Lecture 15 - 5: Maximum Likelihood Estimate

Tabaré Pérez

May 6, 2020

So now we have our parameters. We have ways to compute the likelihood of the documents based on these parameters.

And the question is, how can we utilize our training data to find the best parameters.

So again, we are coming back to supervised scenario that you've seen in the case of classifiers.

We will provide our model with, let's say, positive or negative examples, and the model needs to find the best parameters that fit the data. Here we are going to go back to our favorite principles of maximum likelihood, and we will make the assumption that the best parameters are the parameters which give the highest likelihood to our data.

So you have the training data and you're trying to find parameters for which the probability of these training data will be the highest.

So in other words, what we are trying to do is to find these θs which maximize this expression:

$$\max_{\theta} \mathbb{P}(D|\theta) = \max_{\theta} \prod_{w \in \mathcal{W}} \theta_w^{\text{count}(w)} \tag{1}$$

And again, assume that this is your training document D, over all the θs.

So it turns out, and you will see it both in the lecture and in the exercises, that it's easier for us to this maximization to find the best θs and instead of working with this expression directly, we would actually maximize the log, and we can do it. So what we will need to do is to take the log of this expression:

$$\log \prod_{w \in \mathcal{W}} \theta_w^{\text{count}(w)} = \sum_{w \in \mathcal{W}} \text{count}(w) \cdot \log \theta_w \tag{2}$$

You remember that log of the products is actually sum of logs.

So now we are trying to find θs that are going to maximize the value of this expression. What I will do now, I will stick to the example that I had before where I really looked at the vocabulary of only two symbols.

We will first solve this for the case where the alphabet has just two symbols, and then discussed how we can solve it for general vocabulary. And I know that we all love cats and dogs, but for me it will be easier to write $\mathcal{W}$ with just two words, exactly like in the example above, but this time I'm going to be using just the words 0 and 1, because it's shorter to write it.

$$\mathcal{W} = \{0, 1\} \tag{3}$$

In this particular case, what's so special above this case?

Because in this case, we actually just need it to have a single θ.

So we can write that :

$$\theta_0 = \theta, \theta_1 = 1 - \theta \tag{4}$$

We need a θ for zero, the likelihood of generating the word zero. If we want to know the θ_1, we can just do 1 and subtract θ.

So I would just have one parameter, and then I'm going to rewrite this formula with the single parameter.

$$\sum_{w \in \mathcal{W}} \text{count}(w) \cdot \log(\theta_w) = \text{count}(0) \cdot \log(\theta) + \text{count}(1) \cdot \log(1 - \theta) \tag{5}$$

So my next step is actually to find derivative of this expression with respect to θ and what we will get in this case is:

$$\frac{\partial}{\partial \theta}[\text{count}(0) \cdot \log(\theta) + \text{count}(1) \cdot \log(1 - \theta)] = \frac{\text{count}(0)}{\theta} - \frac{\text{count}(1)}{(1 - \theta)} \tag{6}$$

So this is our new expression, and we need to make it equal to 0 to find our desired θs. So in this case, we can just do this computation:

$$(1 - \theta) \cdot \text{count}(0) - \theta \cdot \text{count}(1) = 0 \tag{7}$$

So if you go and do the algebra, which is pretty straightforward in this case what you would get:

$$\hat{\theta} = \frac{\text{count}(0)}{\text{count}(0) + \text{count}(1)} \tag{8}$$

You see I use this estimate $\hat{\theta}$, special notation, would be actually equal to something that sounds like very logical.

This is kind of an obvious thing, so it looks weird.

You say, for example, I have a document which have 20 zeros and 10 ones. So what is the likelihood of observing 0 here? Would be 20 divided by 30, and that's exactly what this formula says.

But what we've demonstrated here is that, by properly defined our maximum likelihood criteria and making the estimation, we're actually getting the parameters that we are expected to get.

And in some cases, for some model, you can kind of intuitively say what they should look like. In other cases not, but this mechanism always works.