Lecture 2: Probability and Language Models

Intro to NLP, CS585, Fall 2014
Brendan O’Connor (http://brenocon.com)
Admin

- Waitlist
- Moodle access: Email me if you don’t have it
- Did you get an announcement email?
- Piazza vs Moodle?
- Office hours today
Things today

• Homework: ambiguities
• Python demo
• Probability Review
• Language Models
Python demo

- [TODO link ipython-notebook demo]

- For next week, make sure you can run
  - Python 2.7 (Built-in on Mac & Linux)
  - IPython Notebook [http://ipython.org/notebook.html](http://ipython.org/notebook.html)
    - Please familiarize yourself with it.
    - Python 2.7, IPython 2.2.0
  - Nice to have: Matplotlib
  - Python interactive interpreter
  - Python scripts
Levels of linguistic structure

Discourse

Semantics

Syntax

Words

Morphology

Characters

CommunicationEvent(e)  SpeakerContext(s)
Agent(e, Alice)  TemporalBefore(e, s)
Recipient(e, Bob)

S

NP

VerbPast

Prep

NounPrp

Punct

Alice talked to Bob.

talk -ed

Alice talked to Bob.
Levels of linguistic structure

Words are fundamental units of meaning and easily identifiable*.

*in some languages

---

**Words**

Alice talked to Bob.

**Characters**

Alice talked to Bob.
Probability theory
Review: definitions/laws

Conditional Probability

\[ P(A = a) = \sum_{a} P(A = a) \]

Chain Rule

\[ P(AB) = P(A|B)P(B) \]

Law of Total Probability

\[ P(A, B = b) = \sum_{b} P(A, B = b) \]
\[ = \sum_{b} P(A|B = b)P(B = b) \]

Disjunction (Union)

\[ P(A \lor B) = \]

Negation (Complement)

\[ P(\neg A) = \]

Thursday, September 4, 14
Bayes Rule

Want $P(H|D)$ but only have $P(D|H)$
e.g. $H$ causes $D$, or $P(D|H)$ is easy to measure...

H: who wrote this
document?  
Model: authors’ word probs

Bayesian inference

$D$: words

$P(H|D) = \frac{P(D|H)P(H)}{P(D)}$

Likelihood  Prior

Posterior

Normalizer
Bayes Rule and its pesky denominator

Bayesian rule:

\[ P(h|d) = \frac{P(d|h)P(h)}{P(d)} = \frac{P(d|h)P(h)}{\sum_{h'} P(d|h')P(h')} \]

Unnormalized posterior:

\[ P(h|d) \propto P(d|h)P(h) \]

By itself does not sum to 1!

Z: whatever lets the posterior, when summed across \( h \), to sum to 1

\[ Z = \text{Zustandssumme, "sum over states"} \]

"Proportional to" (implicitly for varying H. This notation is very common, though slightly ambiguous.)
Bayes Rule: Discrete

\[
P(E|H = h) \quad \text{Prior}
\]
\[
P(E|H = h)P(H = h) \quad \text{Unnorm. Posterior}
\]
\[
\frac{1}{Z}P(E|H = h)P(H = h) \quad \text{Posterior}
\]
Bayes Rule: Discrete, uniform prior

Uniform distribution: “Uninformative prior”

\[ P(H = h) \]
Prior

\[ P(E|H = h) \]
Likelihood

\[ P(E|H = h)P(H = h) \]
Unnorm. Posterior

Uniform prior implies that posterior is just renormalized likelihood

\[ \frac{1}{Z} P(E|H = h)P(H = h) \]
Posterior

<table>
<thead>
<tr>
<th>Sum to 1?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
</tr>
<tr>
<td>No</td>
</tr>
<tr>
<td>No</td>
</tr>
<tr>
<td>Yes</td>
</tr>
</tbody>
</table>

Thursday, September 4, 14
Bayes Rule for doc classification

Assume a generative process, $P(w \mid y)$

Inference problem: given $w$, what is $y$?

If we knew $P(w \mid y)$
We could estimate $P(y \mid w) \propto P(y)P(w \mid y)$

Assume a generative process, $P(w \mid y)$

Inference problem: given $w$, what is $y$?

If we knew $P(w \mid y)$
We could estimate $P(y \mid w) \propto P(y)P(w \mid y)$

Look at random word.
It is *abracadabra*

Assume 50% prior prob
Prob author is Anna?

<table>
<thead>
<tr>
<th></th>
<th><em>abracadabra</em></th>
<th><em>gesundheit</em></th>
</tr>
</thead>
<tbody>
<tr>
<td>Anna</td>
<td>5 per 1000 words</td>
<td>6 per 1000 words</td>
</tr>
<tr>
<td>Barry</td>
<td>10 per 1000 words</td>
<td>1 per 1000 words</td>
</tr>
</tbody>
</table>
Bayes Rule for doc classification

Assume a generative process, \( P(w \mid y) \)

Inference problem: given \( w \), what is \( y \)?

If we knew \( P(w \mid y) \), we could estimate

\[
P(y \mid w) \propto P(y)P(w \mid y)
\]

Assume a generative process, \( P(w \mid y) \)

Inference problem: given \( w \), what is \( y \)?

If we knew \( P(w \mid y) \), we could estimate

\[
P(y \mid w) \propto P(y)P(w \mid y)
\]

Assume a generative process, \( P(w \mid y) \)

Inference problem: given \( w \), what is \( y \)?

If we knew \( P(w \mid y) \), we could estimate

\[
P(y \mid w) \propto P(y)P(w \mid y)
\]

Assume a generative process, \( P(w \mid y) \)

Inference problem: given \( w \), what is \( y \)?

If we knew \( P(w \mid y) \), we could estimate

\[
P(y \mid w) \propto P(y)P(w \mid y)
\]
Bayes Rule for doc classification

Assume a generative process, \( P(w | y) \)

Inference problem: given \( w \), what is \( y \)?

If we knew \( P(w | y) \)
We could estimate \( \frac{P(w | y)}{P(y)} \propto P(y)P(w | y) \)

Assume conditional independence:

\[
P(w_1, w_2 | y) = P(w_1 | w_2, y) P(w_2 | y)
\]

ASSUME conditional independence:

\[
P(w_1, w_2 | y) = P(w_1 | y) P(w_2 | y)
\]

Look at two random words.

\( w_1 = \text{abracadabra} \)
\( w_2 = \text{gesundheit} \)

Assume 50% prior prob of author being Anna.

<table>
<thead>
<tr>
<th>abracadabra</th>
<th>gesundheit</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anna</td>
<td>5 per 1000 words</td>
</tr>
<tr>
<td>Barry</td>
<td>10 per 1000 words</td>
</tr>
</tbody>
</table>

Chain rule:
Cond indep. assumption: “Naive Bayes”

Assume a generative process, \( P(w \mid y) \)

Inference problem: given \( w \), what is \( y \)?

\[
P(w_1 \ldots w_T \mid y) = \prod_{t=1}^{T} P(w_t \mid y)
\]

each \( w_t \in 1..V \quad V = \text{vocabulary size} \)

Generative story (“Multinom NB” [McCallum & Nigam 1998]):
- For each token \( t \) in the document,
  - Author chooses a word
    by rolling the same weighted V-sided die
This model is wrong!
How can it possibly be useful for doc classification?
Bayes Rule for text inference

Noisy channel model

Original text

Hypothesized transmission process

Inference problem

Observed data

Codebreaking

\[ P(\text{plaintext} \mid \text{encrypted text}) \propto P(\text{encrypted text} \mid \text{plaintext}) \, P(\text{plaintext}) \]

Bletchley Park (WWII)

Enigma machine
Bayes Rule for text inference

Noisy channel model

Original text

Hypothesized transmission process

Inference problem

Observed data

Codebreaking

\[ P(\text{plaintext} \mid \text{encrypted text}) \propto P(\text{encrypted text} \mid \text{plaintext}) P(\text{plaintext}) \]

Speech recognition

\[ P(\text{text} \mid \text{acoustic signal}) \propto P(\text{acoustic signal} \mid \text{text}) P(\text{text}) \]
Bayes Rule for text inference

Noisy channel model

Hypothesized transmission process

Inference problem

Observed data

Codebreaking

\[ P(\text{plaintext} \mid \text{encrypted text}) \propto P(\text{encrypted text} \mid \text{plaintext}) \, P(\text{plaintext}) \]

Speech recognition

\[ P(\text{text} \mid \text{acoustic signal}) \propto P(\text{acoustic signal} \mid \text{text}) \, P(\text{text}) \]

Optical character recognition

\[ P(\text{text} \mid \text{image}) \propto P(\text{image} \mid \text{text}) \, P(\text{text}) \]
Bayes Rule for text inference

Noisy channel model

Bayes' Rule:

\[ P(plaintext \mid encrypted \text{ text}) = \frac{P(encrypted \text{ text} \mid plaintext) P(plaintext)}{P(encrypted \text{ text})} \]

Codebreaking

P(plaintext)

Speech recognition

P(text \mid acoustic \text{ signal})

Optical character recognition

P(text \mid image)

Noisy channel model

Hypothesized transmission process

Inference problem

Observed data

One naturally wonders if the problem of translation could conceivably be treated as a problem in cryptography. When I look at an article in Russian, I say: ‘This is really written in English, but it has been coded in some strange symbols. I will now proceed to decode.’

-- Warren Weaver (1955)
Bayes Rule for text inference

Noisy channel model  \[ \frac{P(\text{plaintext} \mid \text{encrypted text})}{P(\text{encrypted text} \mid \text{plaintext})} \frac{1}{P(\text{plaintext})} \]

Codebreaking
\[ P(\text{plaintext} \mid \text{encrypted text}) \propto P(\text{encrypted text} \mid \text{plaintext}) P(\text{plaintext}) \]

Speech recognition
\[ P(\text{text} \mid \text{acoustic signal}) \propto P(\text{acoustic signal} \mid \text{text}) P(\text{text}) \]

Optical character recognition
\[ P(\text{text} \mid \text{image}) \propto P(\text{image} \mid \text{text}) P(\text{text}) \]

Machine translation?
\[ P(\text{target text} \mid \text{source text}) \propto P(\text{source text} \mid \text{target text}) P(\text{target text}) \]