

LING/COMP 445, LING 645

Problem Set 5

Name: , McGill ID:
Collaborators:

Due before 4:35 PM on Wednesday, November 22, 2023

Please enter your name and McGill ID above.

This problem set consists only of questions involving mathematics or English or or a combination of the two (no coding questions this time). Please put your answers in an answer box like in the example below.

Once you have entered your answers, please compile your copy of this L^AT_EX into a PDF and submit

- (i) the compiled PDF renamed to `ps5-lastname-firstname.pdf` and
- (ii) the raw L^AT_EX file renamed to `ps5-lastname-firstname.tex`

to the Problem Set 5 folder under ‘Assignments’ on MyCourses.

Example Problem: This is an example question using some fake math like this $L = \sum_0^\infty \mathcal{G}\delta_x$.

Example Answer: Put your answer in the box provided, like this:

Example answer is $L = \sum_0^\infty \mathcal{G}\delta_x$.

Problem 1: In class we gave the following equation for the bigram probability of a sequence of words $W^{(1)}, \dots, W^{(k)}$:

$$\Pr(W^{(1)}, \dots, W^{(k)}) \stackrel{\text{def}}{=} \prod_i^k \Pr(W^{(i)} | W^{(i-1)} = w^{(i-1)}) \quad (1)$$

Using this formula, give an expression for the bigram probability of the sentence *abab*, where each character is treated as a word. Try to simplify the formula as much as possible.

Important note: Throughout this problem set, the vocabulary will be $V \stackrel{\text{def}}{=} \{a, b\}$. We will assume the length of the sentence is fixed at some k , and *we will not use the stop symbol*. That is, in a sentence of length k , for $1 \leq i \leq k$, the possible values for the random variable $W^{(i)}$ are just a and b , and we will refer to the beginning of the string as $W^{(0)} = \times$ always. So, $\Pr(W^{(1)} = a | W^{(0)} = \times)$ is the probability that the string starts with a .

Answer 1: Please put your answer in the box below.

Problem 2: There are two possible symbols/words in our language, a and b . There are three conditional distributions in the bigram model for this language, $\Pr(W^{(i)} | W^{(i-1)} = a)$, $\Pr(W^{(i)} | W^{(i-1)} = b)$, and $\Pr(W^{(i)} | W^{(i-1)} = \times)$. These conditional distributions are associated with the parameter vectors $\vec{\theta}_a$, $\vec{\theta}_b$, and $\vec{\theta}_\times$, respectively (these parameter vectors were implicit in the previous problem). For the current problem, we will assume that these parameters are fixed. Use a second subscript notation to denote components of these vectors, so $\theta_{ab} = \Pr(W^{(i)} = b | W^{(i-1)} = a)$.

Suppose that we are given a sentence $W^{(1)}, \dots, W^{(k)}$. We will use the notation $n_{x \rightarrow y}$ to denote the number of times that the symbol y occurs immediately following the symbol x in the sentence. For example, $n_{a \rightarrow a}$ counts the number of times that symbol a occurs immediately following the symbol a . Using Equation 1, give an expression for the probability of a length k sentence in our language:

$$\Pr(W^{(1)}, \dots, W^{(k)} | \vec{\theta}_a, \vec{\theta}_b, \vec{\theta}_\times)$$

The expression should make use of the $n_{x \rightarrow y}$ notation defined above.

Hint: the expression should be analogous to the formula that we found for the likelihood of a corpus under a bag of words model.

Answer 2: Please put your answer in the box below.

Problem 3: Assume the parameter vectors in our bigram model have the following values:

$$\begin{aligned} \vec{\theta}_a &= (0.7, 0.3) \\ \vec{\theta}_b &= (0.2, 0.8) \\ \vec{\theta}_\times &= (0.5, 0.5) \end{aligned}$$

The first vector indicates that if the current symbol a , there is probability 0.7 of transitioning to the symbol a , and probability 0.3 of transitioning to the symbol b . Using your answer to the previous problem and these parameter values, calculate the probability of the string $aabab$.

Answer 3: Please put your answer in the box below.

Problem 4: In Problem 2, you found an expression for the bigram probability of a sentence in our language, which contains the symbols a and b . In that problem, we assumed that there were fixed parameter vectors $\vec{\theta}$ associated with each conditional distribution. In this problem, we will consider the setting in which we have uncertainty about the value of these parameters.

As we did in class, we will use the Dirichlet distribution to define a prior over parameters. Assume each parameter vector is drawn independently given $\vec{\alpha}$:

$$\vec{\theta}_c \mid \vec{\alpha} \sim \text{Dirichlet}(\vec{\alpha}) \quad (2)$$

$$w^{(i)} \mid w^{(i-1)} \sim \text{categorical}(\vec{\theta}_{w^{(i-1)}}) \quad (3)$$

$$w^{(1)} \sim \text{categorical}(\vec{\theta}_{\times}) \quad (4)$$

Suppose that we have a fixed-length sentence $S = W^{(1)}, \dots, W^{(k)}$. Give an expression for the joint probability $\Pr(S, \vec{\theta}_a, \vec{\theta}_b, \vec{\theta}_{\times} \mid \vec{\alpha})$ using the definitions of Dirichlet distributions and likelihoods we defined in class.

Hint: The joint probability $\Pr(S, \vec{\theta}_a, \vec{\theta}_b, \vec{\theta}_{\times} \mid \vec{\alpha}) = \Pr(S \mid \vec{\theta}_a, \vec{\theta}_b, \vec{\theta}_{\times}) \Pr(\vec{\theta}_a, \vec{\theta}_b, \vec{\theta}_{\times} \mid \vec{\alpha})$, because S is conditionally independent of $\vec{\alpha}$ given the parameters. You solved $\Pr(S \mid \vec{\theta}_a, \vec{\theta}_b, \vec{\theta}_{\times})$ in Problem 2.

Answer 4: Please put your answer in the box below.

Problem 5: In the previous problem, you found a formula for the joint probability of a sentence and a set of bigram model parameters. Using this, give a formula for the marginal probability of the sentence $\Pr(S \mid \vec{\alpha})$.

Hint: The formula should be analogous to the formula derived in class for marginal probability of a corpus under a bag of words model. Whereas before there was only a single parameter vector $\vec{\theta}$, now there are three parameter vectors that need to be marginalized away. Otherwise the calculation will be similar.

Answer 5: Please put your answer in the box below.

Problem 6: Let us assume that the parameters of the Dirichlet distribution are $\vec{\alpha} = (1, 1)$. Using your solution to the previous problem, write an expression for $\Pr(S = aabab \mid \vec{\alpha} = (1, 1))$, the marginal probability of the string $aabab$. The expression should contain the [gamma function](#) $\Gamma(\cdot)$. Using the properties of the gamma function discussed in class (i.e., its relationship to the factorial) or an online calculator, compute a numerical value for this expression.

Answer 6: Please put your answer in the box below.

Problem 7: Suppose that we have observed a sentence $S = W^{(1)}, \dots, W^{(k)}$. Find an expression for the posterior distribution over the model parameters, $\Pr(\vec{\theta}_a, \vec{\theta}_b, \vec{\theta}_x \mid S, \vec{\alpha})$.

Hint: Use the joint probability that you computed in Problem 4 and Bayes' rule. The solution should be analogous to the posterior probability for the bag of words model.

Answer 7: Please put your answer in the box below.

Problem 8: Consider the language $L = \{a^*ba^*\}$, that is, the language consisting of some number of a's, followed by a single b, followed by some number of a's. Show that this language is not strictly 2-local.

Hint: use n -Local Suffix Substitution Closure (n -SSC).

Answer 9: Please put your answer in the box below.

Problem 9: Consider the language $L = \{a^n b^m c^n d^m \mid n, m \in \mathbb{N}\}$, that is, the language consisting of n a's followed by m b's followed by n c's followed by m d's where n and m are natural numbers. Show that this language is not strictly 2-local.

Hint: use the same property as in the problem above.

Answer 10: Please put your answer in the box below.

Problem 10: Show that the language $L = \{a^n b^m c^n d^m \mid n, m \in \mathbb{N}\}$ is not strictly k -local, for any value of k .

Answer 11: Please put your answer in the box below.

Problem 11: In class we proved that $L \in \text{SL}_k \implies k\text{-SSC}(L)$. In other words, if a language is k -strictly local, then it satisfies k -Local Suffix Substitution Closure.

Use this theorem to prove that k -strictly local languages are closed under intersection. More precisely, prove that if $L_1 \in \text{SL}_k$ and $L_2 \in \text{SL}_k$, then $L_1 \cap L_2 \in \text{SL}_k$.

Answer 12: Please put your answer in the box below.

LONG FORM READING QUESTION:

(This section is optional for students in LING/COMP 445, but must be completed if taking LING 645.)

You must answer this question on your own.

Delétang et al. (2023) demonstrate the generative capacities of different neural network language models (RNNs, LSTMs, Transformers, and memory enhanced RNNs) by trying to situate them in practice along the Chomsky hierarchy. Large language models like GPT-3 and Bard are transformer-based architectures. On the surface, such models generate natural sounding language (Have you tried ChatGPT?). Given what you have learnt about these types of architectures from Delétang et al.'s experiments, do you think we should expect these models to have learnt grammatical systems for natural language that are similar to the generative systems we think people may have (think generative grammars)? Or are these models likely doing some different entirely? Motivate your answer with at least 2 observations supporting your position.

Answer: Please put your answer in the box below.
