

600.465 — Natural Language Processing

Assignment 2: Probability Exercises

Prof. J. Eisner — Fall 2011
Due date: Monday 19 September, 2 pm

No programming is required for this assignment.

How to hand in your work: Put all notes, documentation, and answers to questions in a `README` file. The file should be editable so that we can insert comments and give it back to you. For this reason, we strongly prefer a plain ASCII file `README`, or a \LaTeX file `README.tex` (in which case please also submit `README.pdf`). If you must use a word processor, please save as `README.rtf` in the portable, non-proprietary RTF format.

Notation: When you are writing the `README` file, you will need some way of typing mathematical symbols. If your file is just plain ASCII text, please use one of the following three notations and stick to it in your assignment. (If you need some additional notation not described here, just describe it clearly and use it.) Use parentheses as needed to disambiguate division and other operators.

	Text	Picts	\LaTeX
$p(x y)$	<code>p(x y)</code>	<code>p(x y)</code>	<code>p(x \mid y)</code>
$\neg x$	<code>NOT x</code>	<code>~x</code>	<code>\neg x</code>
\bar{x} (set complement)	<code>COMPL(x)</code>	<code>\x</code>	<code>\bar{x}</code>
$x \subseteq y$	<code>x SUBSET y</code>	<code>x {=} y</code>	<code>x \subseteq y</code>
$x \supseteq y$	<code>x SUPERSET y</code>	<code>x }= y</code>	<code>x \supseteq y</code>
$x \cup y$	<code>x UNION y</code>	<code>x U y</code>	<code>x \cup y</code>
$x \cap y$	<code>x INTERSECT y</code>	<code>x ^ y</code>	<code>x \cap y</code>
$x \geq y$	<code>x GREATEREQ y</code>	<code>x >= y</code>	<code>x \geq y</code>
$x \leq y$	<code>x LESSEQ y</code>	<code>x <= y</code>	<code>x \leq y</code>
\emptyset (empty set)	<code>NULL</code>	<code>0</code>	<code>\emptyset</code>
\mathcal{E} (event space)	<code>E</code>	<code>E</code>	<code>E</code>

1. These short problems will help you get the hang of manipulating probabilities. Let $\mathcal{E} \neq \emptyset$ denote the event space (it's just a set, also known as the outcome space or 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: $p(\mathcal{E}) = 1$. $p(X \cup Y) = p(X) + p(Y)$ provided that $X \cap Y = \emptyset$.¹

As a matter of notation, remember that the **conditional probability** $p(X | Z) \stackrel{\text{def}}{=} \frac{p(X \cap Z)}{p(Z)}$. For example, singing in the rain is one of my favorite rainy-day activities: so my 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.

- (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 **1d**.
- (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) Under what conditions is it true that $p(\text{singing OR rainy}) = p(\text{singing}) + p(\text{rainy})$?
- (i) Under what conditions is it true that $p(\text{singing AND rainy}) = p(\text{singing}) \cdot p(\text{rainy})$?
- (j) Suppose you know that $p(X | Y) = 0$. Prove that $p(X | Y, Z) = 0$.²
- (k) Suppose you know that $p(W | Y) = 1$. Prove that $p(W | Y, Z) = 1$.

2. All cars are either red or blue. The witness claimed the car that hit the pedestrian was blue. Witnesses are believed to be about 80% reliable in reporting car color (regardless of the actual car color). But only 10% of all cars are blue.

¹In fact, probability functions p are also required to satisfy a generalization of this second axiom: if X_1, X_2, X_3, \dots is an infinite sequence of disjoint sets, then $p(\bigcup_{i=1}^{\infty} X_i) = \sum_{i=1}^{\infty} p(X_i)$. But you don't need this for this assignment.

²More precisely, $p(X | Y, Z)$ could be either 0 or undefined, namely $0/0$. (There do exist advanced ways to redefine conditional probability to avoid this $0/0$ problem. Even then, though, one may want a probability measure p to leave some probabilities or conditional probabilities undefined. This turns out to be important for reasons beyond the scope of this course: e.g. http://en.wikipedia.org/wiki/Vitali_set.)

- (a) Write an equation relating the following quantities and perhaps other quantities:

$$\begin{aligned}
 & p(\text{true} = \text{blue}) \\
 & p(\text{true} = \text{blue} \mid \text{claimed} = \text{blue}) \\
 & p(\text{claimed} = \text{blue} \mid \text{true} = \text{blue})
 \end{aligned}$$

Reminder: Here, *claimed* and *true* are *random variables*, which means that they are functions over some outcome space. For example, the probability that *claimed* = blue really means the probability of getting an outcome x such that $\text{claimed}(x) = \text{blue}$. We are implicitly assuming that the space of outcomes x is something like the set of witnessed car accidents.

- (b) Match the three probabilities above with the following terms: *prior probability*, *likelihood of the evidence*, *posterior probability*.
- (c) Give the values of all three probabilities. (Hint: Use Bayes' Theorem.) Which probability should the judge care about?
- (d) Let's suppose the numbers 80% and 10% are specific to Baltimore. So in the previous problem, you were implicitly using the following more general version of Bayes' Theorem:

$$p(A \mid B, Y) = \frac{p(B \mid A, Y) \cdot p(A \mid Y)}{p(B \mid Y)}$$

where Y is *city* = Baltimore. Just as **1f** generalized **1d**, by adding a “background” condition Y , this version generalizes Bayes' Theorem. Carefully prove it.

- (e) Now prove the more detailed version

$$p(A \mid B, Y) = \frac{p(B \mid A, Y) \cdot p(A \mid Y)}{p(B \mid A, Y) \cdot p(A \mid Y) + p(B \mid \bar{A}, Y) \cdot p(\bar{A} \mid Y)}$$

which gives a practical way of finding the denominator in the question **2d**.

- (f) Write out the equation given in question **2e** with A , B , and Y replaced by specific propositions from the red-and-blue car problem. For example, Y is “*city* = Baltimore” (or just “Baltimore” for short). Now replace the probabilities with actual numbers from the problem, such as 0.8.

Yeah, it's a mickeymouse problem, but I promise that writing out a real case of this important formula won't kill you, and may even be good for you (like, on an exam).

3. 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.

$p(\text{cry} \mid \text{situation})$	Predator!	Timber!	I need help!
bwa	0	0.1	0.8
bwee	0	0.6	0.1
kiki	1.0	0.3	0.1

(a) Notice that each column of the above table sums to 1. Write an equation stating this, in the form $\sum_{\text{variable}} p(\dots) = 1$.

(b) A certain colony of beavers has already cut down all the trees around their dam. As there are no more to chew, $p(\text{timber}) = 0$. Getting rid of the trees has also reduced $p(\text{predator})$ 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.)

$p(\text{cry}, \text{situation})$	Predator!	Timber!	I need help!	TOTAL
bwa				
bwee				
kiki				
TOTAL	0.2	0		

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

- i. This probability is written as: $p(\text{_____})$
- ii. It can be rewritten without the $|$ symbol as: _____
- iii. Using the above tables, its value is: _____
- iv. Alternatively, Bayes' Theorem allows you to express this probability as:

$$\frac{p(\text{_____}) \cdot p(\text{_____})}{p(\text{_____}) \cdot p(\text{_____}) + p(\text{_____}) \cdot p(\text{_____}) + p(\text{_____}) \cdot p(\text{_____})}$$

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

$$\frac{\text{_____} \cdot \text{_____}}{\text{_____} \cdot \text{_____} + \text{_____} \cdot \text{_____} + \text{_____} \cdot \text{_____}}$$

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. (I told you it was a trivial theorem!)

4. (a) $p(\neg\text{shoe} \mid \neg\text{nail}) = 1$ *For want of a nail the shoe was lost,*
 (b) $p(\neg\text{horse} \mid \neg\text{shoe}) = 1$ *For want of a shoe the horse was lost,*
 (c) $p(\neg\text{race} \mid \neg\text{horse}) = 1$ *For want of a horse the race was lost,*
 (d) $p(\neg\text{fortune} \mid \neg\text{race}) = 1$ *For want of a race the fortune was lost,*
 (e) $p(\neg\text{fortune} \mid \neg\text{nail}) = 1$ *And all for the want of a horseshoe nail.*

Show carefully that (e) follows from (a)–(d). *Hint:* Consider

$$p(\neg\text{fortune}, \neg\text{race}, \neg\text{horse}, \neg\text{shoe} \mid \neg\text{nail}),$$

as well as the “chain rule” and problems 1a, 1b, and 1k.

Note: The \neg symbol denotes the boolean operator NOT.

Note: This problem is supposed to convince you that logic is just a special case of probability theory. An excellent and wide-ranging book developing this theme, by the influential statistician E. T. Jaynes, is *Probability Theory: The Logic of Science*. See <http://bayes.wustl.edu/> for more readings.

Note: Be glad I didn’t ask you to prove the correct operation of Figure 1!

5. A **language model** is a probability function p that assigns probabilities to word sequences such as $\vec{w} = (\text{i}, \text{love}, \text{new}, \text{york})$. Think of $p(\vec{w})$ as the probability that if you turned on a radio at an arbitrary moment, its next four words would be “i love new york”—perhaps in the middle of a longer sentence such as “the latest bumper sticker says, i love new york more than ever.” We often want to consider $p(\vec{w})$ to decide whether we like \vec{w} better than an alternative sequence.³

Suppose $\vec{w} = w_1 w_2 \cdots w_n$ (a sequence of n words). A **trigram language model** defines

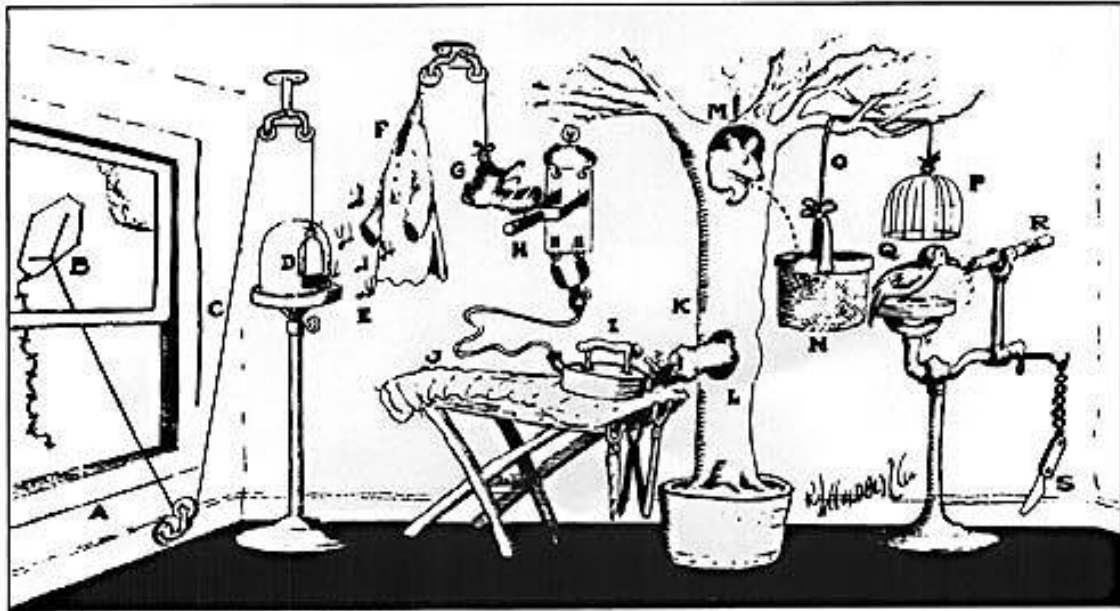
$$p(\vec{w}) \stackrel{\text{def}}{=} p(w_1) \cdot p(w_2 \mid w_1) \cdot p(w_3 \mid w_1, w_2) \cdot p(w_4 \mid w_2, w_3) \cdots p(w_n \mid w_{n-2}, w_{n-1})$$

on the assumption that the sequence was generated in the order $w_1, w_2, w_3 \dots$ (“from left to right”) with each word chosen in a way dependent on the previous two words. (But the first word w_1 is not dependent on anything, since we turned on the radio at a arbitrary moment.)

- (a) Expand the above definition of $p(\vec{w})$ using naive estimates of the parameters, such as

$$p(w_4 \mid w_2, w_3) \stackrel{\text{def}}{=} \frac{c(w_2 w_3 w_4)}{c(w_2 w_3)}$$

³Formally, each element $W \in \mathcal{E}$ of the underlying event space is a possible value of the *infinite* sequence of words that will come out of the radio after you turn it on. $p(\vec{w})$ is really an abbreviation for $p(\text{prefix}(W, |\vec{w}|) = \vec{w})$, where $|\vec{w}|$ denotes the length of the sequence \vec{w} . Thus, $p(\text{i}, \text{love}, \text{new}, \text{york})$ is the total probability of all infinite word sequences W that begin “i love new york”



Pencil Sharpener RUBE GOLDBERG (tm) RGI 038

Figure 1: RUBE GOLDBERG GETS HIS THINK-TANK WORKING AND EVOLVES THE SIMPLIFIED PENCIL-SHARPENER. Open window (A) and fly kite (B). String (C) lifts small door (D) allowing moths (E) to escape and eat red flannel shirt (F). As weight of shirt becomes less, shoe (G) steps on switch (H) which heats electric iron (I) and burns hole in pants (J). Smoke (K) enters hole in tree (L), smoking out opossum (M) which jumps into basket (N), pulling rope (O) and lifting cage (P), allowing woodpecker (Q) to chew wood from pencil (R), exposing lead. Emergency knife (S) is always handy in case opossum or the woodpecker gets sick and can't work.

where $c(w_2w_3w_4)$ denotes the count of times the trigram $w_2w_3w_4$ was observed in a training corpus.

Remark: Naive parameter estimates of this sort are called “maximum-likelihood estimates” (MLE). They have the advantage that they maximize the probability (equivalently, minimize the perplexity) of the training data. But they will generally perform badly on test data, unless the training data were so abundant as to include all possible trigrams many times. This is why we must smooth these estimates in practice.

- (b) One could also define a kind of reversed trigram language model $p_{reversed}$ that instead assumed the words were generated in reverse order (“from right to left”):

$$p_{reversed}(\vec{w}) \stackrel{\text{def}}{=} p(w_n) \cdot p(w_{n-1} | w_n) \cdot p(w_{n-2} | w_{n-1}, w_n) \cdot p(w_{n-3} | w_{n-2}, w_{n-1}) \cdots p(w_2 | w_3, w_4) \cdot p(w_1 | w_2, w_3)$$

By manipulating the notation, show that the two models are identical (i.e., $p(\vec{w}) = p_{reversed}(\vec{w})$ for any \vec{w}) provided that both models use MLE parameters estimated from the same training data (see problem 5a).

- (c) In the data you will use in the *next* assignment, sentences are delimited by `<s>` at the start and `</s>` at the end. For example, the following data set consists of a sequence of 3 sentences:

`<s> do you think so </s> <s> yes </s> <s> at least i thought so </s>`

Given English training data, the probability of

`<s> do you think the </s>`

should be extremely low under any good language model. Why? In the case of the trigram model, which parameter or parameters are responsible for making this probability low?

- (d) You turn on the radio as it is broadcasting an interview. Assuming a trigram model, match up expressions (A), (B), (C) with descriptions (1), (2), (3):

The expression

(A) $p(\text{Do}) \cdot p(\text{you} \mid \text{Do}) \cdot p(\text{think} \mid \text{Do}, \text{you})$

(B) $p(\text{Do} \mid \text{<s>}) \cdot p(\text{you} \mid \text{<s>, Do}) \cdot p(\text{think} \mid \text{Do}, \text{you}) \cdot p(\text{</s>} \mid \text{you}, \text{think})$

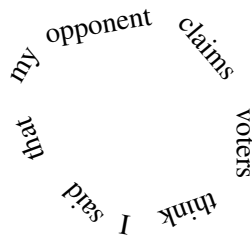
(C) $p(\text{Do} \mid \text{<s>}) \cdot p(\text{you} \mid \text{<s>, Do}) \cdot p(\text{think} \mid \text{Do}, \text{you})$

represents the probability that

- (1) the first complete sentence you hear is `Do you think` (as in, "D'ya think?")
- (2) the first 3 words you hear are `Do you think`
- (3) the first complete sentence you hear starts with `Do you think`

Explain your answers briefly. Which quantity is $p(\vec{w})$? *Remark:* The distinctions matter because "Do" is more probable at the start of an English sentence than in the middle, and because (3) describes a larger event set than (1) does.

6. Some politicians seem to speak in circles. Instead of uttering strings with a start and an end, they blow round linguistic objects out of their mouths, like smoke rings:



I wondered, how would we build a language model to describe the probability of such objects? Well, problems like this show up in computer vision, since images don't have a natural start and end either. Consider the similar problem of modeling the colors of these 4 pixels, using bigrams of adjacent pixel colors:

A	B
D	C

A well-known computer vision paper by J. Besag (1974) proposed approximating $p(A, B, C, D)$ by a product $p(A | B) \cdot p(B | C) \cdot p(C | D) \cdot p(D | A)$.

So here's the question. Can Besag's approximation be justified by using the chain rule plus backoff? If so, show how. If not, fix it as best you can. Discuss.

- Under an bigram language model, knowing an early part of the sentence ("*Horses like ...*") doesn't tell you too much about the end of the sentence. But that overlooks a useful property of real language data. A sentence that starts out "*Horses like ...*" is probably about horses, so it will probably continue to have a lot of horse-related words throughout, even at the end.

Suppose your corpus seems to cover k different topics—horse-racing, Star Trek, knitting, Wall Street, etc. For the most part, each sentence seems to stick to a single topic. **You want to build a better bigram model that also captures the fact the choice of topic persists throughout each sentence.**

Write a formula for $p(w_1 w_2 w_3 w_4)$ under this better model. Your answer should not use the t_{xy} notation, because that refers to the old model. (*Hints:* Latent variable, chain rule, backoff. Make w_4 depend not only on w_3 but also on the topic a .)