Distributed linear regression in high dimensions

Yue Sheng

University of Pennsylvania
E-mail: yuesheng@sas.upenn.edu

Abstract: Distributed statistical learning problems arise commonly when dealing with large datasets. In this setup, datasets are partitioned over machines, which compute locally and communicate short messages. Communication is often the bottleneck. In this paper, we study one-step and iterative
weighted parameter averaging in statistical linear models under data parallelism. We do linear regression on each machine, send the results to a central server, and take a weighted average of the parameters. Optionally, we iterate, sending back the weighted average and doing local ridge regressions centered at it. How does this work compare to doing linear regression on the full data? Here we study the performance loss in estimation and test error, and confidence interval length in high dimensions, where the number of parameters is comparable to the training data size.
We find the performance loss in one-step weighted averaging, and also give results for iterative averaging. We also find that different problems are affected differently by the distributed framework. Estimation error and confidence interval length increases a lot, while the prediction error increases much less.