Lecture 15 - 8: Prior, Posterior and Likelihood

Tabaré Pérez

May 6, 2020

So now what I want to do, I told you, one of the limiting assumptions here is that I assume that the likelihood of being plus or being minus, it's exactly the same.

So sometimes we may have some prior knowledge and we want to enable our models to take advantage of it.

So I will demonstrate to you how we can break this assumption and look at a more general case. OK?

So the first thing that I would like to do is to remind you, the Bayesian rule, so that we are all at the same page. So we can write here:

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(B|A) \cdot \mathbb{P}(A)}{\mathbb{P}(B)} \tag{1}$$

This is just a Bayesian rule. There is nothing very special about it.

So now let's see what we can get when we actually are applying this Bayesian rule to our computation. So again, my goal here is to compute the likelihood that my document is going to get label plus, for instance. So what I will do, I will rewrite this expression:

$$\mathbb{P}(y = +|D) = \frac{\mathbb{P}(D|\theta^+) \cdot \mathbb{P}(y = +)}{\mathbb{P}(D)} \tag{2}$$

So there is a bit clash of notation. Let me explain what I mean, because we're kind of using slightly different notation to mean the same thing. So here I am saying, what is the likelihood that I'm going to assigned to document D its label y, which is plus. So I had rewrite it based on this rule, and we are now looking at two parts. The first part is given that this is the document which has label plus, we got the parameter θ^+, we are going to generate the D. Then we're looking at the likelihood of the label y to be plus. I could have wrote the likelihood of θ^+. But I just want to be consistent with nouns. And here we are dividing it by the likelihood of the documents. So first of all, let me introduce you some notations that you're going to see later on. So

this part, $\mathbb{P}(y = +)$ is typically called **PRIOR**. It's just the likelihood of a certain class to be like the likelihood a document will be spam. Even before we're looking at the document, each one of us have different ratio of spams in the email for example.

So this is even without looking at the document, we can say what is the ratio of spams we expect to see.

This one, $\mathbb{P}(y = +|D)$, is called **POSTERIOR**. Because we are looking now at the likelihood of the class when we already seen the document.

So now what I will do, I will actually look at this expression in more details and again translate it in to a very similar form, connect it again to the linear classification.

So what we can do here, we can, again, do the same story. We're going to take the log:

$$\log\left(\frac{\mathbb{P}(y = +|D)}{\mathbb{P}(y = -|D)}\right) = \tag{3}$$

$$= \log\left(\frac{\mathbb{P}(D|\theta^+) \cdot \mathbb{P}(y = +)}{\mathbb{P}(D|\theta^-) \cdot \mathbb{P}(y = -)}\right) = \tag{4}$$

And you can see the probability of the document disappeared here because we just can cancel it away. So now again, because of the log, I'm going to separate it into two subparts:

$$= \log\left(\frac{\mathbb{P}(D|\theta^+)}{\mathbb{P}(D|\theta^-)}\right) + \overbrace{\log\left(\frac{\mathbb{P}(y = +)}{\mathbb{P}(y = -)}\right)}^{\hat{\theta}_0} = \tag{5}$$

The first part of equation 5 we already seen in the past, correct? So we know already the value of this expression. That's exactly what we already computed here. So I'm just going to take it from here and re-write it:

$$= \sum_{w \in \mathcal{W}} \text{count}(w) \cdot \hat{\theta}_w + \hat{\theta}_0 \tag{6}$$

And then you remember there was a squiggly θ_w. Plus now I have this expression. And the same wave, just for simplicity, to make the notation really crisp, I make this substitution, you remember. I'm going to make the same substitution and just call the second part the equation 5, this whole expression as $\hat{\theta}_0$.

So now what we actually see here that we translated it again to a linear classifier.

But in contrast to the previous case when we had linear classifier that went through origin, now we have linear classifier with an offset, and the offset itself would be actually guided by our priors, which will drive the location of the separator.

So what we've seen here that in this model we can very easily incorporate our prior knowledge about the likelihood of certain classes.

And at the end, what we got, that even though we're talking about generative models and we're using a different mechanism on some ways of estimating the parameters of this multinomial, at the end, we actually are getting the same linear separators that we see in our discriminative modeling.

And before we move and we're kind of done with the multinomials, I just want to make one clarification point about the notations.

Because you remember, θs are actually defined over the alphabet.

So in this case, when we are thinking about the θs for the positive class and negative class, think about our cat and dog examples, that when we did the estimation maybe for the negative class , we get ${\theta^-}_{cat} = 0.7$ and ${\theta^-}_{dog} = 0.3$.

And this will be the θs for the negative class, and maybe, given our estimation procedures that we done, the θs for the positive class are ${\theta^+}_{cat} = 0.5$ and ${\theta^+}_{dog} = 0.5$.

So what you will do, whenever we have computing now, I give you a new document, which maybe only have cats in it or something, we will have a mechanism that would estimate for us the likelihood of it belonging to a certain class and getting the label plus or the label minus.

So here again, we will have multiple θs because we will have a big vocabulary, more than too many times, that can respond to the negative class, and different kind of parameters of θ for the positive class. And the relation between them would determine how this expression look like.