Predicting Correlation of Bagging Estimators in Logistic Regression

Iris Chang

July 28, 2023

Abstract

Logistic regression is used to make a prediction between two different outcomes based on a data set. Typically, when the sample size is much larger compared to a fixed dimension, maximum likelihood estimation is used to estimate the parameters of the model, in which case the estimate has many convenient properties such as being unbiased. It was shown by [5] as well as [4] that these assumptions do not hold in the high dimensional regime where the sample size and dimension are proportional. While these two groups were able to find adequate methods of characterizing model performance in high dimensions, there is an absence of work on the performance and impact of bagging on high dimensional logistic regression models. In our case, bagging refers to the method of dividing the larger data set into two or more overlapping subsets and fitting a logistic regression model on each subset before aggregating the parts together. This work aims to show a single scalar that would be able to predict the correlation between the two unaggregated estimates. Drawing from previous results in linear regression and the unbagged setting, we were able to successfully infer this result.

1 Introduction

Logistic regression is a statistical model that returns a prediction for a binary outcome given a data set. To use an example, researchers in Indonesia used a logistic regression model to predict swing voters' choice in the polls based on certain demographic information such as the voter's age, religion, and education amongst other factors [2]. In this case, the demographics of "loyal voters" for two candidates were used as the training data set to determine our logistic regression model which then could be used to predict swing voters' choices.

There are many ways of further refining the logistic regression model to better fit the training data set and therefore make more useful predictions. Regularization is one such technique in which a penalty of choice is added to the model. This prevents overfitting by limiting the size of the model parameters. In the case of ridge regularized regression or L_2 regularization, for example, the sum of the square of the model parameters is added to the minimization problem as will be discussed more in depth in the following section.

Typically, maximum likelihood estimation is used to estimate parameter values. When the sample size, n, is much larger than the model dimension, p, such that $n \to \infty$ for fixed p, the maximum likelihood estimate (MLE) for the parameters has a number of convenient properties such as being unbiased. However, as we consider a higher dimensional setting in which n and p are proportional, these properties have been shown to break down. Particularly, [5] described the performance of the MLE in the unregularized setting, finding that the MLE is no longer unbiased in the high dimensional regime [5].

Following this, [4] studied the asymptotic performance of regularized logistic regression parameter estimation in high dimensions [4]. To do so, they used the Convex Gaussian Min-max Theorem to uncover

a six equation nonlinear system in six unknowns. Using the solution of this system, various performance measures of interest (e.g. mean squared error) can be explicitly calculated. In the case of unregularized and ridge regularized logistic regression, this system can be further simplified to three equations and therefore a solution of three terms.

After these prior works, we attempt to investigate the impact of bagging on logistic regression. Bagging is a method commonly used in machine learning in which a model is fit to subsets of a total data set which are then aggregated into one ensemble model. Similarly to regularization, it helps to prevent overfitting and provides a more stable model. While there has been work, as described above, on the performance of logistic regression models in high dimensions under various settings, it remains to be thoroughly explored on a bagged model. In particular, we will show a single scalar that predicts the correlation between the estimated parameters of our bagged logistic regression model.

2 Notation and Set-up

The notation will essentially follow that used by [4], with a few modifications and additional terms. All of these will be described in the following section.

We take a data set of n different $(\boldsymbol{x_i}, y_i)$ pairs whose relation is defined by $\boldsymbol{\beta^*}$ and use our logistic regression model to calculate an estimate for $\boldsymbol{\beta^*}$. To formalize this, let $\boldsymbol{x_i} \in \mathbb{R}^p$ for $i = 1, 2, \ldots, n$ be defined as the vector of p terms corresponding to n terms y_i such that $y_i \in \{0, 1\}$. We will take the $\boldsymbol{x_i}$'s to be drawn i.i.d from a normal distribution with mean 0 and covariance matrix $\frac{1}{p}\boldsymbol{I_p}$, where $\boldsymbol{I_p}$ is the identity matrix in p dimensions. Each y_i is a binary outcome with probability depending on $\boldsymbol{x_i}$ and $\boldsymbol{\beta^*}$. It will be defined as a Bernoulli random variable such that $\mathbb{P}[y_i = 1|\boldsymbol{x}_i] = \rho'(\boldsymbol{x}_i^T\boldsymbol{\beta}^*)$ for $i = 1, 2, \ldots, n$ for $\rho'(t) = \frac{e^t}{1+e^t}$.

The MLE for $\boldsymbol{\beta^*}$ denoted as $\hat{\boldsymbol{b}}$ is given by the following:

$$\hat{\boldsymbol{b}} = \underset{\boldsymbol{b}\in\mathbb{R}^p}{\arg\min} \sum_{i=1}^{n} \rho(\boldsymbol{x}_i^T\boldsymbol{b}) - y_i(\boldsymbol{x}_i^T\boldsymbol{b}) \tag{1}$$

for function $\rho(t) = \log(1 + e^t)$. The estimate $\hat{\boldsymbol{b}}$ is the minimizer of the negative log likelihood.

The three random variables Z, Z_1, Z_2 will all be normally distributed with mean 0 and variance 1. The proximal operator associated with convex function $\rho : \mathbb{R}^d \to \mathbb{R}$ and scalar $t > 0$ for vector $\boldsymbol{v}$ is defined as the following:

$$\text{Prox}_{t\rho(\cdot)}(\boldsymbol{v}) = \underset{\boldsymbol{x}\in\mathbb{R}^d}{\arg\min} \frac{1}{2t}\|\boldsymbol{v} - \boldsymbol{x}\|^2 + \rho(\boldsymbol{x}) \tag{2}$$

As mentioned previously, the six equation system whose solution is used to characterize the performance of the regularized logistic regression model in the high dimensional regime can be simplified to a system of three nonlinear equations in an unregularized and L_2 regularized model. The minimization problem corresponding to this is as follows:

$$\hat{\boldsymbol{b}}_{L_2} = \underset{\boldsymbol{b}\in\mathbb{R}^p}{\arg\min} \frac{1}{n} \cdot \left[\sum_{i=1}^{n} \rho(\boldsymbol{x}_i^T\boldsymbol{b}) - y_i(\boldsymbol{x}_i^T\boldsymbol{b})\right] + \frac{\lambda}{2p}\sum_{i=1}^{p} \boldsymbol{b}_i^2 \tag{3}$$

The three equation system has three unknown (σ, α, γ) and is found in Theorem 2 of [4]. It is as follows:

$$\begin{cases} \dfrac{\sigma^2}{2\delta} = \mathbb{E}[\rho'(-\kappa Z_1)(\kappa\alpha Z_1 + \sigma Z_2 - \text{Prox}_{\gamma\rho(\cdot)}(\kappa\alpha Z_1 + \sigma Z_2))^2] \\ -\dfrac{\alpha}{2\delta} = \mathbb{E}[\rho''(-\kappa Z_1)\text{Prox}_{\gamma\rho(\cdot)}(\kappa\alpha Z_1 + \sigma Z_2)] \\ 1 - \dfrac{1}{\delta} + \lambda\gamma = \mathbb{E}\left[\dfrac{2\rho'(-\kappa Z_1)}{1 + \gamma\rho''(\text{Prox}_{\gamma\rho(\cdot)}(\kappa\alpha Z_1 + \sigma Z_2))}\right] \end{cases} \tag{4}$$

Here, δ is $\frac{n}{p}$ and κ is defined as $\frac{\|\boldsymbol{\beta^*}\|}{\sqrt{p}}$. The regularization parameter is denoted by λ, and the function $\rho''(t) = \frac{e^t}{(1+e^t)^2}$.

2.1 Specific Case

In this paper, we consider the specific case in which $\boldsymbol{\beta^*}$ is 0. As a result of this assumption, y_i is independent from $\boldsymbol{x_i}$ and distributed as Bernouilli($\frac{1}{2}$) as $\rho'(0) = \frac{1}{2}$. Additionally, we will look at the unregularized case, so we consider the three equation reduction above in (4) with $\lambda = 0$.

Due to the restrictions described in the previous paragraph, there are a number of additional simplifications to (4), yielding the following two equation system:

$$\begin{cases} \dfrac{\sigma^2}{\delta} = \mathbb{E}[(\sigma Z - \text{Prox}_{\gamma\rho(\cdot)}(\sigma Z))^2] \\ 1 - \dfrac{1}{\delta} = \mathbb{E}\left[\dfrac{1}{1 + \gamma\rho''(\text{Prox}_{\gamma\rho(\cdot)}(\sigma Z))}\right] \end{cases} \tag{5}$$

In particular, the term κ is zero due to the $\|\boldsymbol{\beta^*}\|$ term (which is zero in this setting) included in its definition. The regularization parameter, λ, is also zero as we consider an unregularized model. The $\frac{1}{2}$ in the first equation and factor of 2 in the expectation of the third equation of (4) cancel due to $\rho'(0) = \frac{1}{2}$. The solution (σ, γ) for fixed values of delta of the equations in (5) will be used later to calculate our predictor for the correlation between parameter estimates.

Considering the application of the bagging method, the data set will be divided into two equally sized subsets I and $\tilde{I}$ such that $|I| = |\tilde{I}| = nc$ for sample size n and some constant $c \in (0, 1)$. These two subsets overlap such that $|I \cap \tilde{I}| = c|I|$ meaning the number of observations contained in their intersection is nc^2. In this case, δ will now be defined as $\frac{|I|}{n}$ applying to equations (5) and onward.

We define $\hat{\boldsymbol{\beta}}$ to be the estimator trained over observations in I and $\tilde{\boldsymbol{\beta}}$ to be the estimator trained over observations in $\tilde{I}$:

$$\begin{aligned} \hat{\boldsymbol{\beta}} &= \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\arg\min} \sum_{i\in I} \rho(\boldsymbol{x}_i^T\boldsymbol{\beta}) - y_i(\boldsymbol{x}_i^T\boldsymbol{\beta}) \\ \tilde{\boldsymbol{\beta}} &= \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\arg\min} \sum_{i\in \tilde{I}} \rho(\boldsymbol{x}_i^T\boldsymbol{\beta}) - y_i(\boldsymbol{x}_i^T\boldsymbol{\beta}) \end{aligned} \tag{6}$$

Furthermore, the correlation between these two estimators will be defined as follows:

$$\frac{\hat{\boldsymbol{\beta}}^T\tilde{\boldsymbol{\beta}}}{\|\hat{\boldsymbol{\beta}}\|\|\tilde{\boldsymbol{\beta}}\|} \tag{7}$$

In this paper, we aim to show that the correlation can be predicted by η, a scalar resulting from analysis for

the same setting in linear regression. η is given as follows:

$$\eta = c\frac{\mathbb{E}[\rho'(\text{Prox}_{\gamma\rho(\cdot)}(\sigma G))\rho'(\text{Prox}_{\gamma\rho(\cdot)}(\sigma\tilde{G}))]}{\mathbb{E}[(\rho'(\text{Prox}_{\gamma\rho(\cdot)}(\sigma G)))^2]} \tag{8}$$

for $(G, \tilde{G})$ distributed by the bivariate normal with mean $\begin{pmatrix} 0 \\ 0 \end{pmatrix}$ and covariance matrix $\begin{pmatrix} 1 & \eta \\ \eta & 1 \end{pmatrix}$. As previously described, the solution for η can be found after solving the two equation system (5) for (σ, γ). It is important to note also that values of $\eta \in [0, 1]$ if $I = \tilde{I}$ or $I \cap \tilde{I} = \emptyset$.

3 Methodology and Results

In order to demonstrate that η predicts the correlation, the solution for η was graphed against a scatter plot of the values of equation (7) over δ for fixed values of n and p.

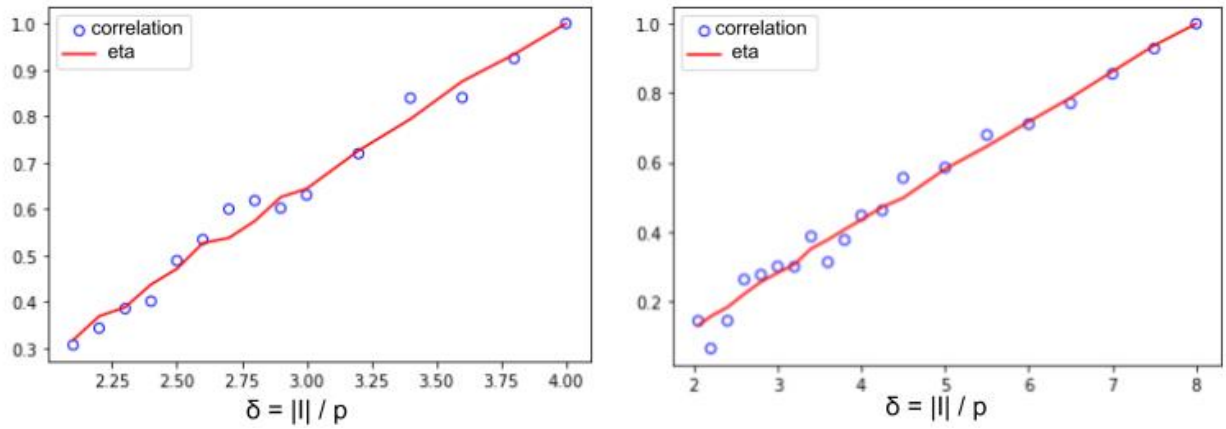
In order to plot these empirical results, the $\boldsymbol{x_i}$ and y_i pairs were randomly generated and then split according to the specifications above for various values of c. In particular, the first cn observations were taken to be in subset I and the next cn with nc^2 overlap were taken to be in subset $\tilde{I}$. This left the last $n(c-1)^2$ observations unused. Then, using the sklearn logistic regression classifying algorithm in [1], the MLE for the model parameters were found for each of the subsets. From this, the correlation was calculated and then plotted as a scatter plot over the δ's corresponding to the values of c.

To solve for η, initially a black box equation solver was used to retrieve a solution for the two equation system in (5). This returned values for σ and γ in terms of δ values. Due to limitations in calculating the MLE, only values of $\delta > 2$ were considered.

Afterward, an approximate solution for the equation for η was found using an iterative method. Essentially, across a grid of values of δ and their corresponding σ and γ values, the value of η was found, calculating G, $\tilde{G}$ with a guess value of η. Then this value of η was used to calculate the values for G, $\tilde{G}$ in the next iteration. This was repeated until the values of η between iterations converged to be within some value ϵ of one another.

In order to calculate η, several approximations were made. The approximate value for the proximal operator $\text{Prox}_{t\rho(\cdot)}$ was found using Newton's method. Next, to find an approximate value for the expectations in equation (8), the law of large numbers was used. Specifically, a simple average of the same quantity within the expectations in equation (8) was taken but over 10,000 trials of G and $\tilde{G}$.

Then, the values of η were plotted over the δ values. Using the procedure described, for $p = 250$ as well as $n = 1000$ and $n = 2000$ respectively, the following figures were produced.



(a) $n = 1000$ (b) $n = 2000$

Figure 1: The blue points represent empirical calculations for correlation between the two MLE estimates for the parameters of subsets I and $\tilde{I}$. The red line represents η over values of δ defined as $\frac{|I|}{p}$. In both figures, $p = 250$.

4 Conclusion and Future work

This paper investigated a method to characterize the performance of a pair of unregularized maximum likelihood estimates for a logistic regression model with bagging. As a result, we were able to use results from [4] and the same setting in linear regression to show a single scalar η that seems to predict the correlation of the two parameter estimates for the bootstrapped subsets.

While these results demonstrate the fit between η and the correlation of the two parameters, there is not yet a rigorous calculation explaining why this result appears. In the future, it may be of interest to derive such a result. One method to do so may be to use the replica method from statistical physics. Previously, [3] were able to come to a similar result on the performance of high dimensional regularized logistic regression models as in [4] using the replica trick. This may be one potential method through which an explanatory calculation may be found.

References

[1] Matthias Feurer, Katharina Eggensperger, Stefan Falkner, Marius Lindauer, Frank Hutter. Auto-sklearn 2.0: hands-free AutoML via meta-learning. *The Journal of Machine Learning Research* Volume 23, Issue 1, Article No.: 261 pp 11936–11996, 2022.

[2] D Irvani, K Sadik, A Kurnia, A Saefuddin and Erfiani. Swing Voters' Vote Choice Prediction Using Multilevel Logit Model to Improve Election Survey Accuracy. *Journal of Physics: Conference Series 1863*, 2021. DOI: 10.1088/1742-6596/1863/1/012021

[3] Bruno Loureiro, Cédric Gerbelot, Hugo Cui, Sebastian Goldt, Florent Krzakala, Marc Mézard, and Lenka Zdeborová. Learning curves of generic features maps for realistic data sets with a teacher-student model. *Journal of Statistical Mechanics: Theory and Experiment*, 2022. DOI 10.1088/1742-5468/ac9825

[4] Fariborz Salehi, Ehsan Abbasi, and Babak Hassibi. The Impact of Regularization on High-dimensional Logistic Regression. *arXiv preprint arXiv:1906.0376*, 2019

[5] Pragya Sur and Emmanuel J Candès. A modern maximum-likelihood theory for high-dimensional logistic regression. *arXiv preprint arXiv:1803.06964*, 2018

Acknowledgement

Thank you to my mentor Pierre C. Bellec for his continued support, guidance, and patience. This work was carried out as a part of the 2023 DIMACS REU program at Rutgers University, supported by NSF grant CNS-2150186.