Mahout: A Scalable Machine Learning and Data Mining library

Jeff Avery
Dependable Computing Systems Lab (DCSL)
Purdue University

PURDUE UNIVERSITY

Outline

- Background
- Uses
- Algorithms
- Strengths
- Example Commands
- Weakness
- Resources
- Questions

PURDUE UNIVERSITY

Background

- "Scalable machine learning and data mining"
- Distributed under the Apache software license
  - MapR Technologies
- Based on Hadoop
- Used in a number of areas
  - Amazon's personalization platform
  - Foursquare
  - Booz Allen Hamilton
  - A number of academic projects and courses also use Mahout

Uses of Mahout

- Clustering
- Classification
- Recommendation mining
  - i.e. recommended purchases on Amazon
- Frequent item set mining
  - i.e. related purchases on Best Buy

Algorithms Implemented

- Classification
  - Logistic Regression (SGD)
  - Bayesian
  - Random Forests
- Clustering
  - Reference Reading
  - K-means
- Pattern Mining
  - Parallel Frequent Pattern Mining

Strengths

- Very easy learning curve
  - Once installed running experiments and learning commands are simple
- Very simple commands
  - trainlogistic – trains clustering engine based on other input values/files
  - runlogistic – runs the newly trained engines on new data that is also provided
- Feedback from experiments are easy to understand
- Plenty of documentation and forums on the web

Example Commands

- Example of training a model
  - bin/mahout trainlogistic --input donut.csv \ --output ./model \ --target color --categories 2 \ --predictors x y --types numeric \ --features 20 --passes 100 --rate 50
- Example of testing a model
  - bin/mahout runlogistic --input donut.csv --model ./model\ --auc --confusion

Weakness

- Installation is not trivial
  - Setting up a single node Hadoop cluster time consuming
    - http://nivirao.blogspot.com/2012/04/installing-apache-mahout-on-ubuntu.html
- Not all examples seem to work
  - Data files don't seem to be included in .tar file
- The MapR virtual machine doesn't work correctly
  - Errors about space constraints

Resources

- Textbook: Mahout In Action
  - Available in PDF form online
- Main website: mahout.apache.org
  - Download .tar files
  - Sign up for mailing lists
  - Instructions for quick start and running examples
- MapR documentation
  - MapR is the company that creates Hadoop-derived products
- Blogs, stackoverflow.com, etc.

Questions, Comments