On Coresets for Logistic Regression

Presentation of work originally published in Advances in Neural Information Processing Systems 31, NeurIPS 2018, [Mu18]

Alexander Munteanu[1], Chris Schwiegelshohn[2], Christian Sohler[3], David P. Woodruff[4]

Abstract: Coresets are one of the central methods to facilitate the analysis of large data. We continue a recent line of research applying the theory of coresets to logistic regression. First, we show the negative result that no strongly sublinear sized coresets exist for logistic regression. To deal with intractable worst-case instances we introduce a complexity measure $\mu(X)$, which quantifies the hardness of compressing a data set for logistic regression. $\mu(X)$ has an intuitive statistical interpretation that may be of independent interest. For data sets with bounded $\mu(X)$-complexity, we show that a novel sensitivity sampling scheme produces the first provably sublinear $(1 \pm \varepsilon)$-coreset.

Keywords: logistic regression, coresets, lower bounds, beyond worst-case analysis

1 Introduction

Scalability is one of the central challenges of modern data analysis and machine learning. Algorithms with polynomial running time might be regarded as efficient in a conventional sense, but nevertheless become intractable when facing massive data sets. As a result, performing data reduction techniques in a preprocessing step to speed up a subsequent optimization problem has received considerable attention. A natural approach is to subsample the data according to a certain probability distribution. In this paper we focus on the logistic regression problem which is an instance of a generalized linear model. We are given data $Z \in \mathbb{R}^{n \times d}$, and labels $Y \in \{-1, 1\}^n$. The optimization task consists of minimizing the negative log-likelihood $\sum_{i=1}^{n} \ln(1 + \exp(-Y_i Z_i \beta))$ with respect to the parameter $\beta \in \mathbb{R}^d$. To tackle scalability issues for logistic regression via sub-sampling we choose a probability distribution based on the *sensitivity* score of each point. Informally, the sensitivity of a point corresponds to the worst-case contribution of the point to the objective function we wish to minimize. If the total sensitivity, i.e., the sum of all sensitivity scores, is bounded by a reasonably small value, there exists a small collection of input points known as a *coreset* with very strong aggregation properties. For any solution $\beta \in \mathbb{R}^d$, the objective function evaluates on the coreset as on the original data up to a small multiplicative error [MS18].

[1] TU Dortmund, Data Science Center, 44221 Dortmund, Germany, alexander.munteanu@tu-dortmund.de
[2] Sapienza University of Rome, CS Department, 00185 Rome, Italy, schwiegelshohn@diag.uniroma1.it
[3] TU Dortmund, CS Department, 44221 Dortmund, Germany, christian.sohler@tu-dortmund.de
[4] Carnegie Mellon University, CS Department, Pittsburgh, PA 15213, USA, dwoodruf@cs.cmu.edu

doi:10.18420/inf2019_37

2 Our contributions

We show that logistic regression has no sublinear streaming algorithm. Due to a standard reduction between coresets and streaming algorithms, this implies that logistic regression admits no sublinear coresets or bounded sensitivity scores in general.

We investigate available sensitivity sampling distributions for logistic regression. For points with large contribution, where $-Y_i Z_i \beta \gg 0$, the objective function increases by a term almost linear in $-Y_i Z_i \beta$. This motivates to use sensitivity scores designed for ℓ_1-related problems. To this end, we propose sampling from a mixture distribution with one component proportional to the *square root* of the ℓ_2^2 leverage scores. The other mixture component is uniform sampling to deal with the remaining domain. Our experiments show that this distribution outperforms uniform and k-means based sensitivity sampling by a wide margin on real data sets. The algorithm is space efficient, and can be implemented in a variety of models used to handle large data sets such as 2-pass streaming, and massively parallel frameworks such as Hadoop and MapReduce, and can be implemented to work in input sparsity time, i.e., proportional to the number of non-zero entries of the data [Wo14].

We analyze our sampling distribution for a parametrized class of instances we call μ-complex, placing our work in the framework of *beyond worst-case analysis* [Ro19]. The parameter μ roughly corresponds to the ratio between the log of correctly estimated odds and the log of incorrectly estimated odds. The condition of small μ is justified by the fact that for instances with large μ, logistic regression exhibits methodological problems. We show that the total sensitivity of logistic regression can be bounded in terms of μ. Moreover, if the data is μ-complex for a small, not necessarily constant μ, then there exists a sampling and reweighting scheme based on the sensitivity framework that produces a $(1 \pm \varepsilon)$-coreset of sublinear size $O(\varepsilon^{-2}\mu\sqrt{n}d^{3/2}\log^2(\mu nd))$ with high probability. A more involved recursive sampling scheme produces a $(1 \pm \varepsilon)$-coreset of size $O(\varepsilon^{-4}\mu^3 d^3 \log^{O(1)}(\mu nd))$, which is beneficial if the data is well-behaved and the input size is particularly large. These are the first provably sublinear coreset constructions for logistic regression.

Bibliography

[MS18] Munteanu, Alexander; Schwiegelshohn, Chris: Coresets-Methods and History: A Theoreticians Design Pattern for Approximation and Streaming. KI, 32(1):37–53, 2018.

[Mu18] Munteanu, Alexander; Schwiegelshohn, Chris; Sohler, Christian; Woodruff, David P.: On Coresets for Logistic Regression. In: Advances in Neural Information Processing Systems (NeurIPS). pp. 6562–6571, 2018.

[Ro19] Roughgarden, Tim: Beyond worst-case analysis. Commun. ACM, 62(3):88–96, 2019.

[Wo14] Woodruff, David P.: Sketching as a Tool for Numerical Linear Algebra. Foundations and Trends in Theoretical Computer Science, 10(1-2):1–157, 2014.