Machine Learning Algorithms in Spark

Abhishek Malali (abhishekmalali@g.harvard.edu)
Neil Chainani (chainani@g.harvard.edu)
Leonhard Spiegelberg (spiegelberg@g.harvard.edu)

November 4, 2015

1 Background

Spark has a machine learning library called MLib which has a missing Neural networks functionality which serves as our main inspiration to implement additional ML algorithms in Spark.

2 Objectives - Functionality and Performance

To implement the machine learning algorithms in Spark while benchmarking the code against Scikit-learn and MLlib which is the standard Machine learning library for Spark. A part of the project will also focus on fitting the data to the problem in the best possible way while ensuring all standard ML rules are followed.

3 Milestones

- 11/13 - Research and benchmark implementations(mord/scikit-learn)
- 11/20 - Preliminary implementations in Spark for ordinal regressions
- 11/27 - Parallelizing regression code and work on bayesian ordinal regressions
- 12/4 - Final analysis of current work and start work on project webpage

4 Division of work

- Leonhard - Design of algorithms for Ordinal regressions
- Neil - Implementation of benchmark code and implementation of algorithms in Spark
- Abhishek - Implementing algorithms in Spark and Parallelizing strategies