Deconstructing Data Science

David Bamman, UC Berkeley

Info 290
Lecture 1: Introduction

Jan 17, 2017
the “data scientist” trope
johannes kepler, data scientist
Florence Nightingale, data scientist
franz bopp, data scientist
Software/Libraries
Data Science

software

algorithms
classification, regression, clustering, network analysis, prediction, hypothesis testing,
data selection, representation, experimental design, validation

critical thinking
Computational Social Science

- Inferring ideal points of politicians based on voting behavior, speeches
- Detecting the triggers of censorship in blogs/social media
- Inferring power differentials in language use

Link structure in political blogs
Adamic and Glance 2005
Digital Humanities

- Authorship attribution (literary texts, paintings, music)
- Genre classification (literary genre, music genre)
- Inferring plot, character types

Predicting reviewed texts
Underwood and Sellers (2015)
Computational Journalism

• Exploratory data analysis for lead generation
• Information extraction from unstructured text
• Data-driven stories

Change in insured Americans under the ACA, NY Times (Oct 29, 2014)
What to expect

• Each class: learn about a technical method (e.g., random forests), and discuss application area that makes use of it.

• As the course goes on, we’ll compare methods with those we’ve already learned to critically assess the assumptions that they make and understand what methods are appropriate for different contexts.

• We will learn by example: Lots of reading.
Themes
1. Validity

How do we assess that a model is valid?
2. Transparency

How do we understand what a model is learning?
3. Fairness

To what degree does a problem translate biases in the input data into biases in its output?

Predictive policing: heat map indicating increased risk of certain crimes
http://magazine.ucla.edu/depts/quicktakes/a-weapon-to-fight-crime-try-math/
Topics

• Overview of methods (classification, regression, clustering)

• Classification: decision trees, random forests, probabilistic models (naive bayes, logistic regression), neural networks

• Clustering: latent variable models (topic models), PCA, factor analysis, K-means, hierarchical clustering

• Linear regression

• Networks (structural properties, diffusion)

• Causal inference
Applications

- Authorship attribution
- Latent attribute prediction
- Predicting movie revenue
- Recommender systems
- Music genre classification

- Word embeddings
- Visual style classification
- Text reuse
- Genre clustering
- Predicting high school dropout rates
... in medias res

- Task: predict political preference of Twitter users.

- Assume access to training data \( <x, y> \) where:
  
  \[
  x = \text{set of Twitter users} \\
  y = \{\text{Democrat, Republican}\}
  \]
Representation

- How can you best represent a data point to enable learning?
David Bamman
@dbamman
Assistant Professor, School of Information, UC Berkeley. Natural language processing, machine learning, computational social science, digital humanities.

Berkeley, CA
people.ischool.berkeley.edu/~dbamman/
Joined October 2009

Tweets
508

Following
400

Followers
799

Likes
133

Lists
2

Tweets
Retweeted

Ted Underwood @Ted_Underwood · 6h
How have the differences between descriptions of men and women in fiction changed over the last 200 yrs? (ICYMI)
tedunderwood.com/2016/01/09/the...

Retweeted

David Bamman @dbamman · Jan 6
"Figure Eights" (Max Roach/Buddy Rich, 1959) is just dazzling. Probably no video of them anywhere? open.spotify.com/track/23EssvWY...

Retweeted

Anders Søgaard @soegaarducph · Jan 6
@stanfordnlp @brendan642 @jacobbeisenstein Here goes: twitter-research.ccs.neu.edu/language/
<table>
<thead>
<tr>
<th>Feature</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>follow clinton</td>
<td>0</td>
</tr>
<tr>
<td>follow trump</td>
<td>0</td>
</tr>
<tr>
<td>“benghazi”</td>
<td>0</td>
</tr>
<tr>
<td>negative sentiment + “benghazi”</td>
<td>0</td>
</tr>
<tr>
<td>“illegal immigrants”</td>
<td>0</td>
</tr>
<tr>
<td>“republican” in profile</td>
<td>0</td>
</tr>
<tr>
<td>“democrat” in profile</td>
<td>0</td>
</tr>
<tr>
<td>self-reported location = Berkeley</td>
<td>1</td>
</tr>
</tbody>
</table>

$x = \text{feature vector}$
\[
\sum_{i=1}^{F} x_i \beta_i = x_1 \beta_1 + x_2 \beta_2 + \ldots + x_F \beta_F
\]

\[
= x^T \beta \quad \text{(dot product, inner product)}
\]

\[
\hat{y}_i = \begin{cases} 
1 & \text{if } \sum_{i}^{F} x_i \beta_i \geq 0 \\
-1 & \text{otherwise}
\end{cases}
\]
\[ \mathbf{x} = \text{feature vector} \]

<table>
<thead>
<tr>
<th>Feature</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>follow clinton</td>
<td>0</td>
</tr>
<tr>
<td>follow trump</td>
<td>0</td>
</tr>
<tr>
<td>“benghazi”</td>
<td>0</td>
</tr>
<tr>
<td>negative sentiment + “benghazi”</td>
<td>0</td>
</tr>
<tr>
<td>“illegal immigrants”</td>
<td>0</td>
</tr>
<tr>
<td>“republican” in profile</td>
<td>0</td>
</tr>
<tr>
<td>“democrat” in profile</td>
<td>0</td>
</tr>
<tr>
<td>self-reported location = Berkeley</td>
<td>1</td>
</tr>
</tbody>
</table>

\[ \mathbf{\beta} = \text{coefficients} \]

<table>
<thead>
<tr>
<th>Feature</th>
<th>( \beta )</th>
</tr>
</thead>
<tbody>
<tr>
<td>follow clinton</td>
<td>-3.1</td>
</tr>
<tr>
<td>follow trump</td>
<td>6.8</td>
</tr>
<tr>
<td>“benghazi”</td>
<td>1.4</td>
</tr>
<tr>
<td>negative sentiment + “benghazi”</td>
<td>3.2</td>
</tr>
<tr>
<td>“illegal immigrants”</td>
<td>8.7</td>
</tr>
<tr>
<td>“republican” in profile</td>
<td>7.9</td>
</tr>
<tr>
<td>“democrat” in profile</td>
<td>-3.0</td>
</tr>
<tr>
<td>self-reported location = Berkeley</td>
<td>-1.7</td>
</tr>
<tr>
<td></td>
<td>“benghazi”</td>
</tr>
<tr>
<td>-------</td>
<td>------------</td>
</tr>
<tr>
<td>(\beta)</td>
<td>1.4</td>
</tr>
<tr>
<td>(x^1)</td>
<td>1</td>
</tr>
<tr>
<td>(x^2)</td>
<td>0</td>
</tr>
<tr>
<td>(x^3)</td>
<td>1</td>
</tr>
</tbody>
</table>

\[(1 \times 1.4) + (1 \times 6.8) + (0 \times -3.1) = 8.2\]
Learning

How do get good values for $\beta$?

<table>
<thead>
<tr>
<th>Feature</th>
<th>$\beta$</th>
</tr>
</thead>
<tbody>
<tr>
<td>follow clinton</td>
<td>-3.1</td>
</tr>
<tr>
<td>follow trump</td>
<td>6.8</td>
</tr>
<tr>
<td>“benghazi”</td>
<td>1.4</td>
</tr>
<tr>
<td>negative sentiment + “benghazi”</td>
<td>3.2</td>
</tr>
<tr>
<td>“illegal immigrants”</td>
<td>8.7</td>
</tr>
<tr>
<td>“republican” in profile</td>
<td>7.9</td>
</tr>
<tr>
<td>“democrat” in profile</td>
<td>-3.0</td>
</tr>
<tr>
<td>self-reported location = Berkeley</td>
<td>-1.7</td>
</tr>
</tbody>
</table>
Online learning

- Go through the training data \( <x, y> \) one data point at a time.

- Make a prediction \( \hat{y} \) with current estimate of \( \beta \); if it’s right \( (y = \hat{y}) \), do nothing.

- If the prediction is wrong \( (y \neq \hat{y}) \), change \( \beta \) to make it slightly less wrong.
\[ \hat{y}_i = \begin{cases} 
1 & \text{if } \sum_{i}^{F} x_i \beta_i \geq 0 \\
-1 & \text{otherwise} 
\end{cases} \]

<table>
<thead>
<tr>
<th>“benghazi”</th>
<th>follows trump</th>
<th>follows clinton</th>
<th>y</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
<td>-1</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
<td>-1</td>
</tr>
</tbody>
</table>

training data
\[ \hat{y}_i = \begin{cases} 
1 & \text{if } \sum F_i x_i \beta_i \geq 0 \\
-1 & \text{otherwise} 
\end{cases} \]

<table>
<thead>
<tr>
<th>“benghazi” follows</th>
<th>follows</th>
<th>y</th>
<th>( \hat{y} )</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>trump</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
<td>-1</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>-1</td>
</tr>
</tbody>
</table>
true \quad y = -1 \\
predicted \quad \hat{y} = 1

\sum_{i}^{F} x_i \beta_i

\frac{\partial}{\partial \beta_i} \sum_{i}^{F} x_i \beta_i = x_i

\beta_{t+1} = \beta_t - x

We want this value (function of $\beta$) to be small

The derivative tells us the direction to go to make it bigger or smaller

Update rule
true \quad y = 1  \quad \sum_{i}^{F} x_i \beta_i

predicted \quad \hat{y} = -1

\frac{\partial}{\partial \beta_i} \sum_{i}^{F} x_i \beta_i = x_i

\beta_{t+1} = \beta_t + x

We want this value (function of \( \beta \)) to be big

The derivative tells us the direction to go to make it bigger or smaller

Update rule
if $\hat{y} = 1$ and $y = -1$

$\beta_{t+1} = \beta_t - x$

<table>
<thead>
<tr>
<th>$\beta_t$</th>
<th>$x$</th>
<th>$\beta_{t+1}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.6</td>
<td>0</td>
<td>3.6</td>
</tr>
<tr>
<td>3.4</td>
<td>1</td>
<td>2.4</td>
</tr>
<tr>
<td>1.2</td>
<td>1</td>
<td>0.2</td>
</tr>
<tr>
<td>0.7</td>
<td>0</td>
<td>0.7</td>
</tr>
</tbody>
</table>

$\sum x_i \beta_i$ 4.6  2.6

$\hat{y}$ 1  1
if $\hat{y} = -1$ and $y = 1$

$$\beta_{t+1} = \beta_t + x$$

<table>
<thead>
<tr>
<th>$\beta_t$</th>
<th>$x$</th>
<th>$\beta_{t+1}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.6</td>
<td>0</td>
<td>3.6</td>
</tr>
<tr>
<td>-3.4</td>
<td>1</td>
<td>-2.4</td>
</tr>
<tr>
<td>1.2</td>
<td>1</td>
<td>2.2</td>
</tr>
<tr>
<td>0.7</td>
<td>0</td>
<td>0.7</td>
</tr>
<tr>
<td>$\sum x_i \beta_i$</td>
<td>-2.2</td>
<td>-0.2</td>
</tr>
<tr>
<td>$\hat{y}$</td>
<td>-1</td>
<td>-1</td>
</tr>
</tbody>
</table>
if $\hat{y} = 1$ and $y = -1$

$$\beta_{t+1} = \beta_t - x$$

if $\hat{y} = -1$ and $y = 1$

$$\beta_{t+1} = \beta_t + x$$

$$\beta_{t+1} = \beta_t + yx$$
Why $\beta_{t+1} = \beta_t + yx$?

[Approximation of stochastic gradient in binary logistic regression (lecture 9)]
Perceptron

**Data:** training data $x \in \mathbb{R}^F$, $y \in \{-1, +1\}$, $i = 1 \ldots N$; initialize $\beta_0 = 0^F$; $k=0$;

while not converged do
  $k = k + 1$;
  $i = k \text{ mod } N$;
  if $\hat{y}_i \neq y_i$ then
    $\beta_{t+1} = \beta_t + y_ix_i$
  else
    do nothing;
  end
end

Rosenblatt 1957
Code
decision boundary in 2 dimensions

$$x_1 \beta_1 + x_2 \beta_2 = 0$$

<table>
<thead>
<tr>
<th># of &quot;hillary&quot;</th>
<th># of &quot;trump&quot;</th>
<th>y</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>8</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>11</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>12</td>
<td>1</td>
</tr>
<tr>
<td>10</td>
<td>2</td>
<td>-1</td>
</tr>
<tr>
<td>11</td>
<td>1</td>
<td>-1</td>
</tr>
<tr>
<td>12</td>
<td>3</td>
<td>-1</td>
</tr>
</tbody>
</table>
Trends

- Counts later points more than earlier points (voted perceptron, averaged perceptron)
- Only linear decision boundaries
- Prone to overfitting
- Extraordinarily simple and accurate classifier
Problem assumptions

• Is this the right task (classification vs. clustering vs. regression, time series forecasting etc.)

• Is the data appropriate for the problem?
Administrivia

• David Bamman
dbamman@berkeley.edu

Office hours: Thursdays 10am-noon, 314 SH — or by appointment

• Rob Kuvinka, TA

Office hours: Tuesday, 5-7pm, 110 South Hall
Grading

• Class participation (10%)

• Homeworks (4 x 12.5%)

• Project (40%)

All deliverables (homeworks, project components) have deadlines; late work not accepted after 2 “free days” used up
Free days

- You have a total of 2 “free days” to use over the entire semester.
- Each free day gives you an extra 24 hours to turn in a homework assignment.
- A free day is used up once you cross the deadline for a homework being due (e.g., 12:01am for a 12:00am deadline).
- Use them wisely!
Homeworks, broadly

A

- Implement a quantitative method and evaluate it on a dataset

B

- Write an analysis/critique of an algorithm and published work that has used it
Homework Example

Binary perceptron classifies into two classes. For inferring political preference, this corresponds to a simple {Democrat, Republican} distinction. Assume rather that the training data you have is hierarchical. Design a perceptron-style algorithm that can exploit this hierarchical structure during learning.

A
Code and evaluate on test data

B
What are the comparative advantages and disadvantages of binary vs. multiclass vs. hierarchical categories? Under what circumstances should either be used? (2 pages, single-spaced)
Participation

• Most classes will include discussion of an application as documented in a research paper.

• While everyone is expected to read these papers, one student each class will act as a discussion leader, coming prepared with questions and discussion topics for the class as a whole to discuss.
Project

• Use methods learned in class to draw inferences about the world and critically assess the quality of the results.

• Collaborative (2-3 students). Choose wisely! Everyone in group will receive the same grade; you will be evaluated both on the empirical methodology and the domain questions you’re asking.
Project

• Milestones:

  • Proposal and literature review (5%). 2 pages, 5 sources.
  • Midterm report (10%). 4 pages, 10 sources.
  • Final report (20%). 10 pages.
  • Presentation (5%). 15-20 min. conference-style talk in front of peers.

• Evaluated according to standards for conference publication—clarity, originality, soundness, substance, evaluation, meaningful comparison, impact.