Large-Scale L1-Regularized and L2-Regularized Logistic Regression via a Stale Synchronous Parallel Parameter Server

Huanchen Zhang (huanche1@andrew.cmu.edu, huanche1)
Yetian Xia (yetianx@andrew.cmu.edu, yetianx)

Motivation

L1-regularized and L2-regularized logistic regression are machine learning models that underlie many important applications in industry. For example, Google [2] and Microsoft [1] have constructed their ad serving/ad prediction systems based on such models due to their good performance. Despite that large-scale regularized logistic regression are widely used in proprietary systems, there is a lack of open source implementation. This project aims to fill in this need.

Prior Work

We plan to use Petuum [4], an open source large-scale ML framework currently being developed at CMU, as the parameter/variable server for our implementation of the models. Petuum is a scalable distributed system designed in particular for iterative ML algorithms [4] in that it follows a Stale Synchronous Parallel (SSP) model of computation instead of the traditional Bulk Synchronous Parallel (BSP) model to minimize the synchronization cost for each computational worker while guarantee correctness [3]. We will leverage this cutting-edge technology in our project for better performance and scalability of our large-scale regularized logistic regression implementation.

Our Contribution

First, we will implement the L1-regularized and L2-regularized logistic regression models using Petuum. In terms of algorithms, we plan to use stochastic gradient descent for L2-regularized logistic regression and subgradient method for L1-regularized logistic regression. We would like to try our implementation on public clusters for both text classification and image classification if time and resource permit.

Second, since Petuum is not a mature system (in fact, it is still under development), we may have to put effort on configuring the system or even modify its source code for our purposes. Nonetheless, our experience with Petuum might be valuable for developing a better ML-specific distributed system.

Preliminary Plan

Phase 1 (2/17-2/23): Study L1-regularized and L2-regularized logistic regression models and algorithms.

Phase 2 (2/24-3/9): Plug and Play with Petuum, and briefly study Petuum's source code

Phase 3 (3/10-3/24): Model implementation

---Project Midterm---

Phase 4 (3/26-4/13): System testing and debugging

Phase 5 (4/14-4/20): Run classification tasks on single machine (simulating multi-worker with multi-thread on multi-core)
Phase 6 (4/21-5/1): If time and resource permit, we would like to run large scale classification tasks on a public cluster to measure our classification accuracy and system performance.

Since we currently only have two people in the group, we will still consider ourselves successful if Phase 6 cannot be carried out for some reason.

Resources Required

1. Petuum's source code: https://github.com/sailinglab/petuum
2. Data set: http://archive.ics.uci.edu/ml/datasets/Reuters-21578+Text+Categorization+Collection for text classification, and http://image-net.org/challenges/LSVRC/2014/index for image classification.
3. Access to a public cluster (e.g. Amazon EC2) to carry out large-scale experiments.

Reference

[1] Richardson, M., Dominowska, E., Ragno, R. (2007). Predicting Clicks: Estimating the Click-Through Rate for New Ads. In *Proceedings of the 16th international conference on World Wide Web (WWW'07).*

[2] Brendan McMahan, H. et al. (2013) Ad Click Prediction: a view from the Trenches. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD'13).*

[3] Ho, Q. et al. (2013). More Effective Distributed ML via a Stale Synchronous parallel Parameter Server. In *Proceedings of NIPS'13*.

[4] http://petuum.org