

Data-driven modeling

APAM E4990

Lecture 1

Jake Hofman

Columbia University

September 11, 2009

Learning by example

Fwd: Yahoo! supercomputing cluster RFP - i have no idea. i have no idea. O
non urgent - whoops! yes that's what i meant, thanks for decoding my questi
SourceForge.net: variational bayes for network modularity - can i get admin |
Byline - iPhone Apps, iPhone 3G apps and iPod touch Applications Gallery a
Laurence J. Peter: Facts are stubborn things, but statistics are more pliable.
Re: JAFOS 2008, Applied Math Session - yes. the listening post dude. On N
Access to over 5,000 Health Plan Choices! - Affordable health insurance. Ins
More effective - If you are having trouble viewing this email click here. Thurs
Special Offer! Cialis, Viagra, VicodinES! - Order all your Favorite Rx~Medica
Financial Aid Available: Find Funding for Your Education - Get the financial a
Find The Perfect School and Financial Aid for your College Degree - HI ! It h
PHARMA_violagra_PHARMA_cialis - Wanted: web store with remedies. N

Learning by example

Fwd: Yahoo! supercomputing cluster RFP - i have no idea. i have no idea. O
non urgent - whoops! yes that's what i meant, thanks for decoding my questi
SourceForge.net: variational bayes for network modularity - can i get admin
Byline - iPhone Apps, iPhone 3G apps and iPod touch Applications Gallery a
Laurence J. Peter: Facts are stubborn things, but statistics are more pliable.
Re: JAFOS 2008, Applied Math Session - yes. the listening post dude. On N
Access to over 5,000 Health Plan Choices! - Affordable health insurance. Ins
More effective - If you are having trouble viewing this email click here. Thurs
Special Offer! Cialis, Viagra, VicodinES! - Order all your Favorite Rx~Medica
Financial Aid Available: Find Funding for Your Education - Get the financial a
Find The Perfect School and Financial Aid for your College Degree - HI ! It h
PHARMA_viagra_PHARMA_cialis - Wanted: web store with remedies. N

- How did you solve this problem?
- Can you make this process explicit (e.g. write code to do so)?

Learning by example



Learning by example



- We learn quickly from few, relatively unstructured examples ... but we don't understand *how* we accomplish this
- **Can we develop algorithms that enable machines to learn by example from large data sets?**

Common applications

- Effective/practical algorithms exist, and impact our daily lives
- Entire industries built around these techniques, e.g.:
 - Spam detection (Email)
 - Information retrieval (Search)
 - Recommendation Systems (“You might also like ...”)
 - Fraud detection (Identity theft)
 - Face recognition (Camera auto-focus)
 - Optical character recognition (Mail routing via ZIP codes)

Netflix prize

The screenshot shows the Netflix homepage for a user named Jake. At the top, there's a red navigation bar with the Netflix logo and buttons for "Browse DVDs", "Watch Instantly", "Your Queue", "Movies You'll", and "Instantly to your TV". A search bar is on the right. Below the navigation bar, there are tabs for "Suggestions (22)", "Rate Movies", "Taste Preferences", and "Movies You've Rated (208)". A large button on the right says "RATED MOVIES 208".

The main section is titled "New Suggestions for Jake" and features three movie recommendations:

- Goodbye Solo**: "Because you enjoyed: Being John Malkovich, Eternal Sunshine of the Spotless Mind, No Country for Old Men". It has a 4.5-star rating and an "Add" button.
- Rushmore**: "Because you enjoyed: The Darjeeling Limited, Lost in Translation, This Is Spinal Tap". It has a 4.5-star rating and an "Add" button.
- It's Always...in Philadelphia: Sen 1 & 2**: "Because you enjoyed: The Big Lebowski, The Office: Series 2, The Office Special". It has a 4.5-star rating and an "Add All" button.

Each recommendation includes a "Not interested" link below the rating.

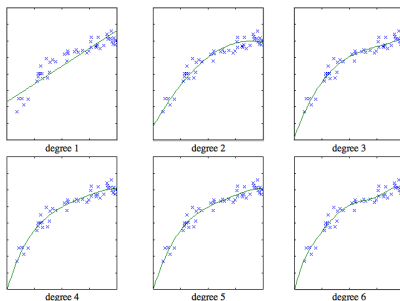
- \$1M for a 10% improvement in predicted rating
- More than 1000 submissions over 2.5 years
- Top two teams within 0.01% of each other (winners announced soon)

Goals

- Many fields ...
 - Statistics
 - Pattern recognition
 - Data mining
 - Machine learning
- ... similar goals
 - Extract and recognize patterns in data
 - Interpret or explain observations
 - Test validity of hypotheses
 - Efficiently search the space of hypotheses
 - Design efficient algorithms enabling machines to learn from data

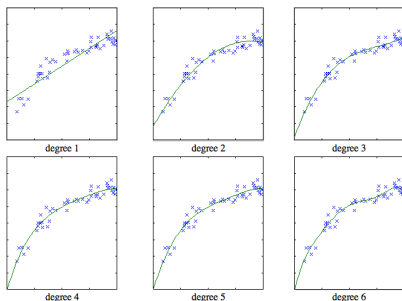
Philosophy

- We would like models that:
 - Provide predictive and explanatory power
 - Are complex enough to describe observed phenomena
 - Are simple enough to generalize to future observations



Philosophy

- We would like models that:
 - Provide predictive and explanatory power
 - Are complex enough to describe observed phenomena
 - Are simple enough to generalize to future observations



- How can we quantify an “optimal” model
 - What to optimize?
 - How to optimize it?

Framework

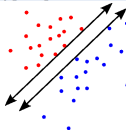
1 Get data

Fwd: Yahoo! supercomputing cluster RFP - I have no idea, I have no idea. O
non urgent - whoopee! yes that's what I meant, thanks for decoding my quest!
SourceForge.net: variational bayes for network modularity - can I get admin |
Byline - iPhone Apps, iPhone 3G apps and iPod touch Applications Gallery |
Laurence J. Peter: Facts are stubborn things, but statistics are more pliable.
Re: JAFOS 2008, Applied Math Session - yes, the listening post dude. On N
Access to over 5,000 Health Plan Choices! - Affordable health insurance. In
More effective - if you are having trouble viewing this email click here. Thru
Special Offer! Cialis, Viagra, VcodinESI - Order all your Favorite Rx-Medic
Financial Aid Available: Find Funding for Your Education - Get the financial |
Find The Perfect School and Financial Aid for your College Degree - Hi I'll h
PHARMA_viagra_PHARMA_cialis - Wanted: web store with remedies. N

Framework

- 1 Get data
- 2 Visualize/perform sanity checks
- 3 Clean/filter observations
- 4 Choose features to represent data

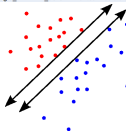
Fwd: Yahoo! supercomputing cluster RFP - I have no idea, I have no idea. O
 non urgent - whoopee! yes that's what I meant, thanks for decoding my quest
 SourceForge.net: variational bayes for network modularity - can i get admin |
 Byline - iPhone Apps, iPhone 3G apps and iPod touch Applications Gallery r
 Laurence J. Peter: Facts are stubborn things, but statistics are more pliable.
 Re: JAFOS 2008, Applied Math Session - yes, the listening post dude. On N
 Access to over 5,000 Health Plan Choices! - Affordable health insurance. In
 More effective - if you are having trouble viewing this email click here. Thurs
 Special Offer! Cialis, Viagra, ViodinESI - Order all your Favorite Rx--Medica
 Financial Aid Available: Find Funding for Your Education - Get the financial i
 Find The Perfect School and Financial Aid for your College Degree - Hi I li h
 "PHARMA_viagra_PHARMA_cialis" - Wanted: web store with remedies. N



Framework

- 1 Get data
- 2 Visualize/perform sanity checks
- 3 Clean/filter observations
- 4 Choose features to represent data
- 5 Specify model
- 6 Specify loss function

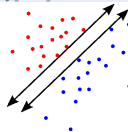
Fwd: Yahoo! supercomputing cluster RFP - I have no idea. I have no idea. O
 non urgent - whoopee! yes that's what I meant, thanks for decoding my quest
 SourceForge.net: variational bayes for network modularity - can I get admin |
 Byline - iPhone Apps, iPhone 3G apps and iPod touch Applications Gallery r
 Laurence J. Peter: Facts are stubborn things, but statistics are more pliable.
 Re: JAFOS 2008, Applied Math Session - yes, the listening post dude. On N
 Access to over 5,000 Health Plan Choices! - Affordable health insurance. In
 More effective - if you are having trouble viewing this email click here. Thru
 Special Offer! Cialis, Viagra, VcodinESI - Order all your Favorite Rx--Medica
 Financial Aid Available: Find Funding for Your Education - Get the financial i
 Find The Perfect School and Financial Aid for your College Degree - Hi I'll h
 "PHARMA_viagra_PHARMA_cialis" - Wanted: web store with remedies. N



Framework

- 1 Get data
- 2 Visualize/perform sanity checks
- 3 Clean/filter observations
- 4 Choose features to represent data
- 5 Specify model
- 6 Specify loss function
- 7 Develop algorithm to minimize loss

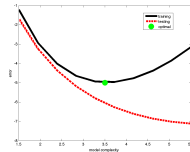
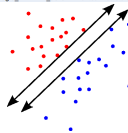
Fwd: Yahoo! supercomputing cluster RFP - I have no idea, I have no idea. O
 non urgent - whoopee! yes that's what I meant, thanks for decoding my quest
 SourceForge.net: variational bayes for network modularity - can I get admin |
 Byline - iPhone Apps, iPhone 3G apps and iPod touch Applications Gallery r
 Laurence J. Peter: Facts are stubborn things, but statistics are more pliable.
 Re: JAFOS 2008, Applied Math Session - yes, the listening post dude. On N
 Access to over 5,000 Health Plan Choices! - Affordable health insurance. In
 More effective - if you are having trouble viewing this email click here. Thurs
 Special Offer! Cialis, Viagra, VcodinESI - Order all your Favorite Rx--Medica
 Financial Aid Available: Find Funding for Your Education - Get the financial i
 Find The Perfect School and Financial Aid for your College Degree - Hi I'll h
 "PHARMA_viagra_PHARMA_cialis" - Wanted: web store with remedies. N



Framework

- 1 Get data
- 2 Visualize/perform sanity checks
- 3 Clean/filter observations
- 4 Choose features to represent data
- 5 Specify model
- 6 Specify loss function
- 7 Develop algorithm to minimize loss
- 8 Choose performance measure
- 9 "Train" to minimize loss
- 10 "Test" to evaluate generalization

Fwd: Yahoo! supercomputing cluster RFP - I have no idea, I have no idea. O
 non urgent - whoopee! yes that's what I meant, thanks for decoding my quest!
 SourceForge.net: variational bayes for network modularity - can i get admin |
 Byline - iPhone Apps, iPhone 3G apps and iPod touch Applications Gallery |
 Laurence J. Peter: Facts are stubborn things, but statistics are more pliable.
 Re: JAFOS 2008, Applied Math Session - yes, the listening post dude. On N
 Access to over 5,000 Health Plan Choices! - Affordable health insurance. In
 More effective - if you are having trouble viewing this email click here. Thus
 Special Offer! Cialis, Viagra, VcodinESI - Order all your Favorite Rx--Medica
 Financial Aid Available: Find Funding for Your Education - Get the financial i
 Find The Perfect School and Financial Aid for your College Degree - Hi I'll h
 "PHARMA_viagra_PHARMA_cialis" - Wanted: web store with remedies. N



Topics

- Supervised

- Linear regression
- Classification / regression trees
- Logistic regression
- Naive Bayes
- k-nearest neighbors
- Support vector machines
- Boosting

- Unsupervised

- K-means
- Mixture models
- Principal components analysis
- Factor analysis
- Topic models
- Collaborative filtering

Topics

- Supervised
 - Linear regression
 - Classification / regression trees
 - Logistic regression
 - Naive Bayes
 - k-nearest neighbors
 - Support vector machines
 - Boosting
 - Data representation: feature space, selection, normalization
 - Model assessment: complexity control, cross-validation, ROC curve, Bayesian Occam's razor, information-theoretic measures
- Unsupervised
 - K-means
 - Mixture models
 - Principal components analysis
 - Factor analysis
 - Topic models
 - Collaborative filtering

Topics

- Supervised
 - Linear regression
 - Classification / regression trees
 - Logistic regression
 - Naive Bayes
 - k-nearest neighbors
 - Support vector machines
 - Boosting
 - Data representation: feature space, selection, normalization
 - Model assessment: complexity control, cross-validation, ROC curve, Bayesian Occam's razor, information-theoretic measures
 - Probabilistic inference: graphical models, variational methods, sampling
 - Large-scale learning (?)
- Unsupervised
 - K-means
 - Mixture models
 - Principal components analysis
 - Factor analysis
 - Topic models
 - Collaborative filtering

Topics

- Simple approaches often do surprisingly well for large problems

Got data?

- Web service APIs expose vast amounts of data



[Subscribe](#)
[Register / Login](#)
[Home](#)
[News](#)
[APIs](#)
[Mashups](#)
[Members](#)
[How-To](#)

[Dashboard](#)
[Directory](#)
[Newest](#)
[Most Popular](#)
[By Category](#)
[API Scorecard](#)
[Add API](#)

Web Services Directory

[Subscribe to get the latest APIs](#)

Filter APIs
 Keywords:
 Category:
 Company:
 Protocols / Styles:
 Data Format:
 Managed By:
 Date:

View by Category

Sort by:

Viewing 1 to 1446 of 1446 APIs

API	Description	Category	Mashups
Google Maps	Mapping services	Mapping	1799
Flickr	Photo sharing service	Photos	476
YouTube	Video sharing and search	Video	413
Amazon eCommerce	Online retailer	Shopping	315
Twitter	Microblogging service	Social	260
eBay	Online auction marketplace	Shopping	178
Microsoft Virtual Earth	Mapping services	Mapping	173
del.icio.us	Social bookmarking	Bookmarks	139
Google Search	Search services	Search	135
Yahoo Maps	Mapping services	Mapping	131
Yahoo Search	Search services	Search	126
411Sync	SMS, WAP, and email messaging	Messaging	120
Last.fm	Online radio service	Music	120
Facebook	Social networking service	Social	107

Got data?

- Many free, public data sets available online

The screenshot shows the Infochimps website. At the top left is the logo "Infochimps" with a small figure icon and the tagline "Find any dataset in the world". To the right is a navigation menu with links for "Sign up", "Home", "About", "Help", "Blog", and "Gallery". A prominent orange banner contains the text: "Infochimps.org is still in beta testing. Anyone can browse and download data, but to upload, edit or add datasets you need an invite code. Request your beta invite now, and follow @infochimps on twitter!". Below this are three main sections: "Search for Data" (with a search input field containing "search for data" and a search button), "Browse Data" (with buttons for "Datasets", "Categories", "Tags", and "Sources"), and "Share Data" (with a green "Sign up" button). A vertical "feedback" button is on the right. At the bottom, there are two sections: "Some Interesting Datasets" with a list of items like "Stock Symbols & Metadata for all three US Stock Exchanges" and "Top Tags" with buttons for "government", "census", "population", "america", "demographics", "state", "selected", "olympics", and "type".

Coding

- Scripting: Python, Ruby, Perl, bash, ...
- Computing: R, SciPy/NumPy, MATLAB, ...
- Wrangling: sed, awk, grep, tr, wc, cut, sort, uniq,

Coding

- Scripting: Python, Ruby, Perl, bash, ...
- Computing: R, SciPy/NumPy, MATLAB, ...
- Wrangling: sed, awk, grep, tr, wc, cut, sort, uniq,

```
$ tr , '\t' < data.csv > data.tsv
```

Coding

- Scripting: Python, Ruby, Perl, bash, ...
- Computing: R, SciPy/NumPy, MATLAB, ...
- Wrangling: sed, awk, grep, tr, wc, cut, sort, uniq,

```
$ tr , '\t' < data.csv > data.tsv
```

```
$ bzcat data.tsv.bz2 | awk -F'\t' 'NF != 16 {print}'
```

Coding

- Scripting: Python, Ruby, Perl, bash, ...
- Computing: R, SciPy/NumPy, MATLAB, ...
- Wrangling: sed, awk, grep, tr, wc, cut, sort, uniq,

```
$ tr , '\t' < data.csv > data.tsv
```

```
$ bzcat data.tsv.bz2 | awk -F'\t' 'NF != 16 {print}'
```

```
$ sed -e 's/<[^>]*>//g' < page.html > page.txt
```