Machine Learning Algorithms in Spark

Abhishek Malali (abhishekmalali@g.harvard.edu)
Neil Chainani (chainani@g.harvard.edu)
Leonhard Speilberg (email)

November 4, 2015

1 Background

Spark has a machine learning library called MLib which has a missing Neural networks functionality which was our inspiration to write ML algorithms in Spark.

2 Benchmarking Data

We intend to use the (small dataset) for initial testing which we would later scale up to (large dataset) to benchmark spark implementations against scikit-learn.

3 Objectives - Functionality and Performance

4 Algorithms to be implemented

- Artificial Neural Networks - ANNs are highly paralellizable and plan to implement a basic implementation which works on Backpropagation for starters which can be extended to other learning algorithms.
- Random Forests - Random forests are paralellizable since we need to create different classification trees which can be done on multiple processors without conflict and later integrated to find results.
- Add one more ML algorithm

5 Design Overview - Technologies and use of Parallelism

6 Verification

7 Schedule, Milestones, and Division of Work