Practical Data Science @ Etsy

Dr. Jason Davis
About me

Ph.D. Machine learning & data mining

Entrepreneur. Founder of ad startup Adtuitive

Engineering Director @ Etsy. Search & Data
Topics

Etsy's data science infrastructure

The Precision / Recall tradeoff

Building products on the web using data science

Data science statistics

Analyzing products with data
Etsy's data science infrastructure
Mapreduce and Elasticity

Scalability

Operability
Elastic MapReduce (EMR) @ Etsy

- Nightly jobs
  - analytics
  - datasets

- Ad-hoc jobs

- ElasticMR
  - Production snapshots
  - Web log & beacons
  - ETL output
  - Dataset output
  - Ad-hoc output
  - Code

- S3

- Prod Dbs

- Akamai
EMR Scalability
0 to 1,000 machines to 0. Instantly.

Simplified job scheduling

Easy to "backfill" and run expensive jobs.
EMR Operability

Barnum does not persist clusters

1. Start job: create new cluster
2. Sync data from S3
3. Run mapreduce analysis
4. Shutdown cluster and sync results to S3

No operability support required
Precision vs. Recall
Precision vs. Recall

Recall = |Target AND Algorithm| / | Target |
Precision = |Target AND Algorithm| / | Algorithm |
Academic Diversion: Distance Metric Learning via LogDet Divergence

Goal: learn distance measure to respect constraints
Goal: learn distance metric subject to constraints

$$\min_A D_{ld}(A|A_0)$$

subject to

$$d_A(x_i, x_j) \leq u \quad (i, j) \in S,$$
$$d_A(x_i, x_j) \geq \ell \quad (i, j) \in D.$$ 

Mahalanobis distance:  $$d_A(x, y) = (x - y)^T A(x - y)$$

Log-det divergence:  $$\mathcal{W}(A_0^{-1}|A^{-1}) = C \exp \left( -\frac{n}{2} D_{ld}(A|A_0) + \frac{p + 1}{2} \log(n|A_0|) \right)$$
Recall vs. Precision
Adtuitive: High precision retail matching

Precision correlates with database size
Scale the db & simple measures work well
Etsy.com search: relevancy vs recency
Search on Etsy: relevancy vs recency

Buyer: "I want to see bananas"
Seller "I want to see my bananas"

Buyers: prefer high precision
Sellers: prefer high recall
Building products with data
SELECT * FROM listings WHERE title LIKE 'silver\%'
ORDER BY creation_tsz;
Automate your Etsy renewing!

The Statsy Clockbot allows you to have your Etsy listing renewals automatically performed at any time of your choosing!

Click Log in to log into Clockbot!

NEW! Try the brand-new Statsy 2!
Recall domination

I was looking for a pink bonnet with blue polka dots. I searched for "polka dot bonnet" but all I got were vintage polka records. Then I wanted a classic rock album and searched for "rock and roll" and got nothing but circular stones.

The switch to search relevancy
The switch to relevancy

Better buyer experience
- Higher precision via better search algorithms

Better advertising for sellers
- Opportunity to pay for recall via Search ads

Better search insight for sellers
- Visibility into search placement via Shop Stats
Step 1: Understand your data
Dig through raw data

Garbage in, garbage out
Visit logs, Raw query logs

80% of data science is data cleaning!
Descriptive statistics

Goal: reject hypotheses as quickly as possible

search depth
query conversion rates
Step 2: Invest in tools
Good tools provide visibility into the intersection of algorithms & data
Which set of search results for **typewriter in vintage items** is better?
search listings per page between 05/10/2012 to 05/25/2012

Description: The number of results to show per page

0.90 40 -
0.05 60 -
0.05 80 -
show filtered

0.09 category
0.88 search
0.03 search_similar_items
show filtered

A/B Analyzer
Step 3: Understand Tradeoffs

(most tradeoffs don't look like this)
Queries a, b, & c are better. Queries x & y are not.
Thinking short-term vs. long-term

How will the change affect behavior?

Affect expectations?

User incentives?

Improved data?
Step 4: Appreciate Product Complexity
Algorithmic Complexity
Improvement / Complexity Tradeoff

*Summer 2011 Search Relevance*
- 4 months
- 30 experiments
- 11 wins
UI Complexity
okay, so I'm just noticing something new in search results and I'm freaking out.

Since relevancy happened listings that had the search phrase in the title and as a single tag were more relevant than listings that had the search phrase words in separate tags. So if someone searched for a crochet hat, all the items on the first many pages had "crochet hat" in the title and "crochet hat" as a single tag.

Well, I have stopped being relevant lately, in all my searches: crochet hat, newsboy hat, earflap hat, etc. So I started doing a little detective work to figure out what is going on. Well, many relevant items no longer have those words as a single tag, but rather those words are often NOT in a single tag.

I kind of want to scream, because my season is approaching, and I'm nowhere to be found in any of my searches. So the question is: is anybody else noticing this change.

Posted at 1:30 am Sep 11, 2012 EDT
Data science statistics (at Etsy)
Popular web metrics

- Conversion rate: \( \frac{\text{# of purchases}}{\text{# of visits}} \)
- Registration rate: \( \frac{\text{# of registrations}}{\text{# of visits}} \)
- Funnel completion rates

Scenario:

- OMG, I just deployed unicorns to the site...
- The registration rate increased from 2.53% to 2.58%
- Where's the champagne?
The binomial distribution

Completion rates are binomial
- Visitor completes task (success)
- Visitor does not complete task (failure)

One parameter: $p = \text{probability of success}$

Other fun facts
- The binomial distribution is not the normal distribution
- $0 \leq p \leq 1$
Confidence Intervals

Hypothesis: my coin is biased and $p = 0.6$

Supporting data: 10 flips, 6 heads, $p_{\text{observed}} = 0.6$

What happens if I rerun the experiment?

What happens if I rerun the experiment 100 times?

- Answer: 95 out of 100 times, $0.262 < p_{\text{observed}} < 0.878$
The binomial confidence interval

Things to know
- Small $p$ values $\rightarrow$ exponentially more data
- $p$ values near 0.5 $\rightarrow$ less data
- Not symmetric

Example
- success = 6, trials = 100
- interval: $[0.022, 0.126]$
Example: Unicorn Search

Experiment setup: 50/50 unicorns / non-unicorns

Hypothesis: unicorns increase conversion rate by 2%

Assumptions
- 500k visits with search per day
- Baseline conversion is 2.0%
- Hypothesis conversion is 2.04%
Example: Unicorn Search

After 7 days
  ● 3.5M search visits, 1.75M per bucket
  ● Non-unicorns: 35,000 conversions
    ○ [0.0198, 0.0202]
  ● Unicorns: 35,700 conversions
    ○ [0.0202, 0.0206]

Conclusion:
We need an entire week to measure this improvement (!)
Analyzing products with data
Web analytics 101

Page view

Visit

Visitor

Bounce

Referer
Funnels

- Navigate
- Explore
- Buy
Example: search funnel

Search => View listing => Add to cart
The product lifecycle
Controlled Experiments
Controlled Experiments

registrations

“vinylgate”

Etsy on Oprah

Facebook registration released

Internet goes down in Portland

Zuckerberg caught stashing nuclear waste in crotch of tree
Experimental Steps

1. Think of something awesome; implement it
2. Release as controlled experiment to small percentage of users
3. Wait a week
4. Analyze on-site behavior comparing control vs test group
5. Apply statistics and determine if the change is a win or a loss
Experimentation @ Etsy

Continuous deployment

Simple experimental configuration

Automated analysis
Continuous deployment

All code is written from "trunk"

Changes are pushed out incrementally

40 times per day
Simple experimental configuration

```php
$server_config['search_infinite'] = array(
    'description' => "Search Infinite Scroll",
    'enabled' => 'rampup',
    'rampup' => array(
        'percent' => 50.0,
        'url_override' => "infinite",
    ),
);
```

Simple configuration => more experimentation
The A/B Analyzer

Indication of statistical significance

Percentage compared to control

Average page requests per visit

Requests/Visit

2.0537  2.0248

CONTROL  -1.41%
Experimental success

Category I: Fail! But you learned something
Category II: Micro-scopic success. Product-level metrics increased, larger Etsy metrics did not
Category III: Macro-scopic success