LING 227/627 — Language and Computation I
Fall 2017

Problem Set 3: Probability and Language Models
Due 10/27/17 at midnight

NOTE: THIS DOCUMENT ONLY INCLUDES THE FIRST PART OF THE PROBLEM SET. MAKE SURE TO CHECK BACK SOON TO GET THE REST.

Part 1: Probability Exercises

The first part of the homework consists of a number of non-programming problems that will give you practice manipulating probabilities in the abstract and in the context of n-gram models.

Problem A: Manipulating Probabilities

Let $\mathcal{E} \neq \emptyset$ denote the event space (it's just a set, also known as the sample space), and p be a function that assigns a real number in $[0, 1]$ to any subset of $\mathcal{E}$. This number is called the probability of the subset.

You are told that p satisfies the following two axioms: $p(\mathcal{E}) = 1$ and $p(X \cup Y) = p(X) + p(Y)$ provided that $X \cap Y = \emptyset$.

As a matter of notation, remember that conditional probability is defined as

$$p(X|Z) \stackrel{\text{def}}{=} \frac{p(X \cap Z)}{p(Z)}$$

For example, if singing in the rain is one of your favorite rainy-day activities, your ratio

$$p(\text{singing}|\text{rainy}) = \frac{p(\text{singing AND rainy})}{p(\text{rainy})}$$

is high. Here the predicate "singing" picks out the set of singing events in $\mathcal{E}$, "rainy" picks out the set of rainy events, and the conjoined predicate "singing AND rainy" picks out the intersection of these two sets—that is, all events that are both singing AND rainy.

Now, answer the following questions:

(a) Prove from the axioms that if $Y \subseteq Z$, then $p(Y) \leq p(Z)$. You may use any and all set manipulations you like. Remember that $p(A) = 0$ does not imply that $A = \emptyset$ (why not?), and similarly, that $p(B) = p(C)$ does not imply that $B = C$ (even if $B \subseteq C$).

(b) Use the above fact to prove that conditional probabilities $p(X|Z)$, just like ordinary probabilities, always fall in the range [0, 1].

(c) Prove from the axioms that $p(\emptyset) = 0$.

(d) Let $\bar{X}$ denote $\mathcal{E} - X$. Prove from the axioms that $p(X) = 1 - p(\bar{X})$. For example, $p(\text{singing}) = 1 - p(\text{NOT singing})$.

(e) Prove from the axioms that $p(\text{singing AND rainy}|\text{rainy}) = p(\text{singing}|\text{rainy})$.

(f) Prove from the axioms that $p(X|Y) = 1 - p(\bar{X}|Y)$. For example, $p(\text{singing}|\text{rainy}) = 1 - p(\text{NOT singing}|\text{rainy})$. This is a generalization of (d).

(g) Simplify: $(p(X|Y) \cdot p(Y) + p(X|\bar{Y}) \cdot p(\bar{Y})) \cdot p(\bar{Z}|X)/p(\bar{Z})$

(h) Suppose you know that $p(X|Y) = 0$. Prove that $p(X|Y, Z) = 0$.

Problem B: Bayes Theorem

Beavers can make three cries, which they use to communicate. *bwa* and *bwee* usually mean something like "come" and "go" respectively, and are used during dam maintenance. *kiki* means "watch out!" The following conditional probability table shows the probability of the various cries in different situations.

p(cry | situation)	Predator!	Timber!	I need help!
bwa	0	0.1	0.8
bee	0	0.6	0.1
kiki	1.0	0.3	0.1

Part (a): Notice that each column of the table sums to 1. Write an equation stating this, in the form:

$$\sum_{\textit{variable}} p(\cdots) = 1$$

Part (b): A certain colony of beavers has already cut down all the trees around their dam. As there are no more to chew, $p(\textit{Timber!}) = 0$. Getting rid of the trees has also reduced $p(\textit{Predator!})$ to 0.2. These facts are shown in the following joint probability table. Fill in the rest of the table, using the previous table and the laws of probability. (Note that the meaning of each table is given in its top left cell.)

p(cry, situation)	Predator!	Timber!	I need help!	TOTAL
bwa				
bee				
kiki				
TOTAL	0.2	0		

Part (c):

A beaver in this colony cries *kiki*. Given this cry, other beavers try to figure out the probability that there is a predator.

1. This probability is written as: $p(\_\_\_\_\_)$
2. It can be rewritten without the | symbol as: _____
3. Using the above tables, its value is: _____
4. Alternatively, Bayes' Theorem allows you to express this probability as:

$$\frac{p(\_\_\_\_\_) \cdot p(\_\_\_\_\_)}{p(\_\_\_\_\_) \cdot p(\_\_\_\_\_) + p(\_\_\_\_\_) \cdot p(\_\_\_\_\_) + p(\_\_\_\_\_) \cdot p(\_\_\_\_\_)}$$

5. Using the above tables, the value of this is:

$$\frac{\_\_\_\_\_ \cdot \_\_\_\_\_}{\_\_\_\_\_ \cdot \_\_\_\_\_ + \_\_\_\_\_ \cdot \_\_\_\_\_ + \_\_\_\_\_ \cdot \_\_\_\_\_}$$

   This should give the same result as in part iii., and it should be clear that they are really the same computation—by constructing table (b) and doing part iii., you were implicitly using Bayes' Theorem.

Problem C: Reverse N-gram models

N-gram models are generally defined as proceeding in a left-to-right direction, with a word depending on the two words that come *before* it.

$$P(w) = \prod_{i=1}^{n+1} p(w_i | w_{i-2}, w_{i-1})$$

By convention $w_{-1} = w_0 =$ BOS (beginning of string) and $w_{n+1} =$ EOS (end of string). So, under this way of representing the generative process, a sentence is generated beginning with two BOS symbols and continuing until an EOS is reached. However, n-gram models can also be defined a right-to-left direction, as follows:

$$P_{rev}(w) = \prod_{i=1}^{n} p(w_i | w_{i+1}, w_{i+2})$$

where $w_{n+1} = w_{n+2} =$ EOS and $w_0 =$ BOS. In this case the generative process starts with two EOS markers, and proceeds by generating the previous word, until a BOS marker is generated.

Show that $P(w) = P_{rev}(w)$ for all w, under the assumption that both models use MLE parameters estimated from the same data. (Hint: try writing out the probability of "i love new york" under both models. To argue that the resulting probabilities are equal, you will have to observe that certain counts are equal.)

Problem D: Smoothing and Probability Distributions

In the *absolute discounting* model of smoothing, all non-zero MLE probabilities are discounted by a constant amount δ where $0 < \delta, 1$.

Absolute discounting: If $C(w_n|w_1 \dots w_{n-1}) = r$

$$P_{abs}(w_n|w_1 \dots w_{n-1}) = \begin{cases} \frac{(r-\delta)}{N} & r > 0 \\ \frac{(V-N_0)\delta}{N_0 \cdot N} & otherwise \end{cases}$$

Here $C(w_n|w_1 \dots w_{n-1})$ is the number of times $w_1 \dots w_n$ has been seen, P_{abs} is the absolute discounting estimate, V is the size of the vocabulary, N is the total number of times the context $w_1 \dots w_{n-1}$ has been seen, and N_0 is the number of word types that were unseen in this context.

Under the *linear discounting* model, the estimated count of seen words is discounted by a certain fraction, defined by a constant α where $0 < \alpha < 1$.

Linear discounting: If $C(w_n|w_1 \dots w_{n-1}) = r$

$$P_{lin}(w_n|w_1 \dots w_{n-1}) = \begin{cases} \frac{(1-\alpha)r}{N} & r > 0 \\ \frac{\alpha}{N_0} & otherwise \end{cases}$$

Show that both absolute discounting and linear discounting yield probability distributions for any context $w_1 \dots w_{n-1}$. In other words, show that either definition defines a probability measure over possible next words w_n that sums to 1.

That's it!

You submission should include:

- A pdf (preferred) or text file containing your answers to the problems in part 1 of the problem set. You can typeset the mathematical formulas using LaTeX, or using a formula editor in Microsoft Word.