Apache Mahout

Scaling Machine Learning

Presented by:
Isabel Drost
Agenda

• Motivation.

• Machine learning?

• Introducing Mahout.

• How can you help?
Some motivation.
Follow news stories

September 10, 2008 by Alex Barth
http://www.flickr.com/photos/a-barth/2846621384

Search through papers.  Automatic topic tracker.
Movie recommendation

IMDB + movie reviews. Aggregate reviews from IMDB, twitter, ...
• Lots and lots of data.

• Structured and unstructured.
Mission

Provide scalable data mining algorithms.
Machine Learning?
Archimedes generates model:

\[
\frac{\text{Density of Object}}{\text{Density of Fluid}} = .
\]

\[
\frac{\text{Weight}}{\text{Weight} - \text{Apparent immersed weight}}
\]
Machine learning generates model

class $+1$ $w^*x + b > 1$

class $-1$ $w^*x + b < -1$

separating hyperplane $w^*x + b = 0$
Machine learning pipeline

1. Gather data. (and meta data).
2. Identify characteristics.
3. Chose right algorithm.
4. Train on the gathered data.
5. Tune parameters of your algorithm.
6. Keep model in sync when nature changes.
Machine learning pipeline

1. Gather data. (and meta data).
2. Identify characteristics.
3. Chose right algorithm.
4. Train on the gathered data.
5. Tune parameters of your algorithm.
6. Keep model in sync when nature changes.
E-Bay

Auction status?

Phishing Spam?

Different topic

Requested password?
One of your mails:

Apache

Hadoop

London

Lucene
Machine learning pipeline

1. Gather data. (and meta data).
2. Identify characteristics.
3. Chose right algorithm.
4. Train on the gathered data.
5. Tune parameters of your algorithm.
6. Keep model in sync when nature changes.
class +1
$w^T x + b > 1$

class -1
$w^T x + b < -1$

separating hyperplane
$w^T x + b = 0$

margin
Machine learning pipeline

Gather data. (and meta data).

Identify characteristics.

Chose right algorithm.

Train on the gathered data.

Tune parameters of your algorithm.

Keep model in sync when nature changes.
Parameter tuning

- Penalty for mistakes.
- Kernel type for data transformation.
- Tune kernel parameters.
Gather data.
(and meta data).

Identify characteristics.

Chose right algorithm.

Train on the gathered data.

Tune parameters of your algorithm.

Keep model in sync when nature changes.
• Build model from data.
Machine learning pipeline

Gather data. (and meta data).

Identify characteristics.

Chose right algorithm.

Train on the gathered data.

Tune parameters of your algorithm.

Keep model in sync when nature changes.
Nature changes?

- Spammers adapt to spam filters.
- Users write mails in different styles.
- Expand to new languages.
- ...
Machine learning pipeline

1. Gather data. (and meta data).
2. Identify characteristics.
3. Chose right algorithm.
4. Train on the gathered data.
5. Tune parameters of your algorithm.
6. Keep model in sync when nature changes.
Introducing Mahout
Classification

• Categorize data.

• Examples:
  • Identify spam mails.
  • Classify movies as “Action”, “Comedy” ...
Classification

- Naive bayes.
- Complementary naive bayes.
- Winnow/Perceptron
- Others upcoming.
Discovering groups of data

• Group data by similarity.

• Examples:
  • News articles by topic.
  • Developers by favorite modules.
Discovering groups of data

- Canopy.
- K-Means.
- Dirichlet based.
- PLSI.
- Others upcoming.
Recommendation mining

- Recommend items.
- Examples:
  - Find books a user may like.
  - Identify movies a user likes.
Upcoming

- More algorithms.
- More examples.
What Mahout can do for you

“Why should I participate?”
Jumpstart your project with proven code.
Discuss with researchers and engineers.
Become a community member.
http://.../pub/mirrors/apache/lucene/mahout/0.1/

Thank you to all those making this possible.
We need You:

- Enthusiasm.
- Mathematical knowledge.
- Proficiency in Hadoop.
- Interest in understanding data.
Some advertising

Berlin - June* at 5p.m.

newthinking store Berlin
Tucholskystr. 48

Hadoop** User/Developer Meeting Germany

* Exact date is set by speaker – that is you!

** Lucene, Tika, Solr, UIMA, Mahout, katta, ... people welcome.