

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

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

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.

Answer 1:

$$\begin{aligned} \Pr(W^{(1)}, \dots, W^{(k)}) &= \prod_i^k \Pr(W^{(i)} | W^{(i-1)}) \\ &= \Pr(a | \times) \cdot \Pr(b | a) \cdot \Pr(a | b) \cdot \Pr(b | a) \cdot \Pr(\times | b) \\ &= \Pr(a | \times) \cdot (\Pr(b | a))^2 \cdot \Pr(a | b) \cdot \Pr(\times | b) \end{aligned}$$

Problem 2: Suppose that 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.

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 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:

$$\begin{aligned}\Pr(W^{(1)}, \dots, W^{(k)} | \vec{\theta}_a, \vec{\theta}_b, \vec{\theta}_\times) &= \Pr(S | \vec{\theta}_a, \vec{\theta}_b, \vec{\theta}_\times) \\ &= \Pr(S | \vec{\theta}_a) \cdot \Pr(S | \vec{\theta}_b) \cdot \Pr(S | \vec{\theta}_\times) \\ &= \prod_{w \in V} \theta_{a \rightarrow w}^{n_{a \rightarrow w}} \cdot \prod_{w \in V} \theta_{b \rightarrow w}^{n_{b \rightarrow w}} \cdot \prod_{w \in V} \theta_{\times \rightarrow w}^{n_{\times \rightarrow w}}\end{aligned}$$

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:

$$\begin{aligned}\Pr(W^{(1)}, \dots, W^{(k)} | \vec{\theta}_a, \vec{\theta}_b, \vec{\theta}_\times) &= \prod_{w \in V} \theta_{a \rightarrow w}^{n_{a \rightarrow w}} \cdot \prod_{w \in V} \theta_{b \rightarrow w}^{n_{b \rightarrow w}} \cdot \prod_{w \in V} \theta_{\times \rightarrow w}^{n_{\times \rightarrow w}} \\ &= \theta_{\times \rightarrow a} \cdot \theta_{a \rightarrow a} \cdot \theta_{a \rightarrow b}^2 \cdot \theta_{b \rightarrow a} \\ &= 0.5 \cdot 0.7 \cdot (0.3)^2 \cdot 0.2 \\ &= 0.0063\end{aligned}$$

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.

In class, we used the Dirichlet distribution to define a prior distribution over parameter vectors:

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

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

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

Suppose that we have a sentence $S = W^{(1)}, \dots, W^{(k)}$. Given 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.

Answer 4:

$$A = a \rightarrow w$$

$$B = b \rightarrow w$$

$$S = \times \rightarrow w$$

$$\begin{aligned} \Pr(S, \vec{\theta}_a, \vec{\theta}_b, \vec{\theta}_\times \mid \vec{\alpha}) &= \Pr(S \mid \vec{\theta}_a, \vec{\theta}_b, \vec{\theta}_\times) \cdot \Pr(\vec{\theta}_a, \vec{\theta}_b, \vec{\theta}_\times \mid \vec{\alpha}) \\ &= \Pr(S \mid \vec{\theta}_a) \cdot \Pr(\vec{\theta}_a \mid \vec{\alpha}) \cdot \Pr(S \mid \vec{\theta}_b) \cdot \Pr(\vec{\theta}_b \mid \vec{\alpha}) \cdot \Pr(S \mid \vec{\theta}_\times) \cdot \Pr(\vec{\theta}_\times \mid \vec{\alpha}) \\ &= \left(\frac{\Gamma(\sum_{w \in V} \alpha_w)}{\prod_{w \in V} \Gamma(\alpha_w)} \right)^3 \left(\prod_{w \in V} \theta_A^{n_A + \alpha_A - 1} \cdot \prod_{w \in V} \theta_B^{n_B + \alpha_B - 1} \cdot \prod_{w \in V} \theta_S^{n_S + \alpha_S - 1} \right) \end{aligned}$$

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 parameters that need to be marginalized away. Otherwise the calculation will be similar.

Answer 5: $C(\vec{\alpha}_w)$ Notation OK'd by O'Donnell

$$\begin{aligned} C(\vec{\alpha}_w) &= \left(\frac{\Gamma(\sum_{w \in V} \alpha_w)}{\prod_{w \in V} \Gamma(\alpha_w)} \right)^3 \\ A &= a \rightarrow w \\ B &= b \rightarrow w \\ S &= \times \rightarrow w \end{aligned}$$

$$\begin{aligned}
\Pr(S|\vec{\alpha}) &= \int_{\theta} \Pr(S|\vec{\theta}_a, \vec{\theta}_b, \vec{\theta}_{\times}) \cdot \Pr(\vec{\theta}_a, \vec{\theta}_b, \vec{\theta}_{\times}|\vec{\alpha}) \\
&= \int_{\theta_a} \Pr(S|\vec{\theta}_a) \cdot \Pr(\vec{\theta}_a|\vec{\alpha}) \cdot \int_{\theta_b} \Pr(S|\vec{\theta}_b) \cdot \Pr(\vec{\theta}_b|\vec{\alpha}) \cdot \int_{\theta_{\times}} \Pr(S|\vec{\theta}_{\times}) \cdot \Pr(\vec{\theta}_{\times}|\vec{\alpha}) \\
&= C(\vec{\alpha}_w) \cdot \left(\int_{\theta_a} \prod_{w \in V} \theta_A^{n_A + \alpha_A - 1} \cdot \int_{\theta_b} \prod_{w \in V} \theta_B^{n_B + \alpha_B - 1} \cdot \int_{\theta_{\times}} \prod_{w \in V} \theta_S^{n_S + \alpha_S - 1} \right) \\
&= C(\vec{\alpha}_w) \cdot \left(\frac{\prod_{w \in V} \Gamma(n_A + \alpha_A)}{\Gamma(\sum_{w \in V} (n_A + \alpha_A))} \cdot \frac{\prod_{w \in V} \Gamma(n_B + \alpha_B)}{\Gamma(\sum_{w \in V} (n_B + \alpha_B))} \cdot \frac{\prod_{w \in V} \Gamma(n_S + \alpha_S)}{\Gamma(\sum_{w \in V} (n_S + \alpha_S))} \right)
\end{aligned}$$

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

Answer 6:

$$\begin{aligned}
C((1, 1)) &= \left(\frac{\Gamma(2)}{\Gamma(1)\Gamma(1)} \right)^3 = 1 \\
A &= a \rightarrow w \\
B &= b \rightarrow w \\
S &= \times \rightarrow w \\
\Pr(S|\vec{\alpha}) &= \frac{\prod_{w \in V} \Gamma(n_A + \alpha_A)}{\Gamma(\sum_{w \in V} (n_A + \alpha_A))} \cdot \frac{\prod_{w \in V} \Gamma(n_B + \alpha_B)}{\Gamma(\sum_{w \in V} (n_B + \alpha_B))} \cdot \frac{\prod_{w \in V} \Gamma(n_S + \alpha_S)}{\Gamma(\sum_{w \in V} (n_S + \alpha_S))} \\
\Pr(aabab|(1, 1)) &= \frac{\Gamma(1+1) \cdot \Gamma(2+1)}{\Gamma(1+1+2+1)} \cdot \frac{\Gamma(1+0) \cdot \Gamma(1+1)}{\Gamma(1+0+1+1)} \cdot \frac{\Gamma(1+0) \cdot \Gamma(1+1)}{\Gamma(1+0+1+1)} \\
&= \frac{\Gamma(2) \cdot \Gamma(3)}{\Gamma(5)} \cdot \frac{\Gamma(1) \cdot \Gamma(2)}{\Gamma(3)} \cdot \frac{\Gamma(1) \cdot \Gamma(2)}{\Gamma(3)} \\
&= \frac{1 \cdot 2}{24} \cdot \frac{1 \cdot 1}{2} \cdot \frac{1 \cdot 1}{2} \\
&= \frac{1}{12} \cdot \frac{1}{2} \cdot \frac{1}{2} \\
&= \frac{1}{48}
\end{aligned}$$

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, $\Pr(\vec{\theta}_a, \vec{\theta}_b, \vec{\theta}_\times \mid S, \vec{\alpha})$.

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: $C(\vec{\alpha}_w)$ Notation OK'd by O'Donnell

$$C(\vec{\alpha}_w) = \left(\frac{\Gamma(\sum_{w \in V} \alpha_w)}{\prod_{w \in V} \Gamma(\alpha_w)} \right)^3$$

$$A = a \rightarrow w$$

$$B = b \rightarrow w$$

$$S = \times \rightarrow w$$

$$\begin{aligned} \Pr(\vec{\theta}_a, \vec{\theta}_b, \vec{\theta}_\times \mid S, \vec{\alpha}) &= \frac{\Pr(S, \vec{\theta}_a, \vec{\theta}_b, \vec{\theta}_\times \mid \vec{\alpha})}{\Pr(S)} \\ &= \frac{\Pr(S \mid \vec{\theta}_a, \vec{\theta}_b, \vec{\theta}_\times) \cdot \Pr(\vec{\theta}_a, \vec{\theta}_b, \vec{\theta}_\times \mid \vec{\alpha})}{\int_{\theta} \Pr(S \mid \vec{\theta}_a, \vec{\theta}_b, \vec{\theta}_\times) \cdot \Pr(\vec{\theta}_a, \vec{\theta}_b, \vec{\theta}_\times \mid \vec{\alpha})} \\ &= \frac{C(\vec{\alpha}_w) \cdot \left(\prod_{w \in V} \theta_A^{n_A + \alpha_A - 1} \cdot \prod_{w \in V} \theta_B^{n_B + \alpha_B - 1} \cdot \prod_{w \in V} \theta_S^{n_S + \alpha_S - 1} \right)}{C(\vec{\alpha}_w) \cdot \left(\frac{\prod_{w \in V} \Gamma(n_A + \alpha_A)}{\Gamma(\sum_{w \in V} (n_A + \alpha_A))} \cdot \frac{\prod_{w \in V} \Gamma(n_B + \alpha_B)}{\Gamma(\sum_{w \in V} (n_B + \alpha_B))} \cdot \frac{\prod_{w \in V} \Gamma(n_S + \alpha_S)}{\Gamma(\sum_{w \in V} (n_S + \alpha_S))} \right)} \\ &= \frac{\prod_{w \in V} \theta_A^{n_A + \alpha_A - 1} \cdot \prod_{w \in V} \theta_B^{n_B + \alpha_B - 1} \cdot \prod_{w \in V} \theta_S^{n_S + \alpha_S - 1}}{\frac{\prod_{w \in V} \Gamma(n_A + \alpha_A)}{\Gamma(\sum_{w \in V} (n_A + \alpha_A))} \cdot \frac{\prod_{w \in V} \Gamma(n_B + \alpha_B)}{\Gamma(\sum_{w \in V} (n_B + \alpha_B))} \cdot \frac{\prod_{w \in V} \Gamma(n_S + \alpha_S)}{\Gamma(\sum_{w \in V} (n_S + \alpha_S))}} \end{aligned}$$

Problem 8: Using your formula for the posterior distribution and setting $\vec{\alpha} = (1, 1)$, calculate the posterior distribution over parameters given the sentence *aabab*. There should be an easy way to interpret the posterior distribution, and how it was derived from the observed sentence. What is this interpretation?

Answer 8:

$$A = a \rightarrow w$$

$$B = b \rightarrow w$$

$$S = \times \rightarrow w$$

$$\begin{aligned}
\Pr(\vec{\theta}_a, \vec{\theta}_b, \vec{\theta}_{\bowtie} \mid S, \vec{\alpha}) &= \frac{\prod_{w \in V} \theta_A^{n_A + \alpha_A - 1} \cdot \prod_{w \in V} \theta_B^{n_B + \alpha_B - 1} \cdot \prod_{w \in V} \theta_S^{n_S + \alpha_S - 1}}{\frac{\prod_{w \in V} \Gamma(n_A + \alpha_A)}{\Gamma(\sum_{w \in V} (n_A + \alpha_A))} \cdot \frac{\prod_{w \in V} \Gamma(n_B + \alpha_B)}{\Gamma(\sum_{w \in V} (n_B + \alpha_B))} \cdot \frac{\prod_{w \in V} \Gamma(n_S + \alpha_S)}{\Gamma(\sum_{w \in V} (n_S + \alpha_S))}} \\
&= \frac{\prod_{w \in V} \theta_A^{n_A + \alpha_A - 1} \cdot \prod_{w \in V} \theta_B^{n_B + \alpha_B - 1} \cdot \prod_{w \in V} \theta_S^{n_S + \alpha_S - 1}}{\frac{1}{48}} \\
&= 48 \cdot \left(\prod_{w \in V} \theta_A^{n_A + \alpha_A - 1} \cdot \prod_{w \in V} \theta_B^{n_B + \alpha_B - 1} \cdot \prod_{w \in V} \theta_S^{n_S + \alpha_S - 1} \right) \\
&= 48 \cdot (\theta_{\bowtie \rightarrow a} \cdot \theta_{a \rightarrow a} \cdot \theta_{a \rightarrow b}^2 \cdot \theta_{b \rightarrow a})
\end{aligned}$$

The posterior and Bayesian inference gives us a probability distribution over our models instead of an estimate of a single model.

Problem 9: 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:

$$\begin{array}{ccccccc}
\times & a & a & b & \mid & a & \times \\
& & & & \times & \mid & \\
& & & & & a & b & a & a & \times
\end{array}$$

Our new string is now $\times aababaa \times \notin L$. Thus, $L \notin SL_2$.

Problem 10: 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:

$$\begin{array}{ccccccc}
\times & a & b & \mid & c & \mid & d & \times \\
& & & \times & \mid & c & \times
\end{array}$$

Our new string is now $\times abc \times \notin L$. Thus, $L \notin SL_2$.

Problem 11: 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:

$$\begin{array}{ccccccc} \times & a^{k-1} & \mid & b^{k-1} & \mid & c^{k-1} & d^{k-1} & \times \\ \times & a^{k-2} & \mid & b^{k-1} & \mid & c^{k-2} & d^{k-1} & \times \end{array}$$

Our new string is now $\times a^{k-1} b^{k-1} c^{k-2} d^{k-1} \times \notin L$. Thus, $L \notin SL_k$.

Problem 12: In class we proved that $k\text{-SSC}(L) \implies L \in SL_k$. In other words, if a language satisfies k -Local Suffix Substitution Closure, then it is k -strictly local.

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

Answer 12: The property says that if $u_1 x v_1 \in L_k$ and $u_2 x v_2 \in L_k$, then $u_1 x v_2 \in L_k$, then $L \in SL_k$. $L_1 \cap L_2$ is the intersection of L_1 and L_2 , ie $L_1 \cap L_2 = \{l \mid l \in L_1 \text{ and } l \in L_2\}$. Let's suppose that $u_1 x v_1 \in L_1$ and $u_2 x v_2 \in L_1$ and also $u_1 x v_2 \in L_1$. Then, let's suppose that $u_3 x' v_3 \in L_2$ and $u_4 x' v_4 \in L_2$ and also $u_3 x' v_4 \in L_2$. $L_1 \cap L_2$ would be what is in common between L_1 and L_2 , ie a subset of $u_1, u_2, u_3, u_4, v_1, v_2, v_3, v_4, x$ and x' . Because the original 'sets' are in SL_k , the intersection or a subset would have to be as well.