L2-Regularized Logistic Regression

Wenbin Hu

May 2024

1 Introduction

Machine learning algorithms have become increasingly important in a wide range of applications, from image recognition to natural language processing to predictive analytics. One of the fundamental algorithms in the machine learning toolkit is Logistic Regression, which is particularly well-suited for binary classification tasks. Logistic Regression models the probability of a binary outcome as a function of one or more predictor variables, allowing it to make predictions about whether a given input belongs to one of two classes.

$$P(y = 1|x) = \frac{1}{1 + e^{-\omega^T x}} \tag{1}$$

$$P(y = 0|x) = 1 - \frac{1}{1 + e^{-\omega^T x}} \tag{2}$$

, where x is data, y is the corresponding label, and ω is the model weight.

However, when faced with a large number of predictor variables, Logistic Regression models can be prone to a problem known as overfitting. Overfitting occurs when a model learns the training data too well, essentially memorizing the specific patterns in the training set rather than capturing the underlying relationships. As a result, the model performs exceptionally well on the training data but fails to generalize to new, unseen data, severely limiting its practical utility.

To address the issue of overfitting in Logistic Regression, researchers have developed various regularization techniques, which aim to simplify the model and improve its ability to generalize. One particularly effective approach is the use of L2 regularization, also known as ridge regression. L2 regularization adds a penalty term to the cost function of the Logistic Regression model, encouraging the model to learn smaller coefficient values and thereby reducing the complexity of the overall model.

$$\min_w C f(x, y, w) + \frac{1}{2}\omega^T \omega \tag{3}$$

, where $f(x, y, \omega)$ is the objective function for logistic regression, $C \in R_{++}$ is the is the regularization hyperparameter.

The goal of this project is to provide a detailed explanation of L2-Regularized

Logistic Regression, exploring its mathematical formulation, the intuition behind its effectiveness, and the practical considerations involved in its implementation. By understanding the principles and mechanics of this powerful technique, readers will be better equipped to apply it in their own machine learning projects and tackle the challenges of overfitting in binary classification tasks.

2 Regularized Logistic Regression

2.1 Logistic Regression

From equations (1), (2), we model the conditional probability.

$$P(y|x) = \frac{1}{1 + e^{-y\omega^T x}} \tag{4}$$

Given a set of instance-label pairs (x_i, y_i), where $x_i \in R^d, y_i \in -1. + 1$, the logistic regression model estimates the model weight with maximum likelihood.

$$\begin{aligned}
\arg\max \prod_{i=1}^{n} P(y_i|x_i) &= \arg\max \prod_{i=1}^{n} \frac{1}{1 + e^{-y_i\omega^T x_i}} \\
&= \arg\max \log \prod_{i=1}^{n} \frac{1}{1 + e^{-y_i\omega^T x_i}} \\
&= \arg\max \sum_{i=1}^{n} \log \frac{1}{1 + e^{-y_i\omega^T x_i}} \\
&= \arg\max - \sum_{i=1}^{n} \log(1 + e^{-y_i\omega^T x_i}) \\
&= \arg\min \sum_{i=1}^{n} \log(1 + e^{-y_i\omega^T x_i})
\end{aligned}$$

2.2 L2-Regularized Logistic Regression

To mitigate overfitting problem of regularization of logistic regression, researchers usually use L2 Regularization method. L2 regularization adds a term to the cost function that penalizes large coefficient values, effectively shrinking the model parameters towards zero and reducing the overall complexity of the model. The objective function is:

$$f(x, y, \omega) := C \sum_{i=1}^{n} \log(1 + e^{-y_i\omega^T x_i}) + \frac{1}{2}\omega^T\omega \tag{5}$$

, where $f(x, y, \omega)$ is the objective function for logistic regression, $C \in R_{++}$ is the is the regularization hyperparameter.

2.3 Convexity Analysis

The L2-regularized logistic regression function can be divided into 2 parts, and we analyse each of them.

- Loss Function: $\sum_{i=1}^{n} \log(1 + e^{-y_i \omega^T x_i})$. $-y_i \omega^T x_i$ is convex for every i because it is affine. $1 + e^{-y_i \omega^T x_i}$ is convex because e^x is convex and increasing. The loss function $\sum_{i=1}^{n} \log(1 + e^{-y_i \omega^T x_i})$ is convex because it is the sum of n convex functions.
- L2 Regularization Term: $\frac{1}{2}\omega^T \omega$. It is the square of l2-norm. Thus, the l2 regularization term is convex.

The L2-regularized logistic regression function is convex because it is exactly the sum of 2 convex functions.

2.4 Optimization

From the equation 5, we formulate the optimization problem:

$$\min_{w} f(x, y, \omega) := C \sum_{i=1}^{n} \log(1 + e^{-y_i \omega^T x_i}) + \frac{1}{2}\omega^T \omega \tag{6}$$

, where $f(x, y, \omega)$ is the objective function for logistic regression, $C \in R_{++}$ is the is the regularization hyperparameter.

Theorem S1. *$x^* \in dom\ f$ is optimal for a unconstrained optimization problem with differentiable objective function if only if $\nabla f(x^*) = 0$*

With theorem 1, we know the optimal point x^* satisfies $\nabla f_w(x, y, \omega) = 0$.

$$\begin{aligned}
\nabla f(x, y, \omega) &= \frac{\partial \frac{1}{2}\omega^T \omega}{\partial \omega} + \frac{\partial C \sum_{i=1}^{n} \log(1 + e^{-y_i \omega^T x_i})}{\partial \omega} \\
&= \omega + C \sum_{i=1}^{n} \frac{\partial \log(1 + e^{-y_i \omega^T x_i})}{\partial \omega} \\
&= \omega + C \sum_{i=1}^{n} \frac{1}{1 + e^{-y_i \omega^T x_i}} \frac{\partial (1 + e^{-y_i \omega^T x_i})}{\partial \omega} \\
&= \omega + C \sum_{i=1}^{n} \frac{-y_i e^{-y_i \omega^T x_i}}{1 + e^{-y_i \omega^T x_i}} \frac{\partial \omega^T x_i}{\partial \omega} \\
&= \omega - C \sum_{i=1}^{n} \frac{y_i e^{-y_i \omega^T x_i}}{1 + e^{-y_i \omega^T x_i}} x_i
\end{aligned}$$

Thus,

$$\nabla f(x, y, \omega) = 0 \Leftrightarrow \omega = C \sum_{i=1}^{n} \frac{y_i e^{-y_i \omega^T x_i}}{1 + e^{-y_i \omega^T x_i}} x_i \tag{7}$$

However, the equation 7 involves exponential functions, which is hard to derive the closed form solution. We use Newton's methods to solve the problem.

2.5 Newton's Method

2.5.1 Newton Step

Theorem S2. *$\nu_{nt} = -\nabla^2 f(x) \nabla f(x)$ is an optimal point for $\min_\nu \hat{f}(x + \nu) := f(x) + \nabla f(x)^T \nu + \frac{1}{2} \nu^T \nabla f(x) \nu$, where $\nu_{nt} = -\nabla^2 f(x) \nabla f(x)$ is called Newton Step.*

With theorem 2, we can know that a Newton step is a descent step.

$$\begin{aligned}
\nabla f_\omega^2(x, y, \omega) &= \frac{\partial \nabla f(\omega)}{\partial \omega} \\
&= \frac{\partial \{\omega - C \sum_{i=1}^{n} \frac{y_i e^{-y_i \omega^T x_i}}{1 + e^{-y_i \omega^T x_i}} x_i\}}{\partial \omega} \\
&= I_d - C \sum_{i=1}^{n} \frac{\partial \frac{y_i e^{-y_i \omega^T x_i}}{1 + e^{-y_i \omega^T x_i}} x_i}{\partial \omega} \\
&= I_d - C \sum_{i=1}^{n} x_i \frac{\partial \frac{y_i e^{-y_i \omega^T x_i}}{1 + e^{-y_i \omega^T x_i}}}{\partial \omega}^T \\
&= I_d - C \sum_{i=1}^{n} x_i \frac{-y_i^2 e^{-y_i \omega^T x_i}}{(1 + e^{-y_i \omega^T x_i})^2} \frac{\partial \omega^T x_i}{\partial \omega}^T \\
&= I_d - C \sum_{i=1}^{n} x_i \frac{-y_i^2 e^{-y_i \omega^T x_i}}{(1 + e^{-y_i \omega^T x_i})^2} x_i^T \\
&= I_d + C \sum_{i=1}^{n} \frac{y_i^2 e^{-y_i \omega^T x_i}}{(1 + e^{-y_i \omega^T x_i})^2} x_i x_i^T
\end{aligned}$$

Thus, the Newton Step of the objective function of L2-regularized logistic regression is:

$$\nu_{nt} = -\nabla^2 f(x) \nabla f(x) \tag{8}$$

$$= -(I_d + C \sum_{i=1}^{n} \frac{y_i^2 e^{-y_i \omega^T x_i}}{(1 + e^{-y_i \omega^T x_i})^2} x_i x_i^T)^T (\omega - C \sum_{i=1}^{n} \frac{y_i e^{-y_i \omega^T x_i}}{1 + e^{-y_i \omega^T x_i}} x_i) \tag{9}$$

2.5.2 Optimization Algorithm

Given a starting point $x \in \operatorname{dom} f$, tolerance $\epsilon > 0$.
Repeat
1. Compute the Newton step and decrement.

$$\nu_{nt} := -\nabla^2 f(x)^{-1} \nabla f(x); \quad \lambda^2 := \nabla f(x)^T \nabla^2 f(x)^{-1} \nabla f(x).$$

2. Stopping criterion. quit if $\lambda^2/2 \leq \epsilon$.
3. Line search. Choose step size t by backtracking line search.
4. Update, $x := x + t\nu_{nt}$.

2.6 Stable Solution

Since the matrix inversion would be numerically unstable in certain condition, we can derive an alternative solution to get without matrix inversion. By SVD theorem, we can find U, Σ, V, s.t. $\nabla_\omega f^2(x, y, \omega) = U\Sigma V^T$. Thus the Pseudo-Inverse of $\nabla_\omega f^2(x, y, \omega)$ is: $\nabla_\omega f^2(x, y, \omega)^\dagger = V\Sigma^{-1}U^T$.

Theorem S3. *Least-Square theorem:* $A^\dagger y \in \arg\min_x \|Ax - b\|_2$

Thus, an alternative solution is $-\nabla_\omega f^2(x, y, \omega)^\dagger \nabla_\omega f(x, y, \omega) = V\Sigma^{-1}U^T \nabla_\omega f(x, y, \omega)$.

3 Discussion

In conclusion, the addition of L2 regularization to the Logistic Regression cost function is a powerful technique for improving the model's ability to generalize and avoid overfitting. By incorporating a penalty term that encourages smaller coefficient values, the regularized version of Logistic Regression learns simpler, more robust models that are less prone to capturing spurious patterns in the training data.

The preservation of the convex nature of the cost function ensures efficient optimization and the reliable identification of the global minimum. However, the choice of the regularization hyperparameter C is critical, as an overly high value may result in an oversimplified model, while a too low value may still allow for overfitting.

While L2 regularization is a popular and effective approach, other regularization methods, such as L1 (Lasso) or elastic net, may be more suitable depending on the specific characteristics of the data and the goals of the analysis. Careful consideration of the strengths, limitations, and appropriate use cases of each technique is essential for achieving optimal results.