

Exercise set 5

Thursday, 26 October 2023 15.08

Exercise 5.1: Bigram probabilities.

- (a) Are the following probabilities possible in an bigram model? $p(w_1 = \text{'rock'}) = 0.01$, $p(w_2 = \text{'band'}) = 0.003$, $p(w_2 = \text{'band'} | w_1 = \text{'rock'}) = 0.4$. Prove why/why not. Derive an inequality between $p(w_1)$, $p(w_2)$ and $p(w_2 | w_1)$ for what probabilities are possible. Hint: consider the Bayes rule.
- (b) Consider the sentence "The whole of science is nothing more than a refinement of everyday thinking." (Albert Einstein, *Physics and Reality*, 1936). Compute the probability of the sentence in a bigram model using the following unigram probabilities: $p(\text{'the'}) = 0.03$, $p(\text{'whole'}) = 0.0001$, $p(\text{'of'}) = 0.01$, $p(\text{'science'}) = 0.0003$, $p(\text{'is'}) = 0.02$, $p(\text{'nothing'}) = 0.0002$, $p(\text{'more'}) = 0.001$, $p(\text{'than'}) = 0.0009$, $p(\text{'a'}) = 0.025$, $p(\text{'refinement'}) = 2 \cdot 10^{-6}$, $p(\text{'everyday'}) = 6 \cdot 10^{-6}$, $p(\text{'thinking'}) = 3 \cdot 10^{-5}$. You need to choose some corresponding bigram probabilities so that they satisfy the condition you derived in (a).

Report your proof and computations.

$$\begin{aligned} a) \quad p(w_1 = \text{'rock'}) &= 0.01 \\ p(w_2 = \text{'band'}) &= 0.003 \\ p(w_2 = \text{'band'} | w_1 = \text{'rock'}) &= 0.4 \\ p(w_1 = \text{'rock'} | w_2 = \text{'band'}) &= \frac{0.4 \cdot 0.01}{0.003} = 1.33 > 1 \end{aligned}$$

→ The following prob is not possible in a bigram model.

$$\begin{aligned} * \quad p(w_1, w_2) &= p(w_2 | w_1) \cdot p(w_1) \\ \rightarrow \quad p(w_2 | w_1) &= \frac{p(w_1, w_2)}{p(w_1)} \leq 1 \\ &> 0 \end{aligned}$$

$$b) \quad p(\text{sentence}) = p(\text{the}) \cdot p(\text{whole} | \text{the}) \cdot p(\text{of} | \text{whole})$$

$$\dots p(\text{everyday} | \text{of}) \cdot p(\text{thinking} | \text{everyday})$$

$$* \quad p(\text{whole} | \text{the}) = \frac{p(\text{the} | \text{whole}) \cdot p(\text{whole})}{p(\text{the})}$$

...

$$\rightarrow p(\text{sentence}) = p(\text{the}) \cdot \frac{p(\text{whole})}{p(\text{the})} \cdot p(\text{the} | \text{whole}) \cdot \frac{p(\text{of})}{p(\text{whole})} \cdot p(\text{whole} | \text{of}) \dots$$

$$\frac{p(\text{of})}{p(\text{everyday})} \cdot p(\text{of} | \text{everyday}) \cdot \frac{p(\text{everyday})}{p(\text{thinking})} \cdot p(\text{everyday} | \text{thinking})$$

$$\begin{aligned} p(\text{sentence}) &= \frac{1}{p(\text{thinking})} \cdot p(\text{the} | \text{whole}) \dots p(\text{of} | \text{everyday}) \cdot p(\text{everyday} | \text{thinking}) \\ &= \frac{1}{3 \cdot 10^{-5}} \cdot \underline{\hspace{10cm}} \end{aligned}$$

The result of $p(\text{sentence})$ should depend on the unknown value of $p(\text{the} | \text{whole})$, ... and $p(\text{thinking})$
and $0 \leq p(\text{sentence}) \leq 1$.

Exercise 5.2: Theoretical n-gram properties.

- (a) Suppose you need to generate a document of length M words. Show that if the n in an n -gram model is at least as large as M , the n -gram model can represent all statistical dependencies that might exist in the language needed to generate the document. So that, for example, a 5-gram model can represent all dependencies needed to generate sentences of 5 words.
- (b) Consider a simplified version of the maximum a posteriori estimation of n -gram probabilities described on the lecture. Suppose all pseudocounts in the Dirichlet priors use the same shared value, $\alpha_{v[[w_1, \dots, w_{n-1}]]} = \alpha_{\text{shared}}$ for all vocabulary terms v and all contexts $[w_1, \dots, w_{n-1}]$ where α_{shared} is the shared value. This results in estimates that are simple smoothed proportional counts. This kind of smoothing is called **Laplace smoothing** when $\alpha_{\text{shared}} = 1$ and **Lidstone smoothing** otherwise.
- Show that in this setting, the maximum a posteriori estimate (as shown on the course slides) for a n -gram probability can be written as a weighted average of two terms: (1) the maximum likelihood estimate of the probability and (2) a uniform distribution over the vocabulary.
 - Show that the mixing weight in the weighted average depends on the number of occurrences of a n -gram context compared to $V\alpha_{\text{shared}}$ where V is the vocabulary size.
 - For an individual n -gram context, how should α_{shared} be chosen so that the weight of the data is greater than the weight of the prior?

Report your proofs.

(exercises continue on the next page)

a) n -gram model, size M

$$p(\text{document}) = p(w_1) \cdot p(w_2|w_1) \cdots p(w_M | w_{M-1} \dots w_1) \quad (\text{chain rule})$$

→ therefore, we could observe the dependency of each word with the next word, leading to the dependencies of the n -gram model

b) maximum a posteriori estimation:

Weighting average of NLE

$$\alpha_{\text{shared}} = \alpha_{v[[w_1, w_2, \dots, w_{n-1}]]}$$

$$\theta_{\text{map}, v} = \frac{n_v + \alpha_{\text{shared}}}{n + \sum \alpha_{\text{shared}}} = \frac{n}{n + \sum \alpha_{\text{shared}}} \cdot \frac{n_v}{n} + \frac{\alpha_{\text{shared}}}{\sum \alpha_{\text{shared}}} \cdot \left(1 - \frac{n}{n + \sum \alpha_{\text{shared}}}\right)$$

• Mixing weight

$$\text{We have } \frac{n}{n + \sum \alpha_{\text{shared}}} = \frac{n}{n + V\alpha_{\text{shared}}} \quad (\text{since } V \text{ is the vocabulary size})$$

• Choosing α_{shared}

$$\text{Assumption: } \frac{n}{n + V\alpha_{\text{shared}}} > 1 - \frac{n}{n + V\alpha_{\text{shared}}} \rightarrow \frac{n}{n + V\alpha_{\text{shared}}} > \frac{1}{2}$$

$$\rightarrow n > V\alpha_{\text{shared}} \rightarrow \frac{n}{V} > \alpha_{\text{shared}}$$

→ choosing α_{shared} to be smaller than $\frac{n}{V}$

Exercise 5.4: N-gram abilities versus AI chatbot abilities

- Can n-grams learn to write text from a corpus where the order of letters in words has been reversed? (For example, "Ti saw eth tseb fo semit, ti saw eth tsrow fo semit.") Why/why not? If they can, what limitations are there in what can be learned?
- Can n-grams learn to write text from a corpus where each letter has been rotated one step backwards in the alphabet (For example "Hs vzs sgd adrs pe shldr, Hs vzs sgd vnqrs pe shldr.") Why/why not? If they can, what limitations are there in what can be learned?
- Can n-grams learn to write text from a corpus where the order of the intermediate letters has been randomly scrambled? (For example, "Dolaonwd the flionwlog ekobos form Pjeocrt Gteubrneg".) Why/why not? If they can, what limitations are there in what can be learned?
- Can n-grams learn the general rule that an opening parenthesis must be followed by a closing parenthesis? Why/why not? If they can, what limitations are there in what can be learned?
- Prove that n-gram models can be used to predict a missing middle word in a string like "the cat w into the box" where w is a missing word, and "the cat"=Left and "into the box"=Right are its contexts to the left and right. To do so, prove that the probability $p(w|Left, Right)$ can be rewritten using only n-gram style probabilities that give the probability of a word given its n-1 previous words. Hint: use the Bayes rule and the chain rule of probabilities.
- Optional extra part: ask a modern deep generative chatbot such as ChatGPT or <https://deepai.org/chat> (or any local offline chatbot), to continue the example sentences "Ti saw eth tseb fo semit, ti saw eth tsrow fo semit.", "Hs vzs sgd adrs pe shldr, Hs vzs sgd vnqrs pe shldr", and "Dolaonwd the flionwlog ekobos form Pjeocrt Gteubrneg". Discuss the results.

- a. Yes, since n-grams are based on the statistical properties of text and do not have any understanding of the meaning of words. In other words, it will generate the words based on the probabilities of the generated words.

However, the generated text may also have its words reserved, which could not make sense when we read.

- b. Yes as discussed in question a.

However, the generated text should be unable to be understood, since, the probabilities of the words are not really related. Therefore, we need to manually check for it. The limitation is the same in question a

- c. No, since the same as question d

- d. No, since as discussed in question a: n-grams are based on the statistical properties of text

- e. We could use Bayes' rule to prove that (below)

- f. The answer from ChatGPT:

"

Certainly! Here are continued sentences for the examples you provided:

"Ti saw eth tseb fo semit, ti saw eth tsrow fo semit. With every step, he felt the weight of time pressing on his shoulders."

"Hs vzs sgd adrs pe shldr, Hs vzs sgd vnqrs pe shldr. The burden of responsibility never seemed to wane, but he carried it with unwavering determination."

"Download the following e-books from Project Gutenberg to enrich your literary collection."

"

The model can generate content, that is mostly similar to the last question, and not familiar with the first and second questions. It may have some struggle with these examples since they are trained to understand the words.

$$e) p(w | \text{left}, \text{right}) = \frac{p(\text{left}^+, w, \text{right})}{p(\text{left}^+, \text{right}^-)}$$

$$= \frac{p(\text{left}^+). p(w | \text{left}^+). p(\text{right}^- | \text{left}^+)}{p(\text{left}^+). p(\text{right}^- | \text{left}^+)} \quad |$$

$$= \frac{p(\text{leg}^+) p(\text{right} | \text{leg}^+)}{p(\text{w} | \text{leg}^+) p(\text{right} | \text{leg}^+, \text{w})}$$

Assume that w does not influence right given legr $\rightarrow p(\text{right} | \text{leg}^+, \text{w}) = p(\text{right} | \text{leg}^+)$

$$\rightarrow = p(\text{w} | \text{leg}^+) \rightarrow \text{QED.}$$