CSE 250B Quiz 6
Tuesday February 19, 2013

Instructions. Do this quiz in partnership with exactly one other student. Write both your names at the top of this page. 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.

Question. Remember that the multinomial distribution is

\[ p(x; \theta) = \left( \frac{\left( \sum_{j=1}^{m} x_j \right)!}{\prod_{j=1}^{m} x_j!} \right) \left( \prod_{j=1}^{m} \theta_j^{x_j} \right). \]

where \( x \) is a vector of word counts and \( \theta \) is a parameter vector.

Consider two different words \( i \) and \( j \) that are rare, but have the same probability: \( \theta_i = \theta_j \). For example, the two words might be “acrobat” and “aphorism.” Now, consider two documents that are identical, except that the first contains each word once, while the second contains word \( i \) only, twice.

Intuitively, which document should have higher probability? Is the multinomial distribution compatible with this intuition?

Answer. Intuitively, if a writer uses a rare word once, it is not surprising that s/he will use the same word again. This is much more likely than using two unrelated rare words in the same document. Hence, the second document should have higher probability.

Looking at words \( i \) and \( j \), in the multinomial distribution the document with both words has the factors

\[ \frac{1}{1! 1!} \theta_i^1 \theta_j^1 \]

while the other document has the factors

\[ \frac{1}{2! 0!} \theta_i^2 \theta_j^0. \]
So the second document has half the probability, the opposite of what intuition expects.

*Additional note.* 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 has the same probability.