Lecture 15 - 6: MLE for Multinomial Distribution

Tabaré Pérez

May 6, 2020

So now the question, you remember when we started talking to you about it, I tell you, let's make an assumption that our $\mathcal{W}$ has just two words in it. And the reason I did it was because when I did this assumption, actually, I am estimating one single θ_0 because θ_1 can be expressed in terms of this θ_0.

If I want to work with vocabulary of any length, then I would have many of these different θs . And in this case, you can do exactly the same story, but you need to use the Lagrange multipliers to find the corresponding θs.

If you do both of these computation, you will get an expression which is very similar to this one and I will just write here this expression:

$$\hat{\theta} = \frac{\text{count}(w)}{\sum_{w' \in \mathcal{W}} \text{count}(w')} \tag{1}$$

So θ for the word w would be the count of this word w divided by the count of all the words in the document.

So again, the expression itself will be exactly the same expression. It's just derivation that is a tiny bit more complex.

So at this point, what we've seen so far, I described to you how given some training data, we can take this training data and estimate these θs, the parameters of the multinomial, to find a multinomial which fit this data the best. And again, I remember you, that we have only a single document.

In reality, you can assume that even if you have a collection of documents, the whole story is exactly the same because whenever we're making our assumption that we are generating the words, these words are generated independently. So pretty much, you can do exactly the same formula by concatenating many of your documents into a single document and then repeating the same story.

So now we're done with the discussion of estimation for multinomial. And with that, we are ready to start talking about how can we use these multinomials to actually do prediction.