CSE 250B Quiz 6

Tuesday February 18, 2014

Instructions. Do this quiz in partnership with exactly one other student. Write both your names at the top of this page. **Circle one name that we will call out when we return the quiz. Choose a first name or last name that is likely to be unique in the class.**

Discuss the answer to the question with each other, and then write your joint answer below the question. Use the back of the page if necessary. It is fine if you overhear what other students say, because you still need to decide if they are right or wrong. You have seven minutes. The maximum score is three points.

Question. Remember that the multinomial distribution is

$$p(x;\theta) = \Big(\frac{(\sum_{j=1}^m x_j)!}{\prod_{j=1}^m x_j!}\Big)\Big(\prod_{j=1}^m \theta_j^{x_j}\Big).$$

where x is a vector of word counts and θ is a parameter vector. Consider two different words i and j that are rare, but have the same probability: $\theta_i = \theta_j$. Assume that the words are unrelated in meaning, such as “stalking” and “echoing.” Now, consider two documents that are identical, except that document E (“each”) contains each word once, while document S (“same”) contains word i only, twice.

Intuitively, which document should have higher probability? Which has higher probability according to the multinomial distribution?

Answer. Intuitively, using one rare word twice is more likely than using two different and semantically unrelated rare words in the same document. Hence, document S should have higher probability.

Looking at words i and j, the probability of E has the factors

$$\frac{1}{1!\,1!}\theta_i^1\theta_j^1$$

while the probability of S has the factors

$$\frac{1}{2!\,0!}\theta_i^2\theta_j^0.$$

So document S has half the probability, the opposite of what intuition expects.

Additional notes. This example reveals a basic weakness of the multinomial distribution as a model for documents. In real documents, words tend to be bursty: if a word appears once, the same word is likely to appear again. Said differently, the second or later appearance of a word is less surprising than the first appearance, and should have higher probability. But with a multinomial distribution, every appearance of a word i has the same probability θ_i.

The question whether S or E should have higher probability is is a question about the human world, where the burstiness of words is an incontrovertible fact. Intuition here means intuition about documents, not about mathematical objects. A mathematical model should conform to the phenomenon being modeled, not the other way round.

This example also reveals a subtle property of the multinomial distribution. The entity assigned a probability is a vector of word counts, not a sequence of words. The multinomial coefficient counts how many different sequences give the same count vector. When two different words each occur once, either can be first, so there are twice as many sequences as if one of the words occurred twice. The count vector for document E has higher probability because it covers twice as many word sequences.