Probability & Language Modeling

CMSC 473/673
UMBC
September 6th, 2017

Some slides adapted from 3SLP, Jason Eisner
Administrivia, II
Online Resources

https://csee.umbc.edu/courses/undergraduate/473/f17

https://piazza.com/umbc/fall2017/cmsc473673
Late Policy

Everyone has a budget of 10 *late days*

If you have them left: assignments turned in after the deadline will be graded and recorded, no questions asked

If you don’t have any left: still turn assignments in. They could count in your favor in borderline cases
Submitting Assignments

UMBC’s submit utility

Course id: cs473_ferraro

“Project” names:
  a1, a2, a3, a4
  paper1, paper2
  project_update
  project_final
Recap from last time...
Three people have been fatally shot, and five people, including a mayor, were seriously wounded as a result of a Shining Path attack today.
Three people have been fatally shot, and five people, including a mayor, were seriously wounded as a result of a Shining Path attack today.
Three people have been fatally shot, and five people, including a mayor, were seriously wounded as a result of a Shining Path attack today.
Three people have been fatally shot, and five people, including a mayor, were seriously wounded as a result of a Shining Path attack today.

\( p_{\theta}(\text{what's a probability?}) \)
Three people have been fatally shot, and five people, including a mayor, were seriously wounded as a result of a Shining Path attack today.

what do we estimate?

Documents? Sentences? Words? Characters?
Three people have been fatally shot, and five people, including a mayor, were seriously wounded as a result of a Shining Path attack today.

what’s a word?
how to deal with morphology and orthography
Tree people have been fatally shot, and five people, including a mayor, were seriously wounded as a result of an Shining Path attack today.

how do we estimate robustly?
Three people have been fatally shot, and five people, including a mayor, were seriously wounded as a result of an ISIS attack today.

\[ p_\theta( ) \]

*how do we generalize?*
Outline

Probability review

Words

Defining Language Models

Breaking & Fixing Language Models
Outline

Probability review

Words

Defining Language Models

Breaking & Fixing Language Models
Probability Takeaways

Basic probability axioms and definitions

Probabilistic Independence

Definition of joint probability

Definition of conditional probability

Bayes rule

Probability chain rule
Kinds of Statistics

Descriptive

Confirmatory

Predictive

The average grade on this assignment is 83.
Kinds of Statistics

Descriptive

The average grade on this assignment is 83.

Confirmatory

Predictive
Kinds of Statistics

Descriptive

The average grade on this assignment is 83.

Confirmatory

Predictive
p(heads)
Probabilities Measure Sets

coin coming up heads

all (known) outcomes
Probabilities Measure Sets

(all known) outcomes

cafeteria serves egg salad

coin coming up heads
Probabilities Measure Sets

- Cafeteria serves egg salad
- Defective minting process
- Coin coming up heads

All (known) outcomes
Probabilities Measure Sets

- cafeteria serves egg salad
- coin coming up heads
- defective minting process
- coin is ancient

All (known) outcomes
Probabilities Measure Sets

all (known) outcomes involving coin being flipped

- cafeteria serves egg salad
- defective minting process
- coin is ancient
- coin coming up heads
(Most) Probability Axioms

\[ p(\text{everything}) = 1 \]
(Most) Probability Axioms

\[ p(\text{everything}) = 1 \]

\[ p(\phi) = 0 \]
(Most) Probability Axioms

\[ p(\text{everything}) = 1 \]

\[ p(\emptyset) = 0 \]

\[ p(A) \leq p(B), \text{ when } A \subseteq B \]
(Most) Probability Axioms

\[ p(\text{everything}) = 1 \]
\[ p(\phi) = 0 \]
\[ p(A) \leq p(B), \text{ when } A \subseteq B \]
\[ p(A \cup B) = p(A) + p(B), \text{ when } A \cap B = \phi \]
(Most) Probability Axioms

\[ p(\text{everything}) = 1 \]

\[ p(\emptyset) = 0 \]

\[ p(A) \leq p(B), \text{ when } A \subseteq B \]

\[ p(A \cup B) = p(A) + p(B), \quad \text{when } A \cap B = \emptyset \]

\[ p(A \cup B) \neq p(A) + p(B) \]
(Most) Probability Axioms

\[ p(\text{everything}) = 1 \]
\[ p(\emptyset) = 0 \]
\[ p(A) \leq p(B), \text{ when } A \subseteq B \]
\[ p(A \cup B) = p(A) + p(B), \quad \text{when } A \cap B = \emptyset \]
\[ p(A \cup B) = p(A) + p(B) - p(A \cap B) \]
Probabilities of Independent Events Multiply

\[ p(\text{ancient coin AND defective minting process}) \]
Probabilities of Independent Events Multiply

\[ p(\text{ancient coin AND defective minting process}) = p(\text{ancient coin}) \times p(\text{defective minting process}) \]
Probabilities of Independent Events Multiply

\[ p(\text{ancient coin, defective minting process}) = p(\text{ancient coin}) \times p(\text{defective minting process}) \]
Joint Probabilities Are (Should Be) Symmetric

\[ p(\text{defective minting process, ancient coin}) = p(\text{defective minting process}) \times p(\text{ancient coin}) \]
Conditional Probabilities *(Also)* Measure Sets

\[ p(\text{heads} \mid \text{defective minting process}) \]
Conditional Probabilities (Also) Measure Sets

\[ p(\text{heads} \mid \text{defective minting process}) = \frac{p(\text{heads AND defective minting process})}{p(\text{defective minting process})} \]
Conditional Probabilities (Also) Measure Sets

$p(\text{heads} \mid \text{defective minting process}) = \frac{p(\text{heads AND defective minting process})}{p(\text{defective minting process})}$
Conditional Probabilities (Also) Measure Sets

All (known) outcomes involving coin being flipped

$p(\text{heads} \mid \text{defective minting process}) = \frac{p(\text{heads AND defective minting process})}{p(\text{defective minting process})}$
Conditional Probabilities Are Probabilities

$p(\text{heads} \mid \text{egg salad})$  vs.  $p(\text{heads} \mid \text{NOT egg salad})$
Conditional Probabilities Are Probabilities

\[ p(\text{heads} | \text{egg salad}) \quad \text{vs.} \quad p(\text{heads} | \text{NOT egg salad}) \]

\[ p(\text{heads} | \text{egg salad}) \quad \text{vs.} \quad p(\text{tails} | \text{egg salad}) \]
Conditional Probabilities Are Probabilities

$p(\text{heads} \mid \text{egg salad})$ vs. $p(\text{heads} \mid \text{NOT egg salad})$

$p(\text{heads} \mid \text{egg salad})$ vs. $p(\text{tails} \mid \text{egg salad})$

$p(\text{heads} \mid \text{egg salad})$ vs. $p(\text{tails} \mid \text{NOT egg salad})$
Interpretations of Probability

Past performance
58% of the past 100 flips were heads

Hypothetical performance
If I flipped the coin in many parallel universes...

Subjective strength of belief
Would pay up to 58 cents for chance to win $1

Output of some computable formula?
\( p(\text{heads}) \) vs \( q(\text{heads}) \)
Bayes Rule

\[ p(\text{heads} \mid \text{defective minting process}) = \frac{p(\text{heads AND defective minting process})}{p(\text{defective minting process})} \]
Bayes Rule

\[ p(\text{heads AND defective minting process}) = p(\text{heads | defective minting process}) \times p(\text{defective minting process}) \]
\[ p(\text{heads AND defective minting process}) = p(\text{heads} \mid \text{defective minting process}) \times p(\text{defective minting process}) \]
Bayes Rule

\[ p(\text{heads AND defective minting process}) = p(\text{defective minting process | heads}) \times p(\text{heads}) \]
Bayes Rule

\[ p(\text{heads AND defective minting process}) = p(\text{heads} | \text{defective minting process}) \cdot p(\text{defective minting process}) \]

\[ p(\text{heads AND defective minting process}) = p(\text{defective minting process} | \text{heads}) \cdot p(\text{heads}) \]
Bayes Rule

\[
\begin{align*}
p(\text{heads} \mid \text{defective minting process}) &= \\
&= \frac{p(\text{defective minting process} \mid \text{heads}) \cdot p(\text{heads})}{p(\text{defective minting process})}
\end{align*}
\]
Bayes Rule

\[ p(X \mid Y) = \frac{p(Y \mid X) \times p(X)}{p(Y)} \]
Bayes Rule

\[ p(X \mid Y) = \frac{p(Y \mid X) \cdot p(X)}{p(Y)} \]

posterior probability
Bayes Rule

\[ p(X \mid Y) = \frac{p(Y \mid X) \cdot p(X)}{p(Y)} \]

posterior probability

likelihood
Bayes Rule

\[ p(X \mid Y) = \frac{p(Y \mid X) \ast p(X)}{p(Y)} \]

posterior probability

likelihood

prior probability
Bayes Rule

\[ p(X \mid Y) = \frac{p(Y \mid X) \ast p(X)}{p(Y)} \]

- posterior probability
- likelihood
- prior probability
- marginal likelihood
  (probability)
Changing the Left

$p(A)$
Changing the Left

$p(A)$

$p(A, B)$
Changing the Left

\[ p(A) \]
\[ p(A, B) \]
\[ p(A, B, C) \]
Changing the Left

\[ p(A, B, C, D) \]

\[ p(A, B, C) \]

\[ p(A, B) \]

\[ p(A) \]

\[ p(A, B, C, D, E) \]
Changing the Right

\[ p(A) \]

\[ p(A \mid B) \]
Changing the Right

\[ p(A \mid B) \]

\[ p(A) \]
Changing the Right

$p(A \mid B)$

$p(A)$
Changing the Right

Bias vs. Variance

Lower bias: More specific to what we care about

Higher variance: For fixed observations, estimates become less reliable
Probability Takeaways

Basic probability axioms and definitions

Probabilistic Independence

Definition of joint probability

Definition of conditional probability

Bayes rule

Probability chain rule
Outline

Probability review

Words

Defining Language Models

Breaking & Fixing Language Models
What Are Words?

Linguists don’t agree

(Human) Language-dependent

White-space separation is a sometimes okay (for written English longform)
What Are Words?

Linguists don’t agree

(Human) Language-dependent

White-space separation is a sometimes okay (for written English longform)

Social media? Spoken vs. written? Other languages?
What Are Words?

bat
What Are Words?

bats
What Are Words?

Fledermaus

*flutter mouse*
What Are Words?

pişirdiler

They cooked it.
What Are Words?

pişmişlermişlerdi

They had it cooked it.
What Are Words?

):
What Are Words?

my leg is hurting nasty ): 
What Are Words?

*add two cups (a pint)*: bring to a boil
Examples of Text Normalization

Segmenting or tokenizing words

Normalizing word formats

Segmenting sentences in running text
The film got a great opening and the film went on to become a hit.

**Type**: an element of the vocabulary.

**Token**: an instance of that type in running text.

How many of each?
What Are Words? Tokens vs. Types

The film got a great opening and the film went on to become a hit.

<table>
<thead>
<tr>
<th>Types</th>
<th>Tokens</th>
</tr>
</thead>
<tbody>
<tr>
<td>The</td>
<td>The</td>
</tr>
<tr>
<td>film</td>
<td>film</td>
</tr>
<tr>
<td>got</td>
<td>got</td>
</tr>
<tr>
<td>a</td>
<td>a</td>
</tr>
<tr>
<td>great</td>
<td>great</td>
</tr>
<tr>
<td>opening</td>
<td>opening</td>
</tr>
<tr>
<td>and</td>
<td>and</td>
</tr>
<tr>
<td>the</td>
<td>the</td>
</tr>
<tr>
<td>went</td>
<td>went</td>
</tr>
<tr>
<td>on</td>
<td>on</td>
</tr>
<tr>
<td>to</td>
<td>to</td>
</tr>
<tr>
<td>become</td>
<td>become</td>
</tr>
<tr>
<td>hit</td>
<td>hit</td>
</tr>
<tr>
<td>.</td>
<td>.</td>
</tr>
</tbody>
</table>
What Are Words? Tokens vs. Types

The film got a great opening and the film went on to become a hit.

**Types**
- The
- film
- got
- a
- great
- opening
- and
- the
- went
- on
- to
- become
- hit
- .

**Tokens**
- The
- film
- got
- a
- great
- opening
- and
- the
- film
- went
- on
- to
- become
- a
- hit
- .
Some Issues with Tokenization

mph, MPH, M.D.

MD, M.D.

Baltimore’s mayor

I’m, won’t

state-of-the-art

San Francisco
CaSE inSensitive?

Replace all letters with lower case version

Can be useful for information retrieval (IR), machine translation, language modeling

\textbf{cat vs Cat} (there are other ways to signify beginning)
CaSE inSensitive?

Replace all letters with lower case version

Can be useful for information retrieval (IR), machine translation, language modeling

   cat vs Cat (there are other ways to signify beginning)

But... case **can** be useful

   Sentiment analysis, machine translation, information extraction

   US vs us
cat ⇔ cats

**Lemma**: same stem, part of speech, rough word sense

cat and cats: same lemma

**Word form**: the fully inflected surface form

cat and cats: different word forms
Lemmatization

Reduce inflections or variant forms to base form

\[ \text{am, are, is} \rightarrow \text{be} \]
\[ \text{car, cars, car's, cars'} \rightarrow \text{car} \]

the boy's cars are different colors \(\rightarrow\)
the boy car be different color
Morphosyntax

Morphemes: The small meaningful units that make up words

Stems: The core meaning-bearing units

Affixes: Bits and pieces that adhere to stems
Morphemes: The small meaningful units that make up words

**Stems:** The core meaning-bearing units

**Affixes:** Bits and pieces that adhere to stems

Inflectional:
(they) look $\rightarrow$ (they) looked
(they) ran $\rightarrow$ (they) run

Derivational:
(a) run $\rightarrow$ running (of the Bulls)
code $\rightarrow$ codeable
Morphosyntax

Morphemes: The small meaningful units that make up words
  - **Stems**: The core meaning-bearing units
  - **Affixes**: Bits and pieces that adhere to stems

Inflectional:
(they) look → (they) looked
(they) ran → (they) run

Derivational:
(a) run → running (of the Bulls)
code → codeable

Syntax: Contractions can rewrite and reorder a sentence

Baltimore’s [mayor’s {campaign} ] →
[ {the campaign} of the mayor] of Baltimore
Words vs. Sentences

!, ? are relatively unambiguous

Period “.” is quite ambiguous
  Sentence boundary
  Abbreviations like Inc. or Dr.
  Numbers like .02% or 4.3

Solution: write rules, build a classifier
Outline

Probability review

Words

Defining Language Models

Breaking & Fixing Language Models
“The Unreasonable Effectiveness of Recurrent Neural Networks”
“The Unreasonable Effectiveness of Recurrent Neural Networks”

“The Unreasonable Effectiveness of Character-level Language Models”
“The Unreasonable Effectiveness of Recurrent Neural Networks”

“The Unreasonable Effectiveness of Character-level Language Models” (and why RNNs are still cool)
Simple Count-Based

\( \rho(\text{item}) \)
Simple Count-Based

$p(\text{item}) \propto \text{count}(\text{item})$

“proportional to”
Simple Count-Based

$p(\text{item}) \propto count(\text{item})$

$= \frac{count(\text{item})}{\sum count(\text{other item})}$
Simple Count-Based

\[ p(\text{item}) \propto \text{count}(\text{item}) \]

\[ = \frac{\text{count}(\text{item})}{\sum \text{count}(\text{other item})} \]

constant
Simple Count-Based

\[ p(\text{item}) \propto \text{count}(\text{item}) \]

sequence of characters \(\rightarrow\) pseudo-words

sequence of words \(\rightarrow\) pseudo-phrases
Shakespearian Sequences of Characters
Novel Words, Novel Sentences

“Colorless green ideas sleep furiously” – Chomsky (1957)

Let’s observe and record all sentences with our big, bad supercomputer

Red ideas? Read ideas?
Probability Chain Rule

\[ p(x_1, x_2) = p(x_1)p(x_2 \mid x_1) \]

Bayes rule
Probability Chain Rule

\[ p(x_1, x_2, \ldots, x_S) = p(x_1)p(x_2 | x_1)p(x_3 | x_1, x_2) \cdots p(x_S | x_1, \ldots, x_i) \]
Probability Chain Rule

\[ p(x_1, x_2, \ldots, x_S) = p(x_1)p(x_2 | x_1)p(x_3 | x_1, x_2) \cdots p(x_S | x_1, \ldots, x_i) = \prod_{i=1}^{S} p(x_i | x_1, \ldots, x_{i-1}) \]
Probability Chain Rule

\[
p(x_1, x_2, ..., x_S) = \prod_{i} p(x_i | x_1, ..., x_{i-1})
\]

An extension of Bayes rule.
N-Grams

Maintaining an entire inventory over sentences could be too much to ask

Store “smaller” pieces?

$p($Colorless green ideas sleep furiously$)$
N-Grams

Maintaining an entire *joint* inventory over sentences could be too much to ask

Store “smaller” pieces?

\[ p(\text{Colorless green ideas sleep furiously}) = p(\text{Colorless}) * \]
N-Grams

Maintaining an entire joint inventory over sentences could be too much to ask

Store “smaller” pieces?

\[ p(\text{Colorless green ideas sleep furiously}) = p(\text{Colorless}) \times p(\text{green} \mid \text{Colorless}) \times \]
N-Grams

Maintaining an entire *joint* inventory over sentences could be too much to ask

Store “smaller” pieces?

\[
p(\text{Colorless green ideas sleep furiously}) = \\
p(\text{Colorless}) * \\
p(\text{green | Colorless}) * \\
p(\text{ideas | Colorless green}) * \\
p(\text{sleep | Colorless green ideas}) * \\
p(\text{furiously | Colorless green ideas sleep})
\]
N-Grams

Maintaining an entire joint inventory over sentences could be too much to ask

Store “smaller” pieces?

\[
p(\text{Colorless green ideas sleep furiously}) = \\
p(\text{Colorless}) \ * \\
p(\text{green} \mid \text{Colorless}) \ * \\
p(\text{ideas} \mid \text{Colorless green}) \ * \\
p(\text{sleep} \mid \text{Colorless green ideas}) \ * \\
p(\text{furiously} \mid \text{Colorless green ideas sleep})
\]
N-Grams

Maintaining an entire joint inventory over sentences could be too much to ask

Store “smaller” pieces?

\[
p(\text{Colorless green ideas sleep furiously}) = p(\text{Colorless}) \times p(\text{green} \mid \text{Colorless}) \times p(\text{ideas} \mid \text{Colorless green}) \times p(\text{sleep} \mid \text{Colorless green ideas}) \times p(\text{furiously} \mid \text{Colorless green ideas sleep})
\]
N-Grams

\[ p(\text{furiously} \mid \text{Colorless green ideas sleep}) \]

How much does “Colorless” influence the choice of “furiously?”
N-Grams

$p(\text{furiously} \mid \text{Colorless green ideas sleep})$

How much does “Colorless” influence the choice of “furiously?”

Remove history and contextual info
N-Grams

\[ p(\text{furiously} \mid \text{Colorless green ideas sleep}) \]

How much does “Colorless” influence the choice of “furiously?”

Remove history and contextual info

\[ p(\text{furiously} \mid \text{Colorless green ideas sleep}) \approx p(\text{furiously} \mid \text{Colorless green ideas sleep}) \]
N-Grams

\[ p(\text{furiously} \mid \text{Colorless green ideas sleep}) \]

How much does “Colorless” influence the choice of “furiously?”

Remove history and contextual info

\[ p(\text{furiously} \mid \text{Colorless green ideas sleep}) \approx p(\text{furiously} \mid \text{ideas sleep}) \]
N-Grams

\[
p(\text{Colorless green ideas sleep furiously}) = \]
\[
p(\text{Colorless}) \times \]
\[
p(\text{green | Colorless}) \times \]
\[
p(\text{ideas | Colorless green}) \times \]
\[
p(\text{sleep | Colorless green ideas}) \times \]
\[
p(\text{furiously | Colorless green ideas sleep})
\]
N-Grams

\[
p(\text{Colorless green ideas sleep furiously}) = \\
p(\text{Colorless}) \times \\
p(\text{green | Colorless}) \times \\
p(\text{ideas | Colorless green}) \times \\
p(\text{sleep | Colorless green ideas}) \times \\
p(\text{furiously | Colorless green ideas sleep})
\]
Trigrams

\[ p(\text{Colorless green ideas sleep furiously}) = p(\text{Colorless}) \times p(\text{green | Colorless}) \times p(\text{ideas | Colorless green}) \times p(\text{sleep | green ideas}) \times p(\text{furiously | ideas sleep}) \]
Trigrams

\[
p(\text{Colorless green ideas sleep furiously}) = \\
p(\text{Colorless}) \ast \\
p(\text{green | Colorless}) \ast \\
p(\text{ideas | Colorless green}) \ast \\
p(\text{sleep | green ideas}) \ast \\
p(\text{furiously | ideas sleep})
\]
Trigrams

\[ p(\text{Colorless green ideas sleep furiously}) = \\
p(\text{Colorless} \mid \text{<BOS> <BOS>}) \times \\
p(\text{green} \mid \text{<BOS> Colorless}) \times \\
p(\text{ideas} \mid \text{Colorless green}) \times \\
p(\text{sleep} \mid \text{green ideas}) \times \\
p(\text{furiously} \mid \text{ideas sleep}) \]

*Consistent notation*: Pad the left with <BOS> (beginning of sentence) symbols
Trigrams

\[
p(\text{Colorless green ideas sleep furiously}) = \\
p(\text{Colorless} \mid \text{<BOS> <BOS>}) \ast \\
p(\text{green} \mid \text{<BOS> Colorless}) \ast \\
p(\text{ideas} \mid \text{Colorless green}) \ast \\
p(\text{sleep} \mid \text{green ideas}) \ast \\
p(\text{furiously} \mid \text{ideas sleep}) \ast \\
p(\text{<EOS>} \mid \text{sleep furiously})
\]

*Consistent notation:* Pad the left with *<BOS>* (beginning of sentence) symbols

*Fully proper distribution:* Pad the right with a single *<EOS>* symbol
# N-Gram Terminology

<table>
<thead>
<tr>
<th>n</th>
<th>Commonly called</th>
<th>History Size (Markov order)</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>unigram</td>
<td>0</td>
<td>p(furiously)</td>
</tr>
</tbody>
</table>
# N-Gram Terminology

<table>
<thead>
<tr>
<th>n</th>
<th>Commonly called</th>
<th>History Size (Markov order)</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>unigram</td>
<td>0</td>
<td>p(furiously)</td>
</tr>
<tr>
<td>2</td>
<td>bigram</td>
<td>1</td>
<td>p(furiously</td>
</tr>
</tbody>
</table>
## N-Gram Terminology

<table>
<thead>
<tr>
<th>n</th>
<th>Commonly called</th>
<th>History Size (Markov order)</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>unigram</td>
<td>0</td>
<td>p(furiously)</td>
</tr>
<tr>
<td>2</td>
<td>bigram</td>
<td>1</td>
<td>p(furiously</td>
</tr>
<tr>
<td>3</td>
<td>trigram (3-gram)</td>
<td>2</td>
<td>p(furiously</td>
</tr>
</tbody>
</table>
## N-Gram Terminology

<table>
<thead>
<tr>
<th>n</th>
<th>Commonly called</th>
<th>History Size (Markov order)</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>unigram</td>
<td>0</td>
<td>p(furiously)</td>
</tr>
<tr>
<td>2</td>
<td>bigram</td>
<td>1</td>
<td>p(furiously</td>
</tr>
<tr>
<td>3</td>
<td>trigram (3-gram)</td>
<td>2</td>
<td>p(furiously</td>
</tr>
<tr>
<td>4</td>
<td>4-gram</td>
<td>3</td>
<td>p(furiously</td>
</tr>
<tr>
<td>n</td>
<td>n-gram</td>
<td>n-1</td>
<td>p(w_i</td>
</tr>
</tbody>
</table>