Lecture 16 - 4: Mixture Model - Observed Case

Tabaré Pérez

May 6, 2020

$$\mathbb{P}(S_n|\theta) = \prod_{i=1}^{n} \sum_{j=1}^{K} \mathbb{P}_j \mathcal{N}(x, \mu^{(j)}, \sigma_j^2 I) \tag{1}$$

MLE for $\mathcal{N}(x, \mu, \sigma^2)$:

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} x^{(i)} \tag{2}$$

$$\hat{\sigma}^2 = \frac{1}{nd} \sum_{i=1}^{n} \left\| x^{(i)} - \mu \right\|^2 \tag{3}$$

So right now, the next question we have to address is how we can find all the parameters that we are interested in, which would be the mixture weight, and then the μs and the σ^2 for each mixture component?

And the easier case here is to start with an example where we know exactly where each point belong and we will call this case, observed case.

Now, what I'm going to do in the observed case, I'm going to introduce some notation which would look so cumbersome.

You may be thinking, why wouldn't we use some simple notation, why we're making all these really complex formulas, when it's so simple?

And the reason we're going to be doing it, it's not because of my cruelty, but because we afterwards, using this annotation, directly translate them to unobserved case, which is more complex. Not much more complex, but still.

So let's start with the first piece of notation:

$$\delta(j|i) = \begin{cases} 1, & x^{(i)} \text{ is assigned to } j \\ 0, & \text{otherwise} \end{cases} \tag{4}$$

So this piece of notation would tell us, δ is an indicator function, which would be equal to 1, when we know that point $x^{(i)}$ is assigned to j, to our mixture component j and 0 otherwise.

Again, as I said earlier, this is an observed case. We have a hard assignment. So for every point i, there will be just one j to which it belongs.

The next point that we are going to make is actually now to use this particular notation to rewrite formula 1.

Because remember, what we're trying to do is to find $\mathbb{P}$, μ, and σ^2. So what I will do now, is to do what we've done before, to say what expression I'm trying to maximize. So we will take this likelihood expression and, as we've done in the past, to do it log likelihood. And I will write it using this notation, you will see. So let's start here:

$$\sum_{i=1}^{n} \left[\sum_{j=1}^{K} \delta(j|i) \log \mathbb{P}_j \mathcal{N}(x^{(i)}, \mu^{(j)}, \sigma_j^2 I) \right] = \tag{5}$$

I'm logging it, so the product is going to become a sum.

Then we go through all the points. And then we are looking, to which clusters do they belong? So we're going to go through every cluster, like for every point we are going to see to which cluster it belongs. So I'm going j from 1 to K, and then within that particular cluster, I'm only going to take points that are assigned to this cluster.

And now, I am going to change this summation. So in this summation, you see, we go through all the points, and find the relevant cluster and then sum up the expression.

What I can do here is actually switch the two and make this computation, instead of going through the cycle, through all the points and through the clusters, instead I independently will do this computation for each individual cluster. So I am going to rewrite this expression, putting now the internal sum on the outside, j from 1 to K, and then we're going from 1 to n, and this is the whole expression.

$$\sum_{j=1}^{K} \left[\sum_{i=1}^{n} \delta(j|i) \log \mathbb{P}_j \mathcal{N}(x^{(i)}, \mu^{(j)}, \sigma_j^2 I) \right] = \tag{6}$$

So let's just look at this expression. So what this expression actually demonstrates to us is that, whenever we are trying to find all the parameters that optimize for each mixture component, we can actually do this computation for each cluster independently. We can independently find the best parameters that fit the cluster, because the fact that these points are observed, we actually know where the point is belonging to.

So we can separately find, here all the stuff and we can separately do the computation here. So now I want you to look back at this specific expression. So we're kind of coming back to the situation that we already have been in.

We separately looking at each one of these Gaussians and for each one of them we can do the computation, similar to what we have done before.

So let's see how it unrolls. And I would like to again use here this notation that hopefully, at this point, already makes sense to you, and to compute different quantities. So let me start with the first quantities that I'm interested in.

The first thing I want to do is to compute how many members belong to each class using this δ notation, to each cluster:

$$\hat{n}_j = \sum_{i=1}^{n} \delta(j|i) \tag{7}$$

We just need to sum up all $\delta(j|i)$, when i goes from 1 to n.

So we will see here, for all the points that belong to cluster j, and only for those point, this would be equal to 1.

So we're going to sum, and it will give us a number. So now, how can I compute my first parameters that I am interested in?

How can I compute their mixture weight for cluster j?

Very simple, using maximum likelihood estimate. We're just going to say, again, we can be really formal here and take the derivatives, make them equal to 0, and maybe we should keep it for the exercise.

But if you go through it, what you will discover, as we've done before, something very intuitive:

$$\hat{\mathbb{P}}_j = \frac{\hat{n}_j}{n} \tag{8}$$

It's just going to be the number of the members in this cluster divided by total number of points. So this will be the weight for the mixture component j.

Now, similarly, we can compute the mean of that cluster. And here comes two ways to think about it. One way to think about it, and again, some of you can just take this internal expression and differentiate it with respect to μ and find it and go through the whole process. And what you will discover, is that it will have a very similar shape as what we've done for individual Gaussians:

$$\hat{\mu}^{(j)} = \frac{1}{\hat{n}_j} \sum_{i=1}^{n} \delta(j|i) x^{(i)} \tag{9}$$

We're just going to sum up all the points and divide by the size of the cluster. In this particular case, what we will get will be sum of all the points that belong to this cluster. And again, how do we say it, we go through all the points from 1 to n, divided by the size of this cluster.

So this would be our estimate, maximum likelihood estimate, for the center of the component j.

And the next thing that we need to do is to compute the variance:

$$\hat{\sigma}_j^2 = \frac{1}{\hat{n}_j} \sum_{i=1}^{n} \delta(j|i) \left\| x^{(i)} - \mu^{(j)} \right\|^2 \tag{10}$$

And again, we can go take the derivatives with it. What you will get would be an equivalent of this formula that was developed for the case of a single Gaussian. So we're going to get 1 divided by the size of that component times d for the dimensions and now we will go select all the points that belong to that mixture.

So again, this is a mechanism where indicator functions just make sure that we are selecting point that really belong to this specific cluster.

So what I've done so far, I've demonstrated to you how, given the observed case, when we know to which component each point belong, I've demonstrated to you how we can estimate all the parameters that we need to define our mixture of Gaussian.