Conditional Probability

Introduction to Data Science Algorithms
Jordan Boyd-Graber and Michael Paul
SLIDES ADAPTED FROM PHILIP KOEHN

Language models

- **Language models** answer the question: *How likely is a string of English words good English?*
- Autocomplete on phones and websearch
- Creating English-looking documents
- Very common in machine translation systems
    - Help with reordering / style

$$p_{\mathrm{lm}}(\text{the house is small}) > p_{\mathrm{lm}}(\text{small the is house})$$

    - Help with word choice

$$p_{\mathrm{lm}}(\text{I am going home}) > p_{\mathrm{lm}}(\text{I am going house})$$

    - Use **conditional probabilities**

N-Gram Language Models

- Given: a string of English words $W = w_1, w_2, w_3, ..., w_n$
- Question: what is $p(W)$?
- Sparse data: Many good English sentences will not have been seen before

$\longrightarrow$ Decomposing $p(W)$ using the chain rule:

$$p(w_1, w_2, w_3, ..., w_n) = \\ p(w_1)\, p(w_2|w_1)\, p(w_3|w_1, w_2) \ldots p(w_n|w_1, w_2, ...w_{n-1})$$

(not much gained yet, $p(w_n|w_1, w_2, ...w_{n-1})$ is equally sparse)

Markov Chain

- **Markov independence assumption**:
  - only previous history matters
  - limited memory: only last k words are included in history (older words less relevant)

  $\longrightarrow$ ***k*th order Markov model**
- For instance 2-gram language model:

$$p(w_1, w_2, w_3, ..., w_n) \simeq p(w_1)\, p(w_2|w_1)\, p(w_3|w_2) ... p(w_n|w_{n-1})$$

- What is conditioned on, here w_{i-1} is called the **history**
- How do we estimate these probabilities?