Lecture 15 - 10: MLE for Gaussian Distribution

Tabaré Pérez

May 6, 2020

From the last lecture we have:

$$\mathbb{P}(x|\mu,\sigma^2) = \frac{1}{(2\pi\sigma^2)^{\frac{d}{2}}} e^{\left(-\frac{1}{2\sigma^2}\|x-\mu\|^2\right)} (1)$$

So now, imagine to yourself you are actually given some training set, and we want to find, as what we've done before when we were talking about the question of estimation, how can they find the best μ and σ squared that will give the highest likelihood to my training data? Again, in this case, the training data will be:

$$S_n = \{x^{(t)}|t = 1 \dots n\} \tag{2}$$

We see that t would go from 1 to n because we have n points in our training data.

Again, since all the points are independent, whenever we are thinking about the likelihood of all these points being generated by our specific Gaussian, we just needed to multiply the likelihood of every training point to be generated by the Gaussian. So we're going to write:

$$\mathbb{P}(S_n|\mu,\sigma^2) = \prod_{t=1}^{n} \mathbb{P}(x^{(t)}|\mu,\sigma^2) \tag{3}$$

The product of t going from $1 \dots n$, because we have n points, and then the likelihood for $x^{(t)}$, μ,σ^2.

And I can continue and just copy 1 here instead of writing $x^{(t)}$.

So what do we want to do for our estimation task? For our estimation task, for given set of points $x^{(t)}$, we want to find the σ and the μ which gives it the highest likelihood.

In the same way as we've done it before when we were talking about multinomial, instead of directly looking at this expression, we're going the log it because it will be easier for us to work with it. So we are looking at log of this expression:

$$\log\left(\prod_{t=1}^{n}\frac{1}{(2\pi\sigma^2)^{\frac{d}{2}}}e^{\left(-\frac{1}{2\sigma^2}\left\|x^{(t)}-\mu\right\|^2\right)}\right) = \tag{4}$$

And what I will do just to reduce the amount of copying I have to do, instead of writing this probability, I am actually going to write this whole expression from 1 above.

So what we can do, as we've already seen many times today, when we are looking at log over the product, it's actually going to be the sum of logs, correct?

So we're going to write now:

$$= \sum_{t=1}^{n}\log\frac{1}{(2\pi\sigma^2)^{\frac{d}{2}}} + \sum_{t=1}^{n}\log e^{\left(-\frac{1}{2\sigma^2}\left\|x^{(t)}-\mu\right\|^2\right)} \tag{5}$$

So now we can actually do some log computations:

$$= \sum_{t=1}^{n}-\frac{d}{2}\log(2\pi\sigma^2) + \sum_{t=1}^{n}-\frac{1}{2\sigma^2}\left\|x^{(t)}-\mu\right\|^2 = \tag{6}$$

So the first thing to notice in the first portion of equation 6 is that we are not having any dependence on the t. We're just going to n times sum of this firts portion of the equation. So instead of writing the sum, we can just multiply it by n:

$$= \underbrace{-\frac{nd}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{t=1}^{n}\left\|x^{(t)}-\mu\right\|^2}_{\mathcal{L}} \tag{7}$$

So now, we have this expression and we're trying to find σ and μ that will give the highest value of this expression.

And we are going to do exactly the same things that we've done in the past.

Let's call the expression in 7 as $\mathcal{L}$. So in this case, we are going to take this expression and differentiate it with respect to μ and make it equal to 0 and similarly, we are going to do the same with respect to our σ and make it equal to 0:

$$\frac{\partial \mathcal{L}}{\partial \mu} = 0, \frac{\partial \mathcal{L}}{\partial \sigma} = 0 \tag{8}$$

And from there, we can compute the μ and the σ and that, actually, will be your homework, to complete this computation.

And what you will get here will be, again, something extremely intuitive.

We will find out that the μ that we are going to get here, would be just the mean of all the observed point.

And similarly, you would find that a σ^2 is what we expect it to be, which is the average variance of the points from the μ.

$$\hat{\mu} = \frac{1}{n} \sum_{t=1}^{n} x^{(t)} \tag{9}$$

$$\hat{\sigma}^2 = \frac{1}{nd} \sum_{t=1}^{n} \left\| x^{(t)} - \mu \right\|^2 \tag{10}$$

So you would go exactly through the same machinery as we went through the multinomials to see that those are the same mechanisms that are used to do this computation.

So in our next lecture, we're going to continue with the Gaussians but we are going to make it slightly more interesting and, instead of telling you that we're just going to always assume we have one mean and one variance, we would actually assume that there can be many different means, and there can be many different clouds.

We will see how this expansion is done and how we can estimate it.