Problem Set 1: Review of probability

Jacob Eisenstein

August 17, 2015

This problem set is graded out of fifteen points, but like all problem sets, will count towards 8% of your final grade. It is due on T-Square at the beginning of class on August 26. It will be accepted up to 72 hours late, with a penalty of 20% per 24 hour period.

Using LaTeX to typeset your answers is encouraged, but not required. All work must be your own.

1 Zombie Bob (4 pts total)

Alice knocks on Bob's door, and hears the reply *graagh!*. There are two possibilities:

1. It is Bob. Alice estimates that Bob's likelihood of saying *graagh* is 10^{-5}.
2. It is Zombie Bob. Having seen many zombie movies, Alice estimates that Zombie Bob's likelihood of saying *graagh* is 0.5.

1.1 Bayes rule (2 pts)

Alice is a cold-eyed realist, and her **prior** probability that Bob could become a zombie was one in a million. After hearing Bob's reply *graagh!*, what is Alice's posterior belief about whether Bob is a zombie? Give an answer that is correct to at least three decimal points.

For the following parts of problem 1, you can round these probabilities to the first two decimal places.

1.2 Expected utility (1 pt)

Having heard Bob say *graagh*, Alice can stay or run away. She has the following *utility* function:

- $u(\text{stay, Bob}) = 0$. Alice gets to talk to Bob.
- $u(\text{stay, Zombie Bob}) = -20$. Zombie Bob eats Alice's brain.

- $u(\text{run, Bob}) = -1$. Alice runs, and is embarrassed.
- $u(\text{run, Zombie Bob}) = 3$. Alice runs, and sells her story to AMC.

What is Alice's expected utility under each strategy? (strategy 1 = stay, strategy 2 = run.)

1.3 The chain rule and marginal probabilities (1 pt)

Now consider that Alice has been studying karate, and has a 50% chance of surviving an encounter with Zombie Bob. If she opens the door, what is the chance that she will live, conditioned on Bob saying *graagh*? Assume there is a 100% chance that Alice lives if Bob is not a zombie.

2 Necromantic Scrolls (4 pts total)

The Necromantic Scroll Aficionados (NSA) would like to know the author of a recently discovered ancient scroll. They have narrowed down the possibilities to two candidate wizards: Anna and Barry. From painstaking corpus analysis of texts known to be written by each of these wizards, they have collected frequency statistics for the words *abracadabra* and *gesundheit*, shown in Table 1

	abracadabra	*gesundheit*
Anna	5 per 1000 words	6 per 1000 words
Barry	10 per 1000 words	1 per 1000 words

Table 1: Word frequencies for wizards Anna and Barry. This means that, for example, for every 1000 words that Anna writes, 5 of them are *abracadabra*.

2.1 Bayes rule (1 pt)

Catherine has a prior belief that Anna is 60% likely to be the author of the scroll. She peeks at a random word of the scroll, and sees that it is the word *abracadabra*. Use Bayes' rule to compute Catherine's posterior belief that Anna is the author of the scroll, $P(\text{Anna} \mid x = \textit{abracadabra})$.

2.2 Breakeven point (1 pt)

What prior probability would Catherine have to assign to Anna being the author, so that her posterior probabilities were equal? Assume the likelihoods remain as shown in Table 1; we are looking for the prior $P(Y = \text{Anna})$ such that $P(Y = \text{Anna} \mid x = \textit{abracadabra}) = P(Y = \text{Barry} \mid x = \textit{abracadabra})$.

2.3 Multiple words (2 pts)

Dante has no prior belief about the authorship of the scrolls (except that the author must be either Anna or Barry). He reads the entire first page, and finds that it contains 100 words, with two counts of the word *abracadabra* and one count of the word *gesundheit*.

1. What is his posterior belief about the probability that Anna is author of the scroll? (1 pt)
2. Does Dante need to consider the 97 words that were not *abracadabra* or *gesundheit*? Why or why not? (1 pt) Assume that he cannot obtain per-author frequency statistics for any additional words — that is, he cannot expand Table 1.

3 Sentence lengths (5 pts total)

Consider the following *model* for writing sentences:

- While not done:
  - With probability $(1 - \lambda)$, stop and break out of the loop.
  - With probability λ, write a word and continue looping.

This model produces sentences of stochastically varying lengths. For example, the probability of a sentence of length 0 is $P(\ell = 0) = 1 - \lambda$.

3.1 Maximum likelihood estimation (2 pts)

Given a corpus of sentences, with lengths $\{\ell_1, \ell_2, \ldots, \ell_N\}$, compute the maximum-likelihood estimate of λ. This means that you want to find the value of λ that maximizes the likelihood function, $P(\ell_{1:N}|\lambda) = \prod_n P(\ell_n|\lambda)$.

It is equivalent (and easier) to maximize the *log* likelihood, $\sum_n \log P(\ell_n|\lambda)$, so you should do this. Specifically, take the derivative with respect to λ, and solve. Write your answer in terms of the average sentence length $\overline{\ell} = \frac{1}{N}\sum_i^N \ell_i$.

3.2 Expectations (3 pts)

1. What is the expected sentence length, given a parameter λ? Since sentence lengths are discrete, the expectation is a weighted sum,

$$E[\ell] = \sum_{\ell=0}^{\infty} \ell P(\ell) \tag{1}$$

   Here is a cheatsheet with some identities: `https://www.tug.org/texshowcase/cheat.pdf`. I believe you will find one of them (on the first page) to be useful. (2 pts)

2. What is the modal (most probable) sentence length, according to this model? (1 pt)

3. **Extra credit**: When the modal value is different from the expected value, the probability distribution is said to be *skewed*. Propose a probability mass function $p(\ell)$, for which the expected sentence length would be no more than one word more or less than the modal sentence length. Sentence lengths must be non-negative integers, so your probability function should have *support* over only these numbers; it must assign $P(x) = 0$ for all $x < 0$ or $x \notin \mathbb{Z}$. The correct answer will be a well-known probability distribution. (1 pt)

4 Part-of-speech tagging accuracy (2 pts total)

Part-of-speech tagging is a key component of the NLP pipeline, in which each word token is assigned an element from a set of syntactic tags, such as noun, verb, adjective, etc. Felicia has built a part of speech tagger with 10% per-word error rate — it gives the wrong tag 10% of the time.

1. Suppose a sentence contains n words, and that the chance of making an error on each word is independent and identically distributed (IID). What is the chance of tagging the entire sentence correctly? Give the answer for $n = 5$, rounding to two decimal places. (1 pt)

2. Felicia's tagger makes errors that are IID. Gregory has also built a tagger that has a 10% per-word error rate, but his tagger makes all of its errors on verbs. Assume that Felicia and Gregory apply their taggers to the same corpus: whose tagger will get more sentences completely correct, and why? (1 pt; hint: you will have to apply a small amount of linguistic intuition to answer this question.)