DYNA-Q in the mountain car, cart-pole, and planar arm tasks. In the case of the acrobot swing up task, where we could not find an accurate model for the domain, MBOA still outperforms all other algorithms. Empirically MBOA outperforms all of the alternatives in terms of data efficiency. This efficiency is gained at the cost of additional computation time for the simulations, which generate sample trajectories of candidate policies during optimization of the expected improvement. Overall, MBOA appears to be a useful step toward combining model-based methods with Bayesian Optimization for purposes of handling inaccurate models and improving data efficiency.

Acknowledgements

We gratefully acknowledge the support of the Army Research Office under grant number W911NF-09-1-0153 and the NSF under grant number IIS-0905678.

References

Efficient and Numerically Stable Sparse Learning

Sihong Xie\textsuperscript{1}, Wei Fan\textsuperscript{2}, Olivier Verscheure\textsuperscript{2}, and Jiangtao Ren\textsuperscript{1}

\begin{itemize}
\item \textsuperscript{1} Sun Yat-Sen University, Guangzhou, China
  \{xiesihong1,issrjt\}@gmail.com
\item \textsuperscript{2} IBM T.J. Watson Research Center, New York, USA
  \{weifan,ov1\}@us.ibm.com
\end{itemize}

Abstract. We consider the problem of numerical stability and model density growth when training a sparse linear model from massive data. We focus on scalable algorithms that optimize certain loss function using gradient descent, with either $\ell_0$ or $\ell_1$ regularization. We observed numerical stability problems in several existing methods, leading to divergence and low accuracy. In addition, these methods typically have weak controls over sparsity, such that model density grows faster than necessary. We propose a framework to address the above problems. First, the update rule is numerically stable with convergence guarantee and results in more reasonable models. Second, besides $\ell_1$ regularization, it exploits the sparsity of data distribution and achieves a higher degree of sparsity with a PAC generalization error bound. Lastly, it is parallelizable and suitable for training large margin classifiers on huge datasets. Experiments show that the proposed method converges consistently and outperforms other baselines using 10% of features by as much as 6% reduction in error rate on average. Datasets and software are available from the authors.

1 Introduction

In this paper, we focus on training a sparse large margin model. Assume that we are given $m$ labeled examples $Z = \{(x_1, y_1), \ldots, (x_m, y_m)\}$, where $x_i \in \mathbb{R}^d, i = 1, \ldots, m$. We aim at solving the following optimization problem:

$$\min_{w \in \mathbb{R}^d} \frac{1}{m} \sum_{i=1}^{m} L(\langle w, x_i \rangle, y_i) + \lambda \|w\|_1$$

$L$ is any smooth and differentiable loss function such as logistic or hinge loss. $\lambda$ is the parameter for trading off between loss and $\ell_1$ regularization. We are interested in scalable algorithm, with parallelizable data accesses and small communication cost. One can find such application in text mining and webspam detection, where the number of features could be in millions.

Sparse learning aims at accurate models using a small number of non-zero elements, with the advantages of efficiency and generalizability [13]. Some existing methods produce sparse models by forward-backward feature selection [13].