MLib Project Proposal

Damian Mastylo

Introduction

Apache Spark is an open-source platform commonly used for massive distributed tasks and data processing. It is a direct competitor to Hadoop, and is a relatively new framework for distributed tasks. Meng, et. al. presents MLib: a machine learning library on top of Apache Spark that includes several primitives to allow for quick usage and deployment. There are APIs in several languages to make the platform very accessible.

Proposal

I would like to test performance/efficiencies of machine learning processes on Spark, Hadoop, and single machines to gauge their effectiveness among multiple data sets. I would most likely use an SVM and basic classification data sets to test these methods.

Previous Work

In the paper, Meng, et.al. briefly describes the efficiency of MLib, but not does delve deep into the results or provide a large amount of results. I would like to present results across various dataset and various setups to really hammer down what is the best library/framework to use for distributed machine learning.

Benefit Distributed Computing

An empirical and practical view of these results will benefit anyone looking to machine learn extremely large data sets that require more processing power than a basic computer or server.

A good benchmark with little confounding variables can provide a lot of guidance for those looking to get into this field.

References

[1] Xiangrui Meng and (2015). MLlib: Machine Learning in Apache Spark. *CoRR, abs/1505.06807*, .