is used to describe data and to explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval or ratio-level independent variables. For machine learning task, Logics Regression model is largely used for unlinear classifiers [Mitchell 05]. As training a Logistic Regression model is a convex optimization, researchers have efficient techniques, such as Gradient Descent, solving it. But, to our best knowledge, a solution in the form of matrix operations has not been revealed. Generally speaking, matrix operation is faster and more convenient than solving optimization problems. In principle, the matrix-operation approximated solution is only applicable to Logistic Regression with very strong L2 regularization, however, it also works as a pretty good approximated solution in our empirical analysis even when the L2 regularization strength is set in a practical range. This method can also generate good parameter initialization efficiently. We propose a second-order approximated solution for Logistic Regression with strong L2 regularization based on matrix operations. The mathematical proof is presented in this paper.

2. Logistic Regression with strong $L2$ regularization

Suppose we are given a data set $X \in R^{n\times m}, y \in \{0, +1\}$ for a binary classification problem. X, a $n \times m$ data matrix, contains n data samples, and each feature vector x^i has m attributes. The target variable vector y is a binary-value column vector with a length of n. Assume we decide to learn a Logistic Regression model mapping from X to y. θ is a column vector parameterizing

Contact: Name, Affiliation, Address, Phone number, Facsimile number, and E-mail address

$$p(y = 1|x) = h_\theta(x) = \frac{e^{-x^T\theta}}{1 + e^{-x^T\theta}} \tag{1}$$

In our discussion, we ignore the intercept term for the simplicity of our notation. The maximum log-likelihood method gives

$$\begin{aligned} l(\theta) &= \sum_{i=1}^{n} lnp(y^{(i)}|x^{(i)}; \theta) \\ &= \sum_{i=1}^{n} (-y_i x^{(i)T}\theta + ln(1 + e^{x^{(i)T}\theta})) \end{aligned} \tag{2}$$

To prevent over-fitting, people often add a regularization term into the log-likelihood. We choose L2-norm regularization and use λ control the strength of regularization, and we obtain the expression of θ as

$$\theta(\lambda) = {}_{\theta\in R^m}(l(\theta) + \lambda\|\theta\|_2^2) \tag{3}$$

where the weight vector θ only depends on λ given a certain data set X, y.

We present the matrix-operation approximated solution first, and then discuss the mathematical proof in next section. With the strong $l2$ regularization limit, the parameter vector θ is a vector with a length of nearly zero, which means $\|\theta\|_2 \to 0$ when $\lambda \to +\infty$. In this situation, only the direction of vector θ still matters. Logistic regression employs the hyperplane $z = \theta x$ to discriminate samples. In logistic regression, $z > 0$ makes positive prediction, and $z < 0$ makes negative prediction. The normalized θ may be written as

$$\begin{aligned} e_\theta(\lambda) &= \frac{\theta(\lambda)}{\|\theta(\lambda)\|_2} \\ e_\theta(\lambda) &= \frac{{}_{\theta\in R^p}(l(\theta) + \lambda\|\theta\|_2^2)}{\|{}_{\theta\in R^p}(l(\theta) + \lambda\|\theta\|_2^2)\|_2} \end{aligned} \tag{4}$$

where the normalization may change the predicted probability but makes no difference on the classification result.

And we present the second-order approximated solution for Logistic Regression as follows:

$$\hat{\theta} \approx (2\lambda I + \frac{1}{4}X^T X)^{-1} X^T (y - \frac{1}{2}) \tag{5}$$

applicable when λ is large, e.g. $\lambda > 1000$. We find out that $\|\theta\|_2 \sim O(\frac{1}{\lambda})$. It indicates the fact that $\|\theta\|_2$ is a first-order small quantity if λ is a first-order large quantity. And its direction vector is given by

$$\hat{e}_\theta \approx \frac{(2\lambda I + \frac{1}{4}X^T X)^{-1} X^T (y - \frac{1}{2})}{\|(2\lambda I + \frac{1}{4}X^T X)^{-1} X^T (y - \frac{1}{2})\|_2} \tag{6}$$

With the limit of strong L2 regularization, we can use the simpler approximated solution

$$\hat{e}_\theta = \lim_{\lambda \to +\infty} \frac{\theta(\lambda)}{\|\theta(\lambda)\|_2} = \frac{X^T (y - \frac{1}{2})}{\|X^T (y - \frac{1}{2})\|_2} \tag{7}$$

where we denote the column vector $(1, \ldots, 1)^T$ as 1.

3. Mathematical Proof

In this section, we organize the proof of the matrix-operation solution. We decide to Taylor expand the regularized log-likelihood function. And we first Taylor expand the second term, a multivariable function, $ln(1+e^{x^T\theta})$, and thus have

$$ln(1 + e^{x^T\theta}) = ln2 + \frac{x^T\theta}{2} + \sum_{i,j=1}^{m} \frac{x_j(x_i - 1)}{8} + O(\theta^3) \tag{8}$$

$$ln(1 + e^{x^T\theta}) = ln2 + \frac{x^T\theta}{2} + \frac{x^T\theta(x^T\theta - 1^T\theta)}{8} + O(\theta^3) \tag{9}$$

using multivariate Taylor Expansion. As we know that θ is a first-order small quantity, we may ignore the three-order small quantities with no impact on the second-order approximated solution. So we write the log-likelihood in the form of

$$l(\theta) = \sum_{i=1}^{n} [ln2 + (\frac{1}{2} - y^{(i)})x^{(i)T}\theta + \frac{x^{(i)T}\theta(x^{(i)T}\theta - 1^T\theta)}{8} + \frac{\lambda}{n}\theta^T\theta + O(\theta^3)] \tag{10}$$

And we have its partial derivative as

$$\frac{\partial l(\theta)}{\partial \theta_j} = \sum_{i=1}^{n} [(\frac{1}{2} - y^{(i)})x_j^{(i)} + \frac{2\lambda}{n}\theta_j + \frac{(2x_j^{(i)} - 1)}{8}x^{(i)T}\theta - \frac{x_j^{(i)}}{8}1^T\theta + O(\theta^2)] \tag{11}$$

Here we assume the data matrix satisfies the mean normalization $\sum_{i=1}^{n} x_j^{(i)} = 0$.

$$\frac{\partial l(\theta)}{\partial \theta_j} = \sum_{i=1}^{n} [(\frac{1}{2} - y^{(i)})x_j^{(i)} + \frac{2\lambda}{n}\theta_j + \frac{1}{4}x_j^{(i)}x^{(i)T}\theta + O(\theta^2)] \tag{12}$$

$$\frac{\partial l(\theta)}{\partial \theta_j} = \sum_{i=1}^{n} [(\frac{1}{2} - y^{(i)})x_j^{(i)} + \frac{2\lambda}{n}\theta_j + \sum_{k=1}^{m} \frac{1}{4}x_j^{(i)}x_k^{(i)T}\theta_k + O(\theta^2)] \tag{13}$$

We ignore high-order small quantities $O(\theta_2)$ in the above system of equations.

$$\frac{\partial l(\theta)}{\partial \theta_j} = \sum_{i=1}^{n} [(\frac{1}{2} - y^{(i)})x_j^{(i)} + \frac{2\lambda}{n}\theta_j + \sum_{k=1}^{m} \frac{1}{4}x_j^{(i)}x_k^{(i)T}\theta_k] \tag{14}$$

We denote the system of linear equations as $M\theta = N$, where $M = 2\lambda I + \frac{1}{4}X^T X$ and $N = X^T(y - \frac{1}{2})$. It simply implies that the second-order approximated matrix-operation solution for Logistic Regression with strong L2-regularization is

$$\hat{\theta} \approx (2\lambda I + \frac{1}{4}X^T X)^{-1} X^T (y - \frac{1}{2}) \tag{15}$$

And its direction vector is given by

$$\hat{e}_\theta \approx \frac{(2\lambda I + \frac{1}{4}X^T X)^{-1} X^T (y - \frac{1}{2})}{\|(2\lambda I + \frac{1}{4}X^T X)^{-1} X^T (y - \frac{1}{2})\|_2} \tag{16}$$

With the limit of strong L2 regularization, we can use the simpler approximated solution

$$\hat{e}_\theta \approx \frac{X^T (y - \frac{1}{2})}{\|X^T (y - \frac{1}{2})\|_2} \tag{17}$$

4. Summary and Conclusion

Logics Regression model is largely used for unlinear classifiers. To solve a Logics Regression model, the efficient way is to use numeric solver such as gradient descent. In this paper, we proposed a Matrix-Operation Fast Approximated Solution for Logistic Regression with Strong L2 Regularization. We conclude that the proposed solution is much faster than the conventional solution. And, in practice, the proposed solution works very closely to the conventional solution.

References

[Mitchell 05] T. Mitchell, Generative and Discriminative Classifiers: Naive Bayes and Logistic Regression. Draft Version, 2005