Machine Learning Project Proposal

Comparison of Robust Logistic Regression with Sklearn Regularized Logistic Regression implementation

Denis Sai dsai@mit.edu, Peijun Xu xup@mit.edu

October 28, 2020

Problem Summary and Motivation

Nowadays, it is already known that adding regularization achieves robustness in features or labels, but is not directly a sparse technique. For instance, robust counterpart to robustifying problem with uncertainty set in the features is (Bertsimas et al., 2019):

$$max_{\beta,\beta_0} - \sum_{i=1}^{n} log(1 + e^{-y_i(\beta^T x_i + \beta_0) + p||\beta||_{q^*}})$$

Where L_{q^*} is the dual norm of L_q.

However, in the most popular Machine Learning toolkit, Scikit Learn, regularization for logistic regression is still being added in a form similar to linear regression using minimization of regularized negative log-likelihood (Yu et al., 2011):

$$min_w \sum_{i=1}^{n} log(1 + e^{-y_i w^T x_i}) + \frac{1}{2\sigma} w^T w$$

The difference between the two approaches is contained in the setting of the regularization in different places of the loss function, which leads to a non-robust formulation in the case of Scikit Learn. In our work, we would like to assess if changing the position of the regularization leads to improved classification metrics (ROC AUC, accuracy). Thus, if it is possible to notice a difference in the quality classification metrics, it would make sense to discuss with the creators of Scikit Learn possible changes to this part of the library.

In addition, Scikit Learn currently does not support robustness in labels, which can also have a positive effect on the classification quality metrics compared to the current library approach, which we also want to evaluate in our work.

Data sets that we will use

To conduct a thorough comparison between performances of these different formulations of logistic regression across different datasets with different sample size n and number of attributes p. We will use the following datasets from UCI Machine Learning Repository:

- Caesarian Section Classification Dataset Data Set($n = 80$, $p = 6$)
- Qualitative Bankruptcy Data Set($n = 259$, $p = 7$)
- Credit Approval Data Set($n = 690$, $p = 15$)
- Breast Cancer Wisconsin (Diagnostic) Data Set($n = 569$, $p = 32$)
- YouTube Spam Collection Data Set($n = 1952$, $p = 5$)
- default of credit card clients Data Set ($n = 30000$, $p = 24$)

Methods that we will use and how these relate to our class

We will use 3 models:

- Scikit Learn Logit with regularization term as a baseline model
- Robust Logistic regression with uncertainty set in features **(implement model from our course)**
- Robust Logistic regression with uncertainty set in labels **(model from our course)(implement model from our course)**
- Robust Logistic regression with both uncertainty sets in features and labels **(implement model from our course)**

Challenges / ideas to overcome them

A key challenge is related to the uncertainty of benefits of state-of-the art robust approaches. In order to tackle that problem we will use new ideas (robustness in labels) so that we will be able to compare not only different placement of the same regularization term, but also add fundamentally new one.

Another challenge is randomness in modeling training, which leads to uncertainty in final evaluation of models' performances. Therefore, we can apply our models to bootstrapped data to generate confidence intervals for performances. By comparing confidence intervals, we can reduce the uncertainty in performance evaluation.

Lastly, we may have missing values in our datasets. We can solve that problem by applying the same preprocessing imputation pipeline for all the models.

References

Bertsimas, D., J. Dunn, C. Pawlowski, and Y. D. Zhuo (2019). Robust classification. *INFORMS Journal on Optimization 1*(1), 2–34.

Yu, H.-F., F.-L. Huang, and C.-J. Lin (2011, October). Dual coordinate descent methods for logistic regression and maximum entropy models. *Mach. Learn. 85*(1–2), 41–75.