Lecture 15 - 3: Simple Multinomial Generative model

Tabaré Pérez

May 6, 2020

So whenever we are talking about multinomials, one commonly example used is talking about documents.

So we will talk about multinomials, and in your head, you can think about texts, documents and what our models, our multinomial models will do, they will generate documents.

Now, whenever we were talking about supervised learning, you remember we discussed how to translate, we thought about the documents, and we translated them into a vector, into bag of words. So there was some particular mechanism that, given a document, we generated a vector of fixed lengths.

Here, we are going to be thinking about it in a different way. We're going to think that our model is going to be generating documents. And some documents, which are good for this model, will have very high probability, and other documents may have low probability.

So let me first show you some notation so that we can ground this discussion in the type of language that we will need to use to discuss our questions of estimation and predictions. And before I go, I want to say that, of course, you're thinking:

- How probabilistic model can be generating documents?
- Is it going to write poems and essays?

So when I'm talking about generate, it's not like generate an essay.

We'll have a very simplistic definition of generate. Specifically, what we would assume that this model have fixed vocabulary, first of all.

We, as humans, also have fixed vocabulary. This is capital $\mathcal{W}$ and then these models would generate one word at a time. So you have this whole bag of possible words, you select one word, put it there, and then again, you go

to the same bag, and you select another word, and all the words that we are selecting are independent of each other.

So it's a very simplistic thing, because clearly, we're not going to get beautiful sentences.

But this is our first model that we will use. So again, the words all will come from the same vocabulary. We draw them one at a time and fully independently.

So what kind of parameters do we need to have to talk about this model? So the first thing in multinomials, we need to decide how likely it is to generate certain words.

Because depending on the model, certain words will be more likely or less likely. So one of the parameters here would be something which is the likelihood of generating the word w, given parameterization of the model:

$$\mathbb{P}(w|\theta) = \theta_w \tag{1}$$

So θ means parameters of the model. So in this case, just for ease of writing, I would write it as θ_w. So θ_w captures what is the likelihood of selecting, generating certain words given all the possibilities.

So what kind of constraints do we need to have on the θ_w to make sure that we have a valid probability distribution?

We need to make sure, because this is probabilities:

1. $\theta_w \geq 0$
2. $\sum_{w \in \mathcal{W}} \theta_w = 1$

So this is our **MULTINOMIAL**.

And if you need another example and the text sounds to you really weird, you can just see about dice:

So you're throwing dice, and your dice doesn't have all the equal sides. Some sides are more likely than others.

And that's exactly what θ_w would record: the likelihood of a particular word or number.