LIGN 165
Problem Set 5

There is no programming required for this assignment. You can either write up your answers by hand and take an image/scan of the document, or you can use a mathematical typesetting tool like LaTeX.

You may collaborate with up to three other students on this problem set. You must submit your work individually. If you do not submit a copy of the problem set under your own name, you will not get credit. When you submit your work, you must indicate who you worked with, and what each of your individual contributions were.

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

$$Pr(W^{(1)}, ..., W^{(k)}) = \prod_{i}^{k} Pr(W^{(i)}|W^{(i-1)} = w^{(i-1)}) \tag{1}$$

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.

Problem 2. Let us 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 $Pr(W^{(i)}|W^{(i-1)} = start)$, where $start$ is the start symbol which begins any sentence. These conditional distributions are associated with the parameter vectors $\vec{\theta}_a$, $\vec{\theta}_b$, and $\vec{\theta}_{start}$, 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 $W^{(1)}, ..., W^{(k)}$. We will use the notation $n_{x \to y}$ to denote the number of times that the symbol y occurs immediately following the symbol x in the sentence. For example, $n_{a \to a}$ 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:

$$Pr(W^{(1)}, ..., W^{(k)}|\vec{\theta}_a, \vec{\theta}_b, \vec{\theta}_{start}) \tag{2}$$

The expression should make use of the $n_{x \to y}$ 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.)

Problem 3. Let us set the parameter vectors in our bigram model as follows:

$$\vec{\theta}_a = (0.7, 0.2, 0.1)$$
$$\vec{\theta}_b = (0.2, 0.7, 0.1)$$
$$\vec{\theta}_{start} = (0.5, 0.5, 0)$$

For example, given the current symbol a, there is probability 0.7 of transitioning to the symbol a, and probability 0.2 of transitioning to the symbol b. The third term in each vector is the probability of sentence ending after that symbol. Thus, given the current symbols a or b, there is probability 0.1 of the sentence ending.

Using your answer to the previous problem and these parameter values, calculate the probability of the string *aabb*.

Problem 4. In the previous problem we assumed that we knew the exact values of the parameter vectors $\vec{\theta}_a$, $\vec{\theta}_b$, and $\vec{\theta}_{start}$. In the current problem, we will assume that there are actually two possible sets of parameter vectors, $\vec{\boldsymbol{\theta}_1}$ and $\vec{\boldsymbol{\theta}_2}$. We do not know ahead of time which is the correct set of parameters.

The first set of parameters $\vec{\boldsymbol{\theta}_1}$ is defined by:

$$\vec{\theta}_a = (0.7, 0.2, 0.1)$$
$$\vec{\theta}_b = (0.2, 0.7, 0.1)$$
$$\vec{\theta}_{start} = (0.5, 0.5, 0)$$

The second set of parameters $\vec{\boldsymbol{\theta}_2}$ is defined by:

$$\vec{\theta}_a = (0.2, 0.7, 0.1)$$
$$\vec{\theta}_b = (0.7, 0.2, 0.1)$$
$$\vec{\theta}_{start} = (0.5, 0.5, 0)$$

We will assume that both sets of parameters have equal prior probability: $P(\vec{\boldsymbol{\theta}_1}) = P(\vec{\boldsymbol{\theta}_2}) = 0.5$.

Compute the marginal probability of the string *aabb* given these possible sets of parameters.

Problem 5. In the current problem, we will try to address a learning problem: determining which of the parameters $\vec{\boldsymbol{\theta}_1}$ or $\vec{\boldsymbol{\theta}_2}$ is the correct one. Using the marginal probability that you computed in Problem 4, compute the posterior probability $P(\vec{\boldsymbol{\theta}_1}|aabb)$.

The quantity that you computed should be greater than $P(\vec{\boldsymbol{\theta}_2}|aabb)$. Why is this true?

Problem 6. Find a string c (consisting of a's and b's) such that $P(\vec{\boldsymbol{\theta}_1}|c)$ is greater than the value $P(\vec{\boldsymbol{\theta}_1}|aabb)$ that you found in the previous problem. How did you construct this string?

Problem 7. Find a string c such that $P(\vec{\boldsymbol{\theta}_1}|c) < P(\vec{\boldsymbol{\theta}_2}|c)$. How did you construct this string?