Department of Computer Science
UNIVERSITY OF COLORADO **BOULDER**

Conditional Probability

Introduction to Data Science Algorithms

Jordan Boyd-Graber and Michael Paul

SLIDES ADAPTED FROM PHILIP KOEHN

How do we estimate a probability?

- Suppose we want to estimate $P(w_n = \text{``}home\text{''} | h = \texttt{go})$.

How do we estimate a probability?

- Suppose we want to estimate $P(w_n = \text{``}home\text{''} | h = \mathtt{go})$.

home	**home**	big	with	to
big	with	to	and	money
and	**home**	big	and	**home**
money	**home**	and	big	to

How do we estimate a probability?

- Suppose we want to estimate $P(w_n = \text{“}home\text{''} | h = \texttt{go})$.

home	**home**	big	with	to
big	with	to	and	money
and	**home**	big	and	**home**
money	**home**	and	big	to

- Maximum likelihood (ML) estimate of the probability is:

$$\hat{\theta}_i = \frac{n_i}{\sum_k n_k} \tag{1}$$

Example: 3-Gram

- Counts for trigrams and estimated word probabilities

the red (total: 225)

word	c.	prob.
cross	123	0.547
tape	31	0.138
army	9	0.040
card	7	0.031
,	5	0.022

  - 225 trigrams in the Europarl corpus start with **the red**
  - 123 of them end with **cross**

$\longrightarrow$ maximum likelihood probability is $\frac{123}{225} = 0.547$.

Example: 3-Gram

- Counts for trigrams and estimated word probabilities

the red (total: 225)

word	c.	prob.
cross	123	0.547
tape	31	0.138
army	9	0.040
card	7	0.031
,	5	0.022

  - 225 trigrams in the Europarl corpus start with **the red**
  - 123 of them end with **cross**

$\longrightarrow$ maximum likelihood probability is $\frac{123}{225} = 0.547$.

- Is this reasonable?

The problem with maximum likelihood estimates: Zeros

- If there were no occurrences of "bageling" in a history go, we'd get a zero estimate:

$$\hat{P}(\text{"bageling"} \mid \texttt{go}) = \frac{T_{\texttt{go}, \text{"bageling"}}}{\sum_{w' \in V} T_{\texttt{go}, w'}} = 0$$

- $\rightarrow$ We will get $P(\texttt{go} \mid d) = 0$ for any sentence that contains go bageling!
- Zero probabilities cannot be conditioned away.

Add-One Smoothing

- Equivalent to assuming a **uniform** prior over all possible distributions over the next word (you'll learn why later)
- But there are many more unseen n-grams than seen n-grams
- Example: Europarl 2-bigrams:
  - $86,700$ distinct words
  - $86,700^2 = 7,516,890,000$ possible bigrams
  - but only about $30,000,000$ words (and bigrams) in corpus

More about this later . . .

- MLE vs. MAP (Estimation)
- Bayesian interpretation: prior of distribution
- Fancier smoothing (Knesser-Ney, neural models)

That's it!

- Next time: Language model lab
- Homework 1