Machine Learning, Predictive Text, and Topic Models

Hanna Wallach
University of Massachusetts Amherst
wallach@cs.umass.edu
Outline

• What is machine learning?
• Examples of machine learning in practice:

$30: Dinner, Cambridge MA
$50: Bus ticket, Cambridge MA
$5000: Hotel suite, Hong Kong
$20: Beer, Amherst MA
$10: Lunch, Amherst MA

Hanna Wallach  Machine Learning, Predictive Text, and Topic Models
Machine Learning (ML)

- There are increasingly large amounts of digital data available:

- Machine learning uses computers to find the most salient features in data to further knowledge and make life easier

... with as little human input as possible
Uncertainty

- There is uncertainty in almost all real world situations:

- ML explicitly represents uncertainty using probability:
  - $\Pr(\text{lemon}) = \text{how certain I am that this is a lemon}$

- Probability provides a framework for reasoning under uncertainty
USPS Digit Recognition

- **Problem:**
  - USPS needs to sort letters by zip code

- **Solution:**
  - Teach a computer to recognize hand-written digits
  - Only ask human when computer is uncertain:

![Images of hand-written digits with 1 or 7 marked]
Credit Card Fraud

- Problem:
  - Want to detect credit card fraud

- Solution:
  - Train a computer to recognise normal and abnormal usages
  - Alert card-holder if abnormal pattern is detected

$30: Dinner, Cambridge MA
$50: Bus ticket, Cambridge MA
$10: Lunch, Amherst MA
$20: Beer, Amherst MA
$10: Lunch, Amherst MA
$30: Dinner, Cambridge MA
$50: Bus ticket, Cambridge MA
$5000: Hotel suite, Hong Kong
$20: Beer, Amherst MA
$10: Lunch, Amherst MA
Sorting News Stories by Genre

Dmitry Medvedev on the Record
BusinessWeek - 7 hours ago
On Dec. 10, Russian President Vladimir Putin revealed the name of the candidate he will back in Russia’s presidential election in March.

Video: Putin will be my Prime Minister: Medvedev RussiaToday
A cautious welcome in Europe for Medvedev
International Herald Tribune
all 1,286 news articles »

Deadlock Stymies Global Climate Talks
New York Times - 50 minutes ago
By T.1OMAS FULLER This article was reported by Thomas Fuller, Peter Gelling and Andrew C. Revkin, and was written by Mr. Fuller. NUSA DUA, Indonesia - As a United Nations conference on global warming here entered its final stretch, the United States and ...

Video: Bali climate meeting: new environmental pact in focus RussiaToday
Can the Planet be Saved in Bali? TIME
Reuters - Los Angeles Times - International Herald Tribune - CNN
all 828 news articles »

Show more stories  Show fewer stories

Fed Lowers Rate by a Quarter Point to 4.25 Percent (Update6)
Bloomberg - 2 hours ago
By Craig Torres Dec. 11 (Bloomberg) -- The Federal Reserve lowered its benchmark interest rate by a quarter point to 4.25

Ask.com Adds Priv Search Data
Wall Street Journal - J
By SCOTT MORRISON SA
Tuesday a "privacy switch
Predictive Text Entry

- e.g., T9 or iTAP
- Used on cell phones
- Enables use of reduced keyboard
- Enter as much text as possible with as few gestures as possible

Text  ➔  Gestures
(as few as possible)
Predictive Text Entry

- This is like the reverse of text compression
- Text compression: want to go from as much text as possible to as small a representation as possible

Text → Bit string (preferably short)
Writing and Text Compression

- **Optimal** text compression

![Diagram](text.png)
Writing and Text Compression

- Optimal text compression and writing with **predictive text entry**

\[\text{Text} \xrightarrow{\text{probabilistic model}} \text{Bit string (preferably short)}\]

\[\text{Text} \xleftarrow{\text{probabilistic model}} \text{Gestures (as few as possible)}\]
Dasher [http://www.dasher.org.uk]

- Driven by 2D continuous gestures
- Uses a model of language
- Available for
  - Windows
  - Linux
  - Mac OS X
  - Pocket PC
  - etc.
Dasher: Screen Layout

- Box sizes are proportional to probabilities
- Probabilities come from a letter-based language model
- \( P(X) = b \)
  \( P(X, Y) = a \)
Dasher: Dynamics

Point where you want to go

- Like driving a car
- Motion sickness?
- Not if you're driving!
Dasher: Benefits

- Keyboards: one gesture per character
- Dasher: some gestures select many characters
- Works with any language
- Inaccurate gestures can be compensated for by later gestures
Topic Models

- Humans can read a document and identify the small number of topics that best characterize that document.
Topic Models

- Topics are mixtures of words and documents are mixtures of topics.
Topic Models

- Infer topic information from word-document co-occurrences
## Example Topics

[Tenenbaum et al.]

<table>
<thead>
<tr>
<th>STORY</th>
<th>FIELD</th>
<th>SCIENCE</th>
<th>BALL</th>
<th>JOB</th>
</tr>
</thead>
<tbody>
<tr>
<td>STORIES</td>
<td>MAGNETIC</td>
<td>STUDY</td>
<td>GAME</td>
<td>WORK</td>
</tr>
<tr>
<td>TELL</td>
<td>MAGNET</td>
<td>SCIENTISTS</td>
<td>TEAM</td>
<td>JOBS</td>
</tr>
<tr>
<td>CHARACTER</td>
<td>WIRE</td>
<td>SCIENTIFIC</td>
<td>FOOTBALL</td>
<td>CAREER</td>
</tr>
<tr>
<td>CHARACTERS</td>
<td>NEEDLE</td>
<td>KNOWLEDGE</td>
<td>BASEBALL</td>
<td>EXPERIENCE</td>
</tr>
<tr>
<td>AUTHOR</td>
<td>CURRENT</td>
<td>WORK</td>
<td>PLAYERS</td>
<td>OPPORTUNITIES</td>
</tr>
<tr>
<td>READ</td>
<td>COIL</td>
<td>RESEARCH</td>
<td>PLAY</td>
<td>WORKING</td>
</tr>
<tr>
<td>TOLD</td>
<td>POLES</td>
<td>CHEMISTRY</td>
<td>FIELD</td>
<td>TRAINING</td>
</tr>
<tr>
<td>SETTING</td>
<td>IRON</td>
<td>TECHNOLOGY</td>
<td>PLAYER</td>
<td>SKILLS</td>
</tr>
<tr>
<td>TALES</td>
<td>COMPASS</td>
<td>MANY</td>
<td>BASKETBALL</td>
<td>CAREERS</td>
</tr>
<tr>
<td>PLOT</td>
<td>LINES</td>
<td>MATHEMATICS</td>
<td>COACH</td>
<td>POSITIONS</td>
</tr>
<tr>
<td>TELLING</td>
<td>CORE</td>
<td>BIOLOGY</td>
<td>PLAYED</td>
<td>FIND</td>
</tr>
<tr>
<td>SHORT</td>
<td>ELECTRIC</td>
<td>FIELD</td>
<td>PLAYING</td>
<td>POSITION</td>
</tr>
<tr>
<td>FICTION</td>
<td>DIRECTION</td>
<td>PHYSICS</td>
<td>HIT</td>
<td>FIELD</td>
</tr>
<tr>
<td>ACTION</td>
<td>FORCE</td>
<td>LABORATORY</td>
<td>TENNIS</td>
<td></td>
</tr>
</tbody>
</table>
Transfer between Topics [Mimno]
Entities and Topics [Newman et al.]

<table>
<thead>
<tr>
<th>Sept. 11</th>
<th>Fear</th>
<th>US Pride</th>
<th>Defense</th>
<th>Agencies</th>
</tr>
</thead>
<tbody>
<tr>
<td>attack</td>
<td>0.017</td>
<td>fear 0.023</td>
<td>american 0.062</td>
<td>defense 0.039</td>
</tr>
<tr>
<td>victim</td>
<td>0.016</td>
<td>public 0.019</td>
<td>flag 0.046</td>
<td>missile 0.039</td>
</tr>
<tr>
<td>tragedy</td>
<td>0.015</td>
<td>threat 0.011</td>
<td>country 0.035</td>
<td>system 0.032</td>
</tr>
<tr>
<td>missing</td>
<td>0.013</td>
<td>concern 0.010</td>
<td>war 0.028</td>
<td>administration 0.019</td>
</tr>
<tr>
<td>lost</td>
<td>0.012</td>
<td>anger 0.008</td>
<td>nation 0.022</td>
<td>arms 0.019</td>
</tr>
<tr>
<td>families</td>
<td>0.012</td>
<td>crisis 0.008</td>
<td>history 0.012</td>
<td>weapon 0.019</td>
</tr>
<tr>
<td>lives</td>
<td>0.010</td>
<td>support 0.007</td>
<td>feel 0.010</td>
<td>nuclear 0.015</td>
</tr>
<tr>
<td>memorial</td>
<td>0.010</td>
<td>sense 0.007</td>
<td>symbol 0.009</td>
<td>test 0.014</td>
</tr>
<tr>
<td>happened</td>
<td>0.009</td>
<td>seen 0.007</td>
<td></td>
<td>missiles 0.013</td>
</tr>
<tr>
<td>dead</td>
<td>0.009</td>
<td>changed 0.006</td>
<td></td>
<td>treaty 0.012</td>
</tr>
<tr>
<td>E130</td>
<td>0.980</td>
<td>E55 0.720</td>
<td>E55 1.000</td>
<td>E145 0.780</td>
</tr>
<tr>
<td>NY</td>
<td>0.188</td>
<td>BUSH 0.290</td>
<td>AMERICA 0.164</td>
<td>RUSSIA 0.113</td>
</tr>
<tr>
<td>WTC</td>
<td>0.091</td>
<td>CLINTON 0.133</td>
<td>US 0.102</td>
<td>PENTAGON 0.073</td>
</tr>
<tr>
<td>AMERICA</td>
<td>0.071</td>
<td>WHITE HSE 0.094</td>
<td>WASH. DC 0.064</td>
<td>CHINA 0.057</td>
</tr>
<tr>
<td>GOD</td>
<td>0.036</td>
<td>WASH DC 0.075</td>
<td>BUSH 0.037</td>
<td>CLINTON 0.055</td>
</tr>
<tr>
<td>WASH. DC</td>
<td>0.035</td>
<td>CONGRESS 0.062</td>
<td>WW2 0.024</td>
<td>BUSH 0.052</td>
</tr>
<tr>
<td>NYC</td>
<td>0.027</td>
<td>POWELL 0.032</td>
<td>CIVIL WAR 0.021</td>
<td>PUTIN 0.046</td>
</tr>
<tr>
<td>GIULIANI</td>
<td>0.023</td>
<td>UN 0.014</td>
<td>WEST 0.012</td>
<td>N. KOREA 0.033</td>
</tr>
<tr>
<td></td>
<td></td>
<td>PRESIDENT 0.014</td>
<td>RIGHT 0.012</td>
<td>IRAQ 0.029</td>
</tr>
</tbody>
</table>
Topics and Email

- Enron email corpus:
  - 250k email messages, 23k people

Sally -
Attached are the hypertiles from the final report out at yesterday’s ASE Studio Workshop. The CD is finished and on its way to Houston. The files are organized by team:
Hammer - Sales and Marketing, Vision Stmt, Mission Stmt, Target Market, How to Approach, Pricing, SLA
Pliers - Producst and Services - Consulting Based
Saw - Infrastructure Transition Plan
Wrench - Producst and Services - Basic Outsourcing
I hope these help with your meeting tomorrow. Let me know if there is anything else I can do to help.
Lisa P
Selecting Email Keywords [Dredze et al.]

Sally -
Attached are the hypertiles from the final report out at yesterday’s ASE Studio Workshop. The CD is finished and on its way to Houston. The files are organized by team:
- **Hammer** - Sales and Marketing, Vision Stmt, Mission Stmt, Target Market, How to Approach, Pricing, SLA
- **Pliers** - Product and Services - Consulting Based
- **Saw** - Infrastructure Transition Plan
- **Wrench** - Product and Services - Basic Outsourcing

I hope these help with your *meeting* tomorrow. Let me know if there is anything else I can do to help.

Lisa P

- **Without topics:** producst pliers stmt hammer wrench
- **With topics:** team meeting services lisa ase
### Senders, Recipients, Topics

[McCallum et al.]

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>operations 0.0321</td>
<td>market 0.0567</td>
<td>state 0.0404</td>
<td>blackberry 0.0726</td>
</tr>
<tr>
<td>team 0.0234</td>
<td>power 0.0563</td>
<td>california 0.0367</td>
<td>net 0.0557</td>
</tr>
<tr>
<td>office 0.0173</td>
<td>price 0.0280</td>
<td>power 0.0337</td>
<td>www 0.0409</td>
</tr>
<tr>
<td>list 0.0144</td>
<td>system 0.0206</td>
<td>energy 0.0239</td>
<td>website 0.0375</td>
</tr>
<tr>
<td>bob 0.0129</td>
<td>prices 0.0182</td>
<td>electricity 0.0203</td>
<td>report 0.0373</td>
</tr>
<tr>
<td>open 0.0126</td>
<td>high 0.0124</td>
<td>davis 0.0183</td>
<td>wireless 0.0364</td>
</tr>
<tr>
<td>meeting 0.0107</td>
<td>based 0.0120</td>
<td>utilities 0.0158</td>
<td>handheld 0.0362</td>
</tr>
<tr>
<td>gas 0.0107</td>
<td>buy 0.0117</td>
<td>commission 0.0136</td>
<td>stan 0.0282</td>
</tr>
<tr>
<td>business 0.0106</td>
<td>customers 0.0110</td>
<td>governor 0.0132</td>
<td>fyi 0.0271</td>
</tr>
<tr>
<td>houston 0.0099</td>
<td>costs 0.0106</td>
<td>prices 0.0089</td>
<td>named 0.0260</td>
</tr>
</tbody>
</table>

| S.Beck 0.2158         | J.Dasovich 0.1231        | J.Dasovich 0.3338               | R.Haylett 0.1432   |
| L.Kitchen 0.0826      | J.Steffes 0.1133         | R.Shapiro 0.2440                | T.Geaccone 0.0737  |
| S.Beck 0.0826         | J.Dasovich 0.1133        | J.Dasovich 0.2440                | R.Haylett 0.0737   |
| J.Lavorato 0.0530     | R.Shapiro 0.1133         | J.Steffes 0.1394                 | T.Geaccone 0.0737  |
| S.Beck 0.0530         | M.Taylor 0.0218          | J.Dasovich 0.1394                | R.Haylett 0.0420   |
| S.White 0.0530        | E.Sager 0.0218           | R.Sanders 0.1394                 | D.Fossum 0.0260    |

“Chief Operations Officer”

“Government Relations Executive”

Hanna Wallach

Machine Learning, Predictive Text, and Topic Models

24
Summary

- Machines can learn a lot from unstructured digital data
- We can use machine learning to build useful applications, some of which you are already using!
Questions?