Mammoth Scale Machine Learning

Speaker: Robin Anil, Apache Mahout PMC Member

OSCON’10
Portland, OR
July 2010
Quick Show of Hands

- Are you fascinated about ML?
- Have you used ML?
- Do you have Gigabytes and Terabytes of data to analyze?
- Do you have Hadoop or MapReduce experience?

- Thanks for the survey!
Little bit about me

- Apache Mahout PMC member
- A ML Enthusiast 😊
- Software Engineer @ Google
- Google Summer of Code Mentor
- Previous Life: Google Summer of Code student for 2 years.
Agenda

- Introducing Mahout
- Different classes of problems
- And their Mahout based solutions
- Basic data structure
- Usage examples
- Sneak peek at our Next Release
The Mission

To build a *scalable* machine learning library
Scale!

- Scale to large datasets
  - Hadoop MapReduce implementations that scales linearly with data.
  - Fast sequential algorithms whose runtime doesn’t depend on the size of the data
  - Goal: To be as fast as possible for any algorithm
- Scalable to support your business case
  - Apache Software License 2
- Scalable community
  - Vibrant, responsive and diverse
  - Come to the mailing list and find out more
The Mission

To build a scalable machine learning library
Why a new Library

- Plenty of open source Machine Learning libraries either
  - Lack community
  - Lack scalability
  - Lack documentations and examples
  - Lack Apache licensing
  - Are not well tested
  - Are Research oriented
Agenda

- Introducing Mahout
- Different classes of problems
- And their Mahout based solutions
- Basic data structure
- Usage examples
- Sneak peek at our Next Release
ML on Twitter

- Collection of tweets in the last hour
- Each 140 character or token stream
- We will keep using this example throughout this talk
What is Clustering

- Call it fuzzy grouping based on a notion of similarity
Mahout Clustering

- Plenty of Algorithms: K-Means, Fuzzy K-Means, Mean Shift, Canopy, Dirichlet

- Group similar looking objects

- Notion of similarity: Distance measure:
  - Euclidean
  - Cosine
  - Tanimoto
  - Manhattan
Clustering Tweets

“Identify tweets that are similar and group them”
Topic modeling

- Grouping similar or co-occurring features into a topic
  - Topic “Lol Cat”:
    - Cat
    - Meow
    - Purr
    - Haz
    - Cheeseburger
    - Lol
Mahout Topic Modeling

- Algorithm: Latent Dirichlet Allocation
  - Input a set of documents
  - Output top K prominent topics and the features in each topic
Filtering Topics from Tweets

“Identify emerging topics in a collection of tweets”
Classification

- Predicting the type of a new object based on its features
- The types are predetermined

Dog                   Cat
Mahout Classification

- Plenty of algorithms
  - Naïve Bayes
  - Complementary Naïve Bayes
  - Random Forests
  - Logistic Regression (Almost done)
  - Support Vector Machines (patch ready)

- Learn a model from a manually classified data
- Predict the class of a new object based on its features and the learned model
Detect OSCON Tweets

“Tweets without #OSCON”

Use tweets mentioning #OSCON to train and Classify incoming tweets
Recommendations

- Predict what the user likes based on
  - His/Her historical behavior
  - Aggregate behavior of people similar to him

Customers Who Bought This Item Also Bought

- **Pattern Recognition and Machine Learning** by Christopher M. Bishop
  - Rating: ★★★★☆ (50)
  - Price: $76.10

- **The Elements of Statistical Learning: Data Mining, Inference, and Prediction** by Trevor Hastie
  - Rating: ★★★★★ (38)
  - Price: $71.95

- **Pattern Classification (2nd Edition)** by Richard O. Duda
  - Rating: ★★★★★ (29)
  - Price: $88.52
Mahout Recommenders

- Different types of recommenders
  - User based
  - Item based
- Full framework for storage, online and offline computation of recommendations
- Like clustering, there is a notion of similarity in users or items
  - Cosine, Tanimoto, Pearson and LLR
Recommended Tweets

“Discover interesting tweets without Re-Tweeting or Replying”
Frequent Pattern Mining

- Find interesting groups of items based on how they co-occur in a dataset
Mahout Parallel FP Growth

- Identify the most commonly occurring patterns from
  - Sales Transactions
    buy “Milk, eggs and bread”
  - Query Logs
    ipad -> apple, tablet, iphone
  - Spam Detection
    Yahoo! http://www.slideshare.net/hadoopusergroup/mail-antispam
Frequent patterns in Tweets

“Identify groups of words that occur together”
Or
“Identify related searches from search logs”
Mahout is Evolving

- Mapreduce enabled fitness functions for Genetic programming
  - Integration with Watchmaker
  - Solves: Travelling salesman, class discovery and many others
- Singular Value decomposition [SVD] of large matrices
  - Reduce a large matrix into a smaller one by identifying the key rows and columns and discarding the others
  - Mapreduce implementation of Lanczos algorithm
Agenda

- Introducing Mahout
- Different classes of problems
- And their Mahout based solutions
- Basic data structure
- Usage examples
- Sneak peek at our Next Release
Vector
The vector denoted by point (5, 3) is simply
Array([5, 3]) or HashMap([0 => 5], [1 => 3])
Representing Vectors – The basics

- Now think 3, 4, 5, ….. n-dimensional
- Think of a document as a bag of words.
  “she sells sea shells on the sea shore”
- Now map them to integers
  
  she => 0
  sells => 1
  sea => 2
  and so on
- The resulting vector [1.0, 1.0, 2.0, … ]
Vectorizer tools

- Map/Reduce tools to convert text data to vectors
  - Use collate multiple words (n-grams) eg: “San Francisco”
  - Normalization
  - Optimize for sequential or random access
  - TF-IDF calculation
  - Pruning
  - Stop words removal
Agenda

- Introducing Mahout
- Different classes of problems
- And their Mahout-based solutions
- Basic data structure
- Usage examples
- Sneak peek at our Next Release
How to use mahout

- Command line launcher `bin/mahout`
- See the list of tools and algorithms by running `bin/mahout`
- Run any algorithm by its shortname:
  - `bin/mahout kmeans --help`
- By default runs locally
- `export HADOOP_HOME = /path/to/hadoop-0.20.2/`
  - Runs on the cluster configured as per the conf files in the hadoop directory
- Use driver classes to launch jobs:
  - `KMeansDriver.runjob(Path input, Path output ...)"
Clustering Walkthrough (tiny example)

- Input: set of text files in a directory
- Download Mahout and unzip
  - mvn install
  - bin/mahout seqdirectory –i <input> –o <seq-output>
  - bin/mahout seq2sparse –i seq-output –o <vector-output>
  - bin/mahout kmeans –i<vector-output> –c <cluster-temp> –o <cluster-output> –k 10 –cd 0.01 –x 20
Clustering Walkthrough (a bit more)

- Use bigrams: `-ng 2`
- Prune low frequency: `--s 10`
- Normalize: `--n 2`

- Use a distance measure: `--dm org.apache.mahout.common.distance.CosineDistanceMeasure`
Clustering Walkthrough (viewing results)

- bin/mahout clusterdump
  
  `-s cluster-output/clusters-9/part-00000
  `-d vector-output/dictionary.file-*
  `-dt sequencefile -n 5 -b 100

- Top terms in a typical cluster

  comic           => 9.793121272867376
  comics          => 6.115341078151356
  con             => 5.015090566692931
  sdcc            => 3.927590843402978
  webcomics       => 2.916910980686997
Agenda

- Introducing Mahout
- Different classes of problems
- And their Mahout-based solutions
- Basic data structure
- Usage examples
- Sneak peek at our Next Release
Mahout 0.4 (trunk)

- New breed of classifiers:
  - Stochastic Gradient Descent (SGD)
  - Pegasos SVM (Order of magnitude faster than SVM Perf)
  - Lib Linear (Winner, ICML 2008)

- New Recommenders:
  - Restricted Boltzmann Machine (RBM) based recommender
  - SVD++ recommender

- New Clustering algorithms:
  - Spectral Clustering
  - K-Means++

- Full Hadoop 0.20 API compliance and performance improvements
Get Started

- [http://mahout.apache.org](http://mahout.apache.org)
- [dev@mahout.apache.org](mailto:dev@mahout.apache.org) - Developer mailing list
- [user@mahout.apache.org](mailto:user@mahout.apache.org) - User mailing list
- Check out the documentations and wiki for quickstart
Resources

- “Mahout in Action” Owen, Anil, Dunning, Friedman
  [http://www.manning.com/owen](http://www.manning.com/owen)

- “Taming Text” Ingersoll, Morton, Farris
  [http://www.manning.com/ingersoll](http://www.manning.com/ingersoll)

- “Introducing Apache Mahout”
Thanks to

- Apache Foundation
- Mahout Committers
- Google Summer of Code Organizers
- And Students
- OSCON
- Open source!
References

- news.google.com
- Dog [http://www.flickr.com/photos/30800139@N04/3879737638/](http://www.flickr.com/photos/30800139@N04/3879737638/)
- Amazon Recommendations
- twitter