IIT Madras

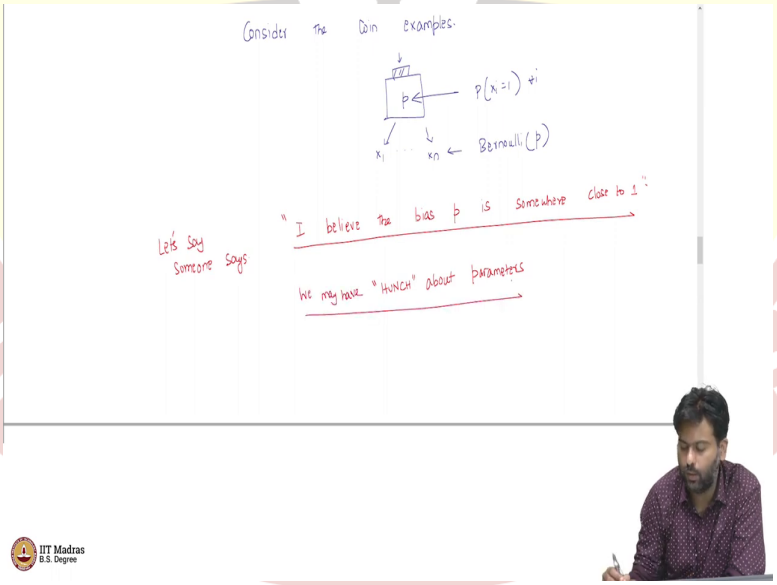
ONLINE DEGREE

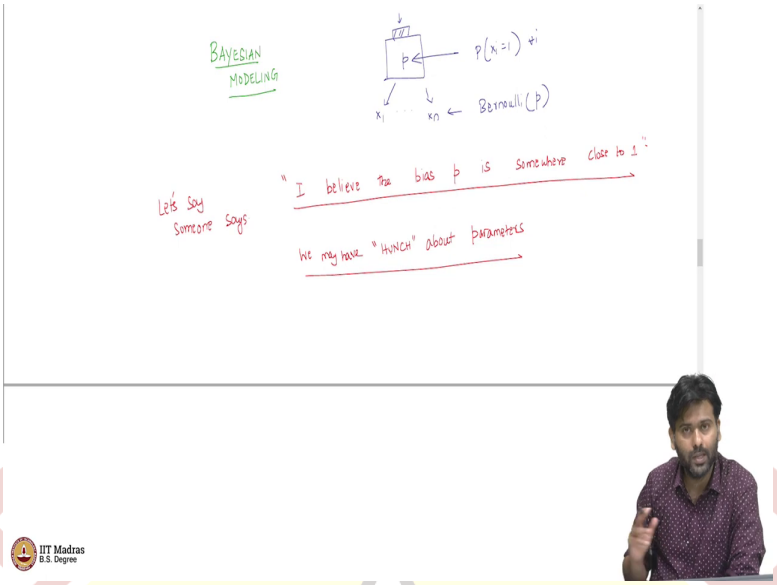
Machine Learning Techniques
Professor Arun Raj Kumar
Department of Computer Science and Engineering
Indian Institute of Technology Madras
Bayesian estimation

So, this is a good estimator given data, it gives me good estimators. Is there something else that might be typically available in practice? And if so, is the maximum likelihood estimator still the best thing to do? Or is there something else that one can do? Of course, I am being vague here. So, let me make that a little bit more precise. And then and then we'll try to see how you can potentially come up with different estimators, which might be in some cases better than the maximum likelihood estimators.

And to do this, let us again, go back and revisit our simple example of coin toss. So, our model is still the coin toss model. You observe data, which is 0 and 1. And you are going to make an assumption that it comes from a box with a coin with some unknown bias P, of generating heads and you press it n times, you get the coin. And you get n data points, n observations all that is the same.

(Refer Slide Time: 01:22)





But now, the extra thing, piece of information that you have is that somebody comes and says. So, let us put me this topic, consider the coin example. By the way, the coin tosses are called as Bernoulli trials. So, it like how Gaussian. So, the corresponding random variable in the coin are called as Bernoulli trials.

So, you have a coins x1 to xn. So, this is basically all my xi's or Bernoulli that is what means with some parameter P, this is the probability of xi = 1 for all i and it is the same setup. We have this in addition, let us say somebody came and told you the following statement, "I believe the bias p is somewhere is close to 1." Let us say someone said this.

Now, imagine a situation where you have this box. And the box has a coin inside about which at this point, you do not know anything. Let us say you have not even seen the data. But then somebody walks in and says, I believe that the bias p is somewhere close to 1. Now, you have this extra piece of information, which is an English sentence. So, somebody says that they believe that p is close to 1, that is an English sentence. Nevertheless, it is a statement that gives you some information about the coin inside the box. And this is what we will call as, in general domain knowledge, let us say.

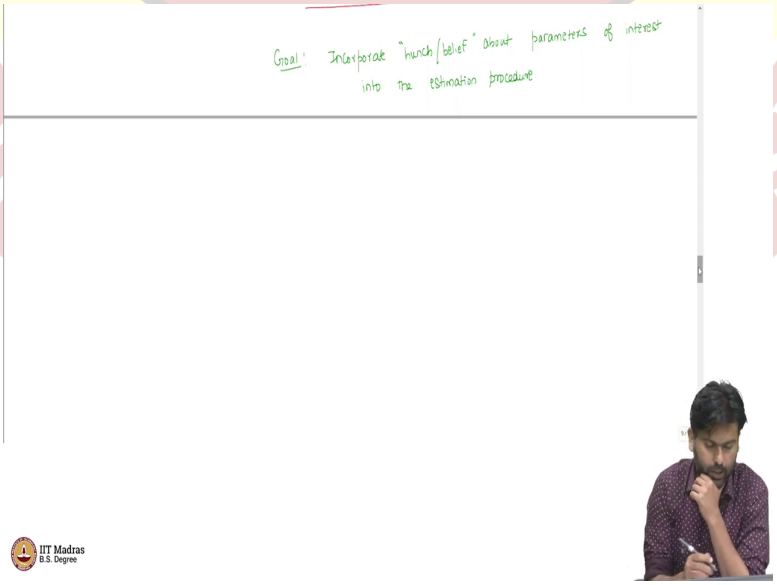
So, the statement could be anything. It could be closer to 1 not maybe they are saying something like, I know, it is small, the p is either small or either too high. But then I am sure that it is not close to 0.5. That is also a statement somebody could make it. So, all these are statements that people could make or you might have as a practitioner from your experience, you might have what I am going to call, as hunch about data about the parameters. Not the data about the parameters. We may have a hunch about the parameters.

So, remember, these hunches have nothing to do with the data. So, we have not even seen a single data point it. Even before seeing the data points in the previous case of maximum likelihood estimators. If I did not give you data and then I asked you, well, what would be your guess for the underlying model that generates the data? You go blank. So, because there is no way you could make a guess, because your method depends only on data. Only if you give me data, I will be able to see what might be a good guess.

If there is no data, there is no other information that I have. Maximum likelihood estimator depends completely on data. But in practice, you might have something more than the data, which is the hunch that I am talking about here. And now, it might be good and it will be very good if we had a principled way to incorporate our hunch into our estimation process. So, is there a way we can somehow codify our hunch into mathematically more precise mechanisms that can be incorporated into our estimation procedure itself. If so, how can we do this?

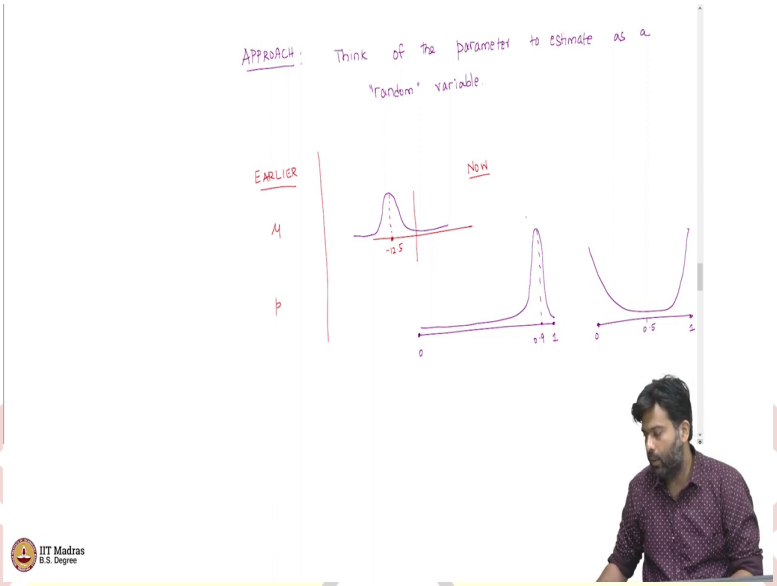
So, this is what we are going to see next. And this will take us to what is known as a Bayesian modelling approach. Again, for people who have seen this, this might be a recap, otherwise, this can be thought of as a primer in Bayesian modelling. So, the goal is to incorporate these hunches that we have that we might have. So, how can we do that? So, that is the question.

(Refer Slide Time: 05:44)



So, goal incorporate hunch or belief about parameters of interest into the estimation procedure and how do we do this? The way we are going to think of this is as follows.

(Refer Slide Time: 06:26)



So, the approach that we will take to do this is as follows. So, we are going to think of the parameter that we are trying to estimate as a random variable. And I will tell you what that means, intuitively, but that is the, that is the approach that we are going to take. So, basically, we have this hunch. So, earlier we were thinking of the parameter as some μ or some p. So, this is what we were trying to estimate.

Now, we are seeing. So, we are going to not, in some sense, we are going to encode our hunch as a distribution over this underlying parameter. So, earlier, this is what our goal was to get a μ or now we are going to say I have some hunch, which is to say that, well, if I say, for example, for the case of p, let us look at the case of p I know that the value of p can take any value between 0 and 1.

Now, if I if somebody said that, well, I think I believe that the value, true value of p is somewhere close to 0.9 or something like that, or it is close to 1, I might potentially put a hunch like this. So, what is this tell us ? This tells us that I believe even before seeing the data, that my true value of P is most likely around 0.9 or closer to 1. So, in this case, I am saying 0.9. I mean, that is just an example. So, this could be 0.9. Depending on our hunch, we can modify this. So, this might be one way to encode your hunch, or a statement of the form, well, I know that the p is not close to 0.5, it is either too small or too big.

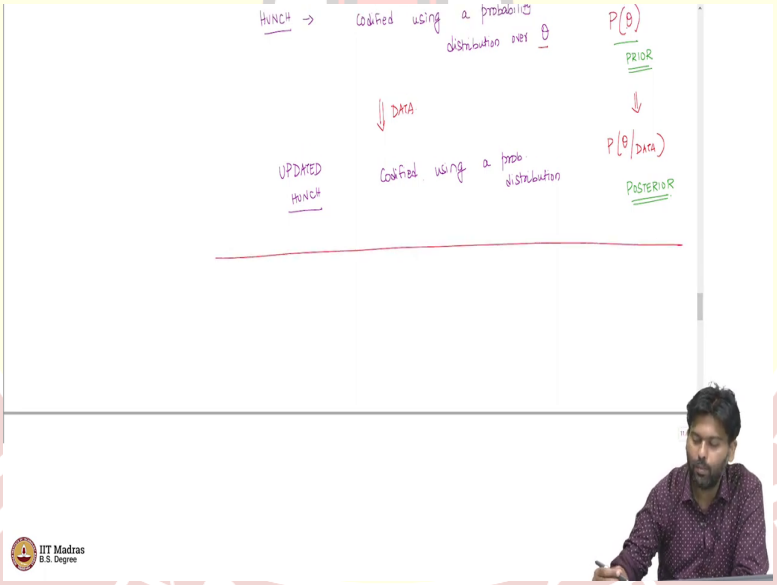
So, maybe you want a hunch that encodes that, maybe you would put a distribution like this. So, this is still between 0 and 1. This is 0.5. And essentially, you can think of it as putting weights on each of these values of p. So, because p can take in this case, any value between 0

and 1, you can encode these weights in terms of a distribution itself. So, which means I can treat this hunch itself as a random variable as a parameter itself as a random variable, which has some associated distribution with it.

Now, what does the distribution tell me? Well, without seeing the data, it kind of tells me that it gives me what is the chance, I believe, before seeing the data that my true parameter falls in a particular interval. That is what these things are encoding. So, for p it could be like this. For μ, you could say that well, I think the mean μ, the parameter I am trying to estimate is probably around -12.5. So, maybe that is the guess that I want to encode. Maybe I will encode that something like this.

So, it is around 12.5. It could be other things also, but then I believe it is more around 12.5 than anything else, if that is what I believe, then this is a way I could have encoded that hunch, before I see the data.

(Refer Slide Time: 10:09)



So, now, what happens is, so, what are we saying? We are saying that we have a hunch, which is what I am going to call as codified using a probability distribution over θ. Let us say θ is the parameter that I am trying to estimate. Let us not fix μ or p specific values or specific parameters. But in general, it is some parameter θ, which means that there is some p of θ. What is p of θ give me? Well, if θ is a continuous parameter, like your p or μ, then it means that it is a continuous probability distribution, where the support the values that θ can take or any values that you could have potentially guessed.

For p, this could be any value between 0 and 1 and the shape of this p, the PDF determines, what is our belief about this the parameter of interest even before we see the data. So, the hunch that we have can be codified using a probability distribution over θ, which can be said as P(θ), which simply tells us that, if θ is takes value in a continuous range. For example, like the p that we were trying to estimate in case of Bernoulli random variables, then this P(θ) would actually be a PDF. So, it tells us that what do we believe about this underlying p in terms of probabilities.

In other words, if I did not see the data and I asked you, well, what is the chance that this p takes a particular value in a particular range, for else takes a value in a particular range, then you can integrate this PDF and then give me the probability and so on. So, essentially, we are treating θ as a random variable. That is what it means.

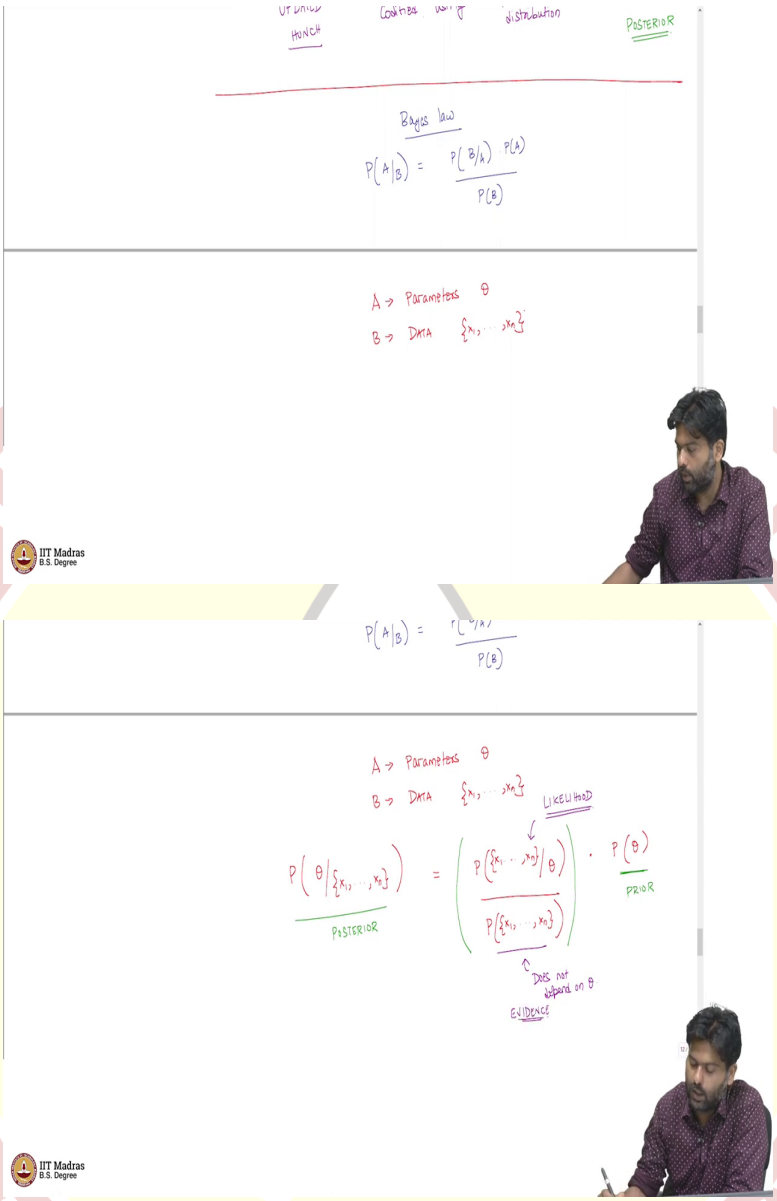
Now, what do we do with this hunch? Well, of course, we see the data next. So, after this, we have the data. And once we see the data, our belief system needs to be updated. So, we may believe that the p is around 0.9, which means that we believe that, the chance of heads is much higher than the chance of tails.

But then if we observe 1000 data points and it so happens that, 900 of them are tails, then it is against our belief system. So, our belief system said that we are expecting 90 percent heads, but then we actually are seeing 90 percent heads on average, but then we are actually seeing, let us say 90 percent tails, which means that we have to update our belief system accordingly, as how the data dictates.

Also, if, our data adheres to our belief system, then that strengthens our belief system. So, then, we might still want to update our belief system where we might be more confident about our guesses and so on. So, in any case, we after looking at this data, we need to move from hunch to an updated hunch. So, this needs to happen. And how can we codify the updated hunch? Well, this can be codified using again a probability distribution, but then what distribution is this? This is no longer p(θ), but then it is p(θ) given data.

So, this idea of going from what is called as prior distribution over θ to what is called as a posterior distribution of θ, but then after observing data, is what is called as the Bayesian way of doing things. So, you have a hunch, you see data, you update your hunch. So, this is the Bayesian modeling.

(Refer Slide Time: 14:15)



But then what is so Bayesian about it. So, where is Bayes coming into the picture? Well, that is that precisely happens to describe how you go from p(θ) to p(θ |data). So, where is the base coming in? So, if you remember the Bayes law, or the Bayes theorem, now, we know that P (A | B) is from high school, we know that this is P(B| A) . P (A) / P(B). So, this is our standard Bayes Rule, which says that A conditioned on B the probability of a condition on B can be gotten as with using P(A), P (B) and P (B | A). So, if we know these three things, I can get P (A | B).

Now how does this help in our case? We are going to think of a as parameters, so, which is simply our θ and B as our data, so, which is x1 to xn. So, which means simply by using the Bayes law, we can do the following, we can see that p (θ | {x1,.., xn}), which is our updated

hunch, after seeing data can be written as P (B | A), which is P({x1,..., xn}). The data given the parameter into P (A), which is the parameter divided by the data P({x1,..., xn}). This is just exactly analogous to our Bayes law.

And now, if you notice this, it gives you an excellent way to go from your prior to your posterior. And what Bayes theorem is saying is simply the following. So, you have some initial belief about your prior distribution. Now, to go to a posterior distribution, you have to review your prior distribution. You have to make a multiplicative update to this prior and that multiplicative update is given by this specific term, which also is something that we have encountered earlier.

So, now, at least the numerator is something that should be familiar to you. Think about what the numerator is. The numerator is saying, well, if I give you θ, we told you what the parameter is, what is the chance that I see this data? Now, this is what we have been calling as the likelihood. So, this is something that we already have seen.

Now, the denominator is something that is independent of θ. So, this does not depend on θ. So, this is the technical term for this is called as evidence, this is the chance that you actually observe the data itself, but then notice that it does not depend on θ. So, which means that you can think of your posterior as proportional to your likelihood times your prior. So, you are re-weighing your previous belief, using the likelihood and then that will give you the posterior. So, that is what Bayesian modelling essentially is telling.

(Refer Slide Time: 17:20)

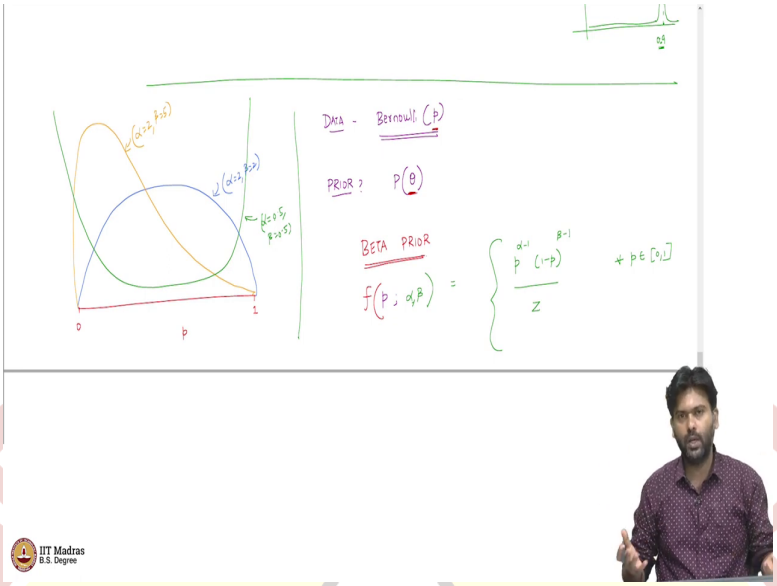
So, for example, I might have a hunch, like this. Example. Maybe I had a hunch somebody came and told me that be that I am trying to get data from is close to 0.9, then I could have, incorporated that hunch by using a PDF. Let me draw this carefully. So, PDF something like this. So, which peaks at 0.9, let us say. Now I see my data. And let us say my I say 10 data points, which are many of them are 0, let us say. So, eight of them are 0 and two of them are 1s. Again, all of these representative images, it is not exact numbers that I am plotting, but then just to give a feel.

So, now the data is kind of the likelihood is kind of telling me suggesting me that, the p value should actually be 2 by 0.2 whereas my belief is saying it is 0.9. So, now, if I somehow combine these, what I might get is something like this. So, I might get something like this as the updated hunch. So, this is my hunch. This is my updated hunch. It might peak somewhere, perhaps at 0.27, I do not know. These are just numbers that I am making up. But you get the idea.

So, you start with some distribution over your possible parameter values. You see the data and then you get a new distribution, which is perhaps a different distribution. Now, if I had a different set of data points, now, this is a case where the, the data does not know correspond to my hunch.

On the other hand, if I had the flipped version of this data, let us say something like this, where you had nine 1s and one 0. Now, in this particular case, the same hunch might translate to an even sharper rise at 0.9. So, I am believing in 0.9, even further, strongly, so, because my data also in some sense corresponds to my hunch. So, this is the idea of Bayesian modelling.

(Refer Slide Time: 19:41)



Now, let us take one simple example, to talk about Bayesian modelling. And then we will move on to other types of setups. So, again, we will talk about how to encode a hunch when data is setup is data as Bernoulli, which is the coin toss example, Bernoulli of p. So, this is the basically when the assumption that we are making about the data is that it is Bernoulli of p. So, there is a box with a coin, all the things that we have discussed. So, the likelihood is Bernoulli.

Now, what is a good prior? Which means, P(θ)? In this case, θ is just P, which means how can I encode my prior? Well, of course, here I have been drawing pictures, but then mathematically how can we encode this prior? One way to encode a good way to encode this

prior is using what is called as the beta distribution. It is called a beta prior. And we will see why this is a good way to encode prior.

Basically, you want some continuous distribution whose values are between 0 and 1. We cannot use a Gaussian here. So, because Gaussian can give me any value between $-\infty$ and ∞, but then the parameter I am trying to estimate which P, I know takes value between 0 and 1, so any distribution that I use as a prior to encode my prior knowledge should be supported only in 0 and 1. And a good choice is what is called as a beta prior.

So, once I say a beta prior, I should put down the density of this distribution. So, the density of the beta distribution looks like as follows. The density is defined for every value of p between 0 and 1. And like the Gaussian is parameterized using the mean and the variance, the beta distribution is parameterized using two values, α and β, which are both positive numbers. And it is given something like this.

So, this is $p^{(\alpha - 1)}$. $1 - p^{(\beta - 1)}$ for all p is 0 , 1. Of course, divided by some normalizing constant z, which does not depend on p, such that this is a PDF, it integrates to 1 and so on. That is not so important for us, what is more important is the functional forms, how it depends on p.

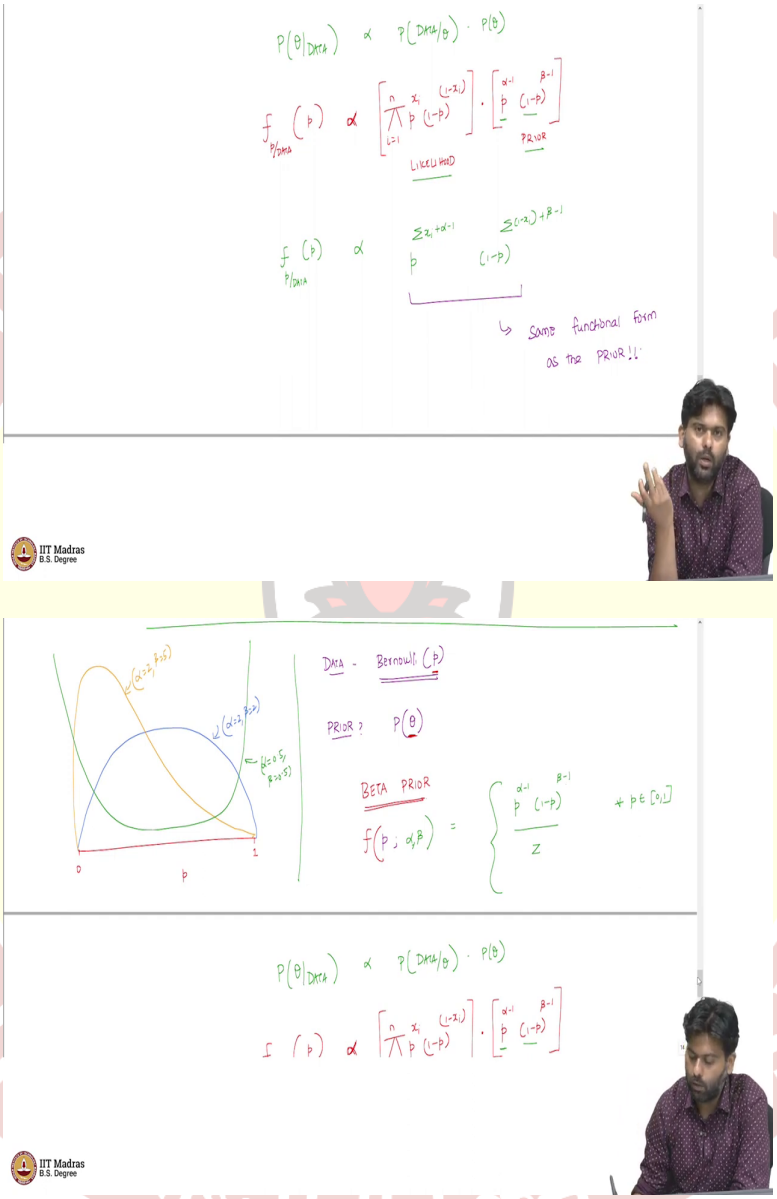
So, what can this beta prior do? Let us, take some examples and see what kind of PDF this looks like. So, if you take 0 to 1 which is the value of p and then if I use, let us say p as α as 2, 1 beta as 2, I get something like this. So, this is my PDF, when $\alpha = 2$ and $\beta = 2$, meaning my this is essentially the function $p \cdot (1- p)$ divided by some normalizing constant such that the area under this curve is 1, because it has to be a probability distribution. And we know $p \cdot (1 - p)$ looks something like this. So, this is the case $\alpha = 2$ and $\beta = 2$.

Now, when α is, let us say 2 and $\beta = 5$, it looks something like this. Again, these are representative images, this is $\alpha = 2$, $\beta = 5$. It kind of says that when you have small α and big β, then you are kind of believing that you are true p is somewhere closer to 0 than 1. If $\alpha = 2$ and $\beta = 2$, then it is like I am still believing that the coin is pretty much unbiased. I mean, my highest value is 0.5, but then I am spreading out my bets, so to say.

Now, for a different choice, for example, if $\alpha = 0.5$ and $\beta = 0.5$, then this picture actually looks something like this. This function looks like this. It is a very flexible distribution. Now, this is a case where I can encode my belief that well, my p is either small or large. But it is definitely not close to 0.5. So, it is a biased coin. So, it is a skewed coin that much I know.

So, now that can be encoded using the choices of α = 0.5 and β = 0.5. So, this can capture different types of intuition. Let us, say there is some intuition that we have, which can be put down using some choice of α and β.

(Refer Slide Time: 24:11)



So, now what do we do with this prior? Well, of course, we need to write down the posterior now. So, we need to write down the well, we need to write on p (θ | data). And we said that well, this is proportional to p (data | θ) p (θ). And all this prior is telling us is like is the PDF for p (θ).

Now, how does the p (θ | data) look like if your data comes from a Bernoulli likelihood, that is what we want to find out. In other words, we want to find out the PDF of p given data at

some value of p. Now this is proportional to the Bernoulli likelihood, which we know looks like $\prod_{i=1}^{n} p^{xi} . (1-p)^{(1- xi)}$. You have already seen while we discuss maximum likelihood that this is our Bernoulli likelihood. So, this is our likelihood function, our foreseeing, of course, the data x1 to xn and p (θ), we are saying can be encoded as $p^{(\alpha - 1)} . (1 - p)^{(\theta - 1)}$.

So, this is our prior. So, this is prior, this is likelihood. And if I multiply this I should get, of course, I have to divide by the evidence, but then I am that is why I am not saying equal to but then proportional to what I would get as the PDF of the posterior.

Now, the interesting thing that you might already be noticing is that the prior and likelihood both have the form p power or something into 1 - p power something. So, they look similar, at least functionally. So, which means I can somehow combine this and say that this is proportional to p power sum over xi, this guys, $p^{(\sum xi + \alpha - 1)} . (1 - p)^{(\sum(1 - xi) + (\beta - 1))}$. So, this is what my PDF of my posterior looks is proportional to.

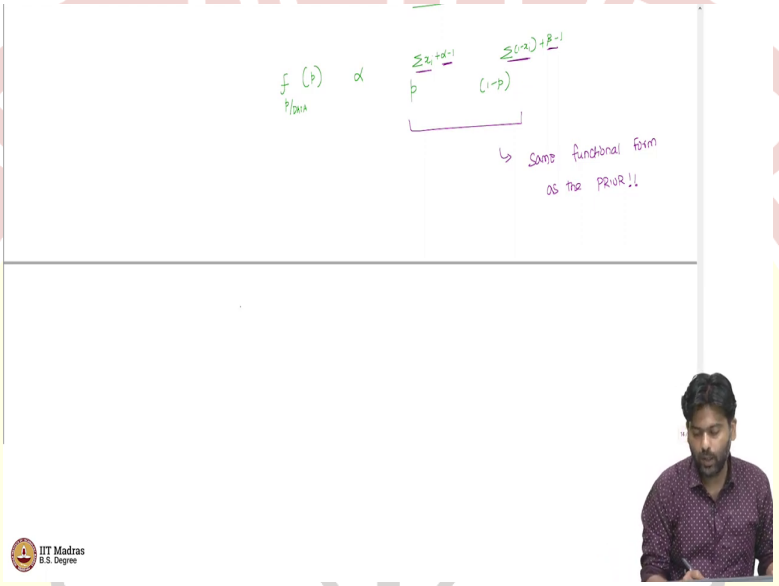
So, it is just a simplification of this thing. Of course, I am seeing proportional to because there is some normalizing constant, which is needed to make this a real PDF, it has to integrate to 1 and so on. But we will see that that is not so important, as we will see in a minute. So, what does this tell us? So, the very interesting thing that this tells us is that f (p) was of the form p power something into 1 - p power something. Our prior shape was of the form $p^{(\alpha - 1)}(1- p)^{(\alpha - 1)}$.

Now our posterior also, this distribution also has the same functional form. So, same and that does not happen for all choices of prior. I will comment about that in a minute. So, this is some functional form, same functional form as the prior. So, the posterior also has the functional form p power something into 1 minus p power something, which means that we know that the posterior distribution is also a beta distribution.

Now, the parameters are no longer α and beta. For the prior it was α and beta. But now the posteriors parameter are not α and beta. They depend on the priors parameter α and beta. But then they also depend on the data. It is a beta distribution but then the parameters have not changed. It is as if you are only updating the parameter of the distribution. And you do not have to exactly calculate the density for at all values of p. You do not have to completely calculate the entire function, so, entire density. If you know the parameter, then you get the function for free.

So, interestingly, this has not happened by for all choices of priors. So, if I did not choose a beta distribution, if I had chosen some other complicated distribution between 0 and 1 that encoded my prior knowledge, it is not necessary that if I multiply it with the Bernoulli likelihood, I will get a beta posterior. That is not at all necessary. So, you will still get a posterior. It might be useful and all that, but it is need not be convenient in the sense that your prior and the posterior of the same functional form.

(Refer Slide Time: 28:35)



In this particular case, it happens. So, we start with the beta prior, which means we have some parameters, α and β . And now, after seeing data, the updated hunch is a beta posterior, it is also beta and the data is Bernoulli, so, Bernoulli. Now, what are the parameters of this posterior? Well we pause and think about it, I will tell you now. So, this is just $\alpha + \sum xi$, $\beta + \sum(1 - xi)$. Now, one way to think about $\sum xi$ is just the number of heads and nh number of heads in my data and $\sum(1 - xi)$ is the number of tails.

So, what is this kind of tells us? It tells us that if I started with some value, α and β, I observed in nh tails and nh heads and nt tails in my data. Then my updated belief about the parameters looks like a beta distribution with parameters, number of heads + α and number of tails + β.

Now, if you had used a simple maximum likelihood, then it would be only number of heads and number of tails which would have determined our guess. So, it would be nh /nh + nt, which is just the n. So, nh / n would be our guess which is what we saw as our guess for the maximum likelihood estimator. Now here we are seeing, one way to guess, after getting this

better posterior could be guessing as α plus nh divided by one possible guess. Guess could be to look at α + nh / α + nh + β + nt, which is α + nh / α + β + n, nh + nt is just n.

Now, this might be our guess. If you had to commit to a single guess, then we might commit to this guess. And in fact, this turns out for this particular case to be the expected value of the posterior. So, is expected value of any beta distribution with parameters α and β is α / α + β. In this case, the posterior is a beta with (α + nh , β + nt). And the expected value of this turns out to be exactly this value.

Now, this is kind of telling us essentially that, let us think about this for a second. So, we have some data which have nh tails and the nt, nh heads and nt tails. If you had to do a maximum likelihood estimator, we would have said that number of heads by n that is our estimate.

Now, we are saying number of heads + α / n + α + β . It is as if we are saying that we have our data. And we also have this extra ghost data, which have α heads and β tails. If we had our data and an extra α +β data points where we had α heads and beta tails, well, then we would have guessed our maximum likelihood estimator as α + nh / α + β + n, which is exactly what is the expected value of the posterior.

Now, what does this tell us, this tells us that in this particular example, the prior using the beta prior can be thought of as if thinking that we did some ghost experiments or pseudo experiments, which are not real data, but then these pseudo data ghost data which have α+β data points, out of which α showed up heads and β should up tails that is what is giving us our beliefs. So, it is as if somebody came and said, I did 100 experiments, 80 of them happened to be heads, 20 of them happen to be tails.

Now, if somebody told us that, then we converted that into a hunch using a beta prior with α as 80 and β as 20. Now, it is as if we can once we do this, the posterior will take into account these ghost data samples also and then will give us a maximum likelihood estimator, it could be thought of as that way. So, that is that is one way to interpret what is going on here.

So, in general, you could, this might be one way to make a single guess from the posterior, which is using the expected value of the posterior. There are other ways you can make guesses, which is what is called as a map estimator. Which is to say that well, in maximum likelihood, we wrote down the likelihood function and we pick the value which maximizes the likelihood.

In the map estimator, which is a maximum Aposteriori estimator. We look at the maximizer of the posterior distribution. So, now you have a posterior distribution. Now, you look at which value has the maximum PDF. So, PDF value or the likelihood and then you make that as your guess. So, that is typically called as $\hat{p}_{MAP}$. And in this case, it will be slightly different from α by $\alpha + \beta$. So, the point is that now you have an updated hunch, which gives you belief system over all possible values that your parameter could take. Not committing to a single value.

If you want to commit to a single value, you can either take the mod of this distribution, which is the maximum Aposteriori estimator. You can take the expected value. You can take whatever you want. So, you can take some samples, average them, everything is possible, because you have an entire distribution in your hand. So, this way of modeling things is what is called as a Bayesian way of modeling things.

So, to summarize, at a very high level, we have looked at two different types of estimation procedures. One is called as the maximum likelihood procedure, which just writes down the likelihood function and then tries to maximize it to get a point estimate. The Bayesian world starts with a hunch about our parameter to estimate and then converts that hunch into an updated hunch via the data and using the Bayes theorem.

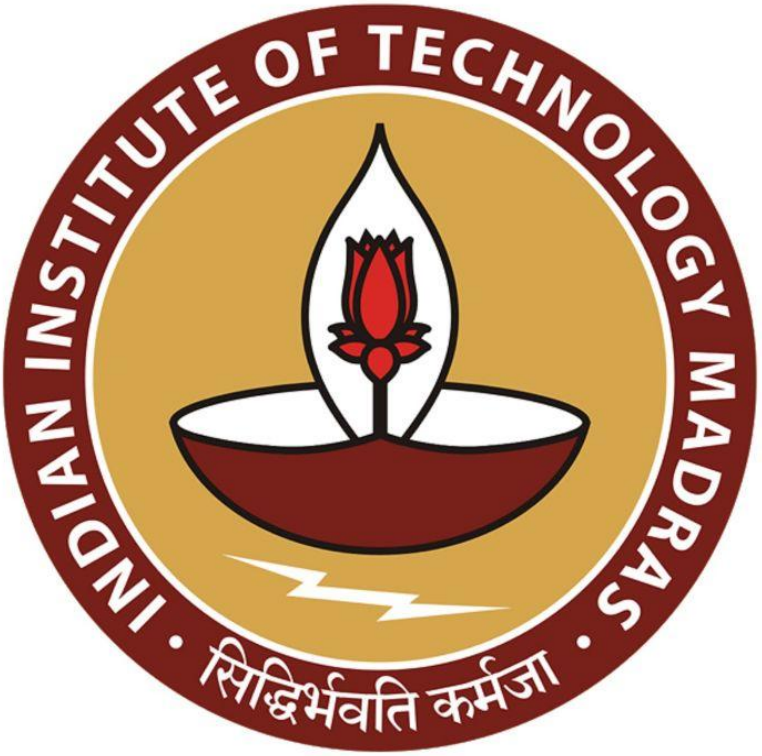
And once you earn a updated hunch, which is a distribution over all possible choices our parameters can take. Then you can either convert that into a single estimator if you want by taking the Aposteriori maximum Aposteriori estimate or you can take the expected value or you can do whatever you want. So, these are two broad ways of doing estimations.

Now, this is all you know basic statistical ideas. What we are going to do next is use these ideas specifically the principle of maximum likelihood and see how that can be used for a very, very specific unsupervised learning problem, which will give us a probabilistic twist to a clustering algorithm that we have already seen, which is the K-means algorithm. So, for K-means functions did not assume any probabilistic model for data.

So, now, if you assume a reasonable probabilistic model for data, can you come up with a probabilistic version or a probabilistic counterpart to clustering algorithm? Or can we come up with a probabilistic counterpart to the representation learning algorithms? So, these are the type of questions we are going to ask next. And that will give us the real power of using all these techniques that we have learned here including maximum likelihood or even Bayesian

methods, when we want to apply it to specific unsupervised learning machine learning problems like representation learning or clustering.

And the next thing that we will see is how to use this for estimation ideas for coming up with a probabilistic version of the clustering algorithm. We will see that next time. Hope you enjoyed this video. Thank you.

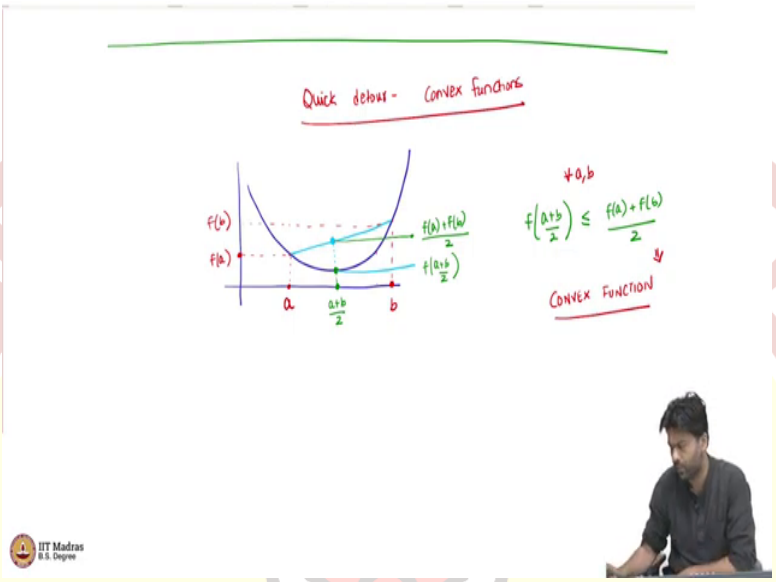


IIT Madras

ONLINE DEGREE

Machine Learning Techniques
Professor Arun Rajkumar
Department of Computer Science and Engineering
Indian Institute of Technology Madras
Convex functions and Jensen's Inequality

(Refer Slide Time: 0:10)



So, it is a quick detour, a very quick high level primer about convex functions. Convexity is a fascinating topic, it has its own, you can do a whole course on convexity, but what we are going to do is only look at the bare essentials, which are necessary for our purposes.

So, what is convexity? Well, can we go first let me put down a picture and say what is a convex function, we want to understand certain types of class of functions which are very popular in machine learning in general, which have a lot of practical applications and one of that set of functions is the convex function. Convex functions have this property, so let us it the picture that one can think of is, something like this, where let us say you have two points, a and a point we will take a different point may be b . And the corresponding values that this function gives or let us say f (a) and f (b) .

Now, what I want to look at is some point here in between, which has value (a + b)/ 2. And I want to understand how does the f ((a+b) / 2) look like? So, it is the functions value is here. So, now, this is one value of interest.

The other value we look at is you imagine as if the function at this point from a to b was not that convex or not that function that we are looking at, but then like a straight line, look at a linear interpolation of this. And then that will give you some value here.

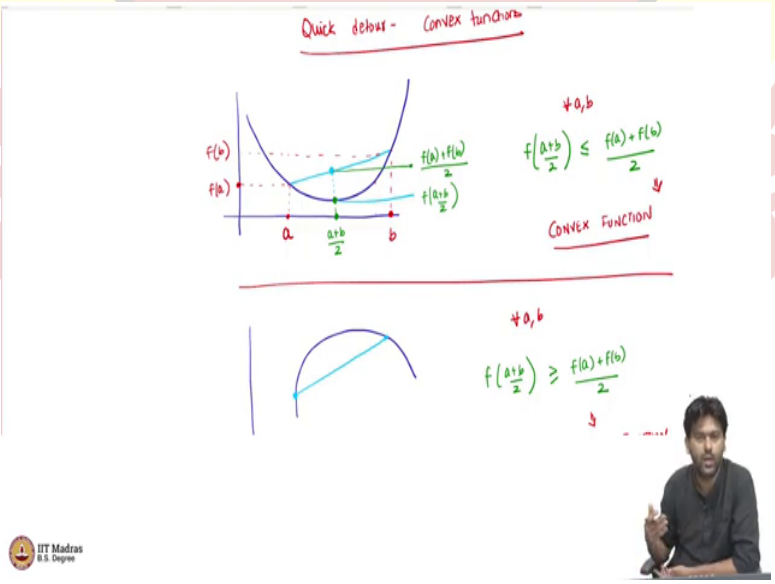
So, that at the point (a + b)/ 2, you get a value for this modified function, which is like interpolating between a and b linearly. So, now, what is well, this is, this value we know is f of let me write it in green f((a + b)/ 2).

What is this value? Can you can guess what this value would be? Well, this is linearly interpolating between f (a) and f(b) , so this is sitting exactly bang in the middle of f(a) and f(b) . So, this is going to be f(a)+ f(b) /2, halfway from f(a) to f(b) / 2.

Now, from this picture, we see that f((a+ b) / 2) ≤ (f(a)+ f(b)) /2. Now, if this happens for every a , b , that the linear interpolation at two points has a strictly higher value than the function itself, then such a function is called a convex function.

If this happens, then it implies that this function is a convex function, it should happen for every every choice of a and b , I have just shown two choices of a and b now, try to convince yourself that for this curve that you have that I have drawn, you can take any two points, and then this property will hold.

(Refer Slide Time: 03:48)

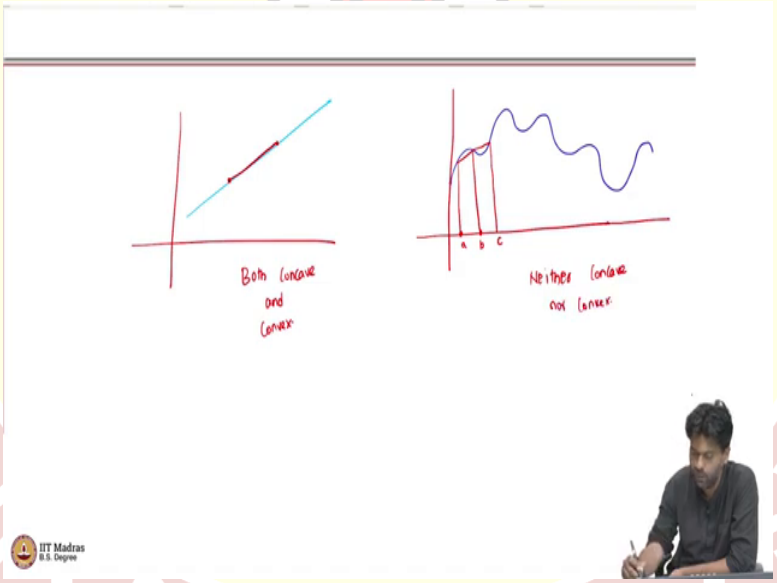


If, if the other side holds, so, if the other way holds that if you have a function like this, which goes the other way, where the linear interpolation has value strictly less than the functions value, then such a function is called a concave function.

So, it is our usual convex mirrors and concave mirrors, if you remember from your high school physics. So, that is the idea. So, for all a , b , what should happen in a concave function is $f((a + b)/ 2) \geq (f(a)+ f(b) / 2)$. If this happens, then we will call this a concave function.

Now from this, an immediate question is are there functions which are both concave convex and concave? Well, if you look at the definition, it says less than or equal to and greater than or equal to, which means that if there is a function where the inequalities were actually equalities, then It means that it is both convex and concave.

(Refer Slide Time: 04:58)



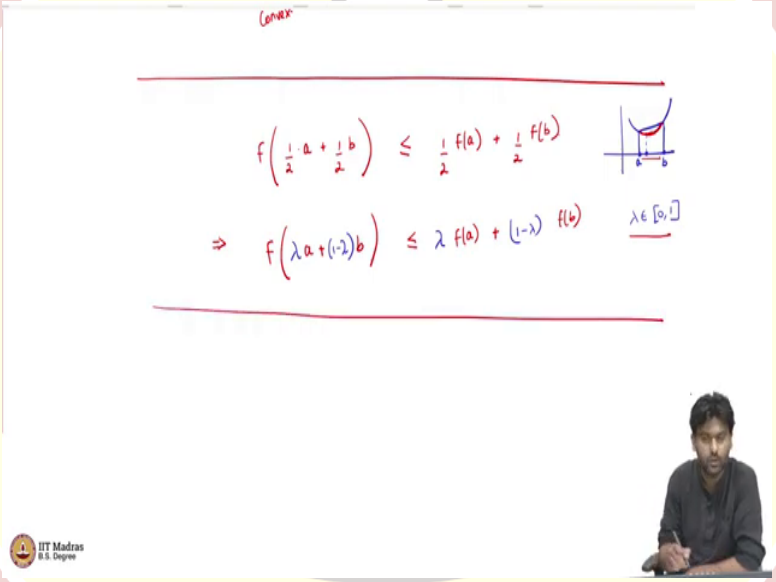
But what does it mean to say that the inequality is equality, it means that $f((a + b)/ 2) = (f(a)+ f(b) /2)$ for all a,b. And what function satisfies that? Well, that means that the function is linear.

So, the input, you divide by 2, now the output also get added and divided by 2. So, it means it is a, that is the property of a linear function. So, if you have a linear function, well, then if you have a function like this, then I take any 2 points. And then if I linearly interpolate, well, of course, I am gonna get the same value. So, this is both concave, and convex.

The next question is, are there functions which are neither concave nor convex? Well, of course, there are functions which are neither concave or convex, can you think of shape, such a function will help? Well, I am just giving you some shape here for a function, maybe a function like this.

So, it is neither concave or convex, because I can choose 2 points, maybe I will choose 2 points here, where the linear interpolation is below the function value, so it can not be convex. And I will choose maybe a and b, and then I will choose a, b and c, where the linear interpolation is above the function value between b and c. So, it can not be convex also. So, it is neither concave or convex.

(Refer Slide Time: 06:43)



So, what we are interested in is convex or concave functions. And one way to think of this is as follows the definition that I wrote down, I am going to write it in a slightly different way, I wrote (a+ b) / 2 functions value evaluated at (a+ b) / 2, I can write it as a/2+ b/2 . So, now if it is convex, then this is at most, f(a)/2+ f(b)/2.

Now, it happens that if this holds for every a and b, you can actually argue that this happens for any λ, $f(\lambda a+(1 - \lambda)b) \leq \lambda f(a)+ (1 - \lambda) f(b)$, where λ is any value between 0 and 1. So, if the first if you can argue that a function satisfies the first row first inequality, it also should satisfy the second inequality in pictures, this just says that, well, I drew the picture earlier.

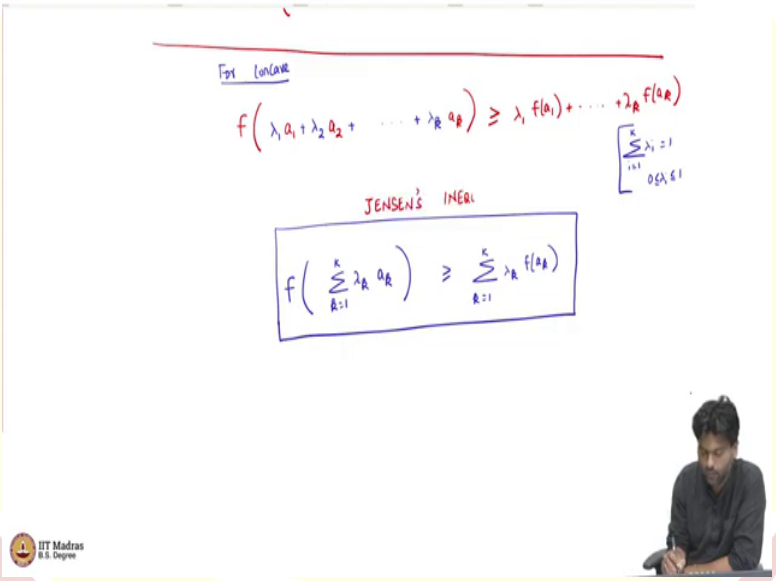
So, you have this picture like this, you have a and b. And now we are saying, you can also alternatively, think of the definition of a convex function, as you know, I can take any point in

the line segment joining a and b, in the interval joining a and b, maybe this point here, that point will also be lower than this line.

So, every point here in this region, so every value in that region is lower than the line itself. So, the line segment joining a and b, f(a) and f(b). And so you can generally use any λ the as you vary λ, you are traveling from a to b, if λ is 0.

So, then this is b, if λ is 1, it is a and as you change λ, you move from a to b , and then the functions value is always strictly less, I mean less than or equal to the linear interpolation. So, of course, concavity also has a similar property, I would not write it, but then you just reverse the inequality.

(Refer Slide Time: 09:07)

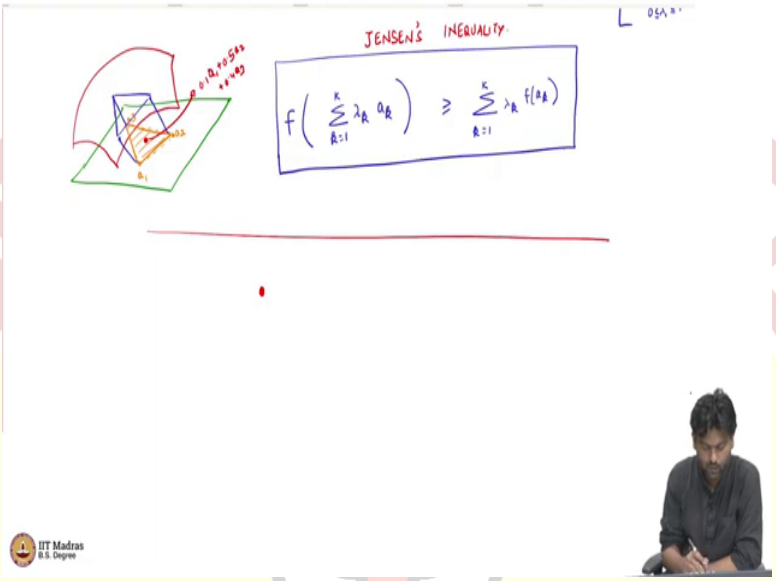


Now, you can extend this to multiple points as well. So, this once you have this λ now, it also is true. So, if I have not just 2 points a and b, if I have a1, a2, …, ak, now I ask λ1 a1+ λ2 a2+,..., λk ak, where the λs will sum to 1, they are between 0 and 1.

Now for concave functions, I mean, I am writing it for concave but it is true for convex also the other way around and that is implied. So, this will be greater than or equal to λ1 f(a1)+, …,+ λk f(ak). Of course we will assume that $f\sum_{i=1}^{k}$ λi = 1, i = 1 to k and 0 < λi ≤1.

So, essentially what in a slightly compact notation for concave functions, we are saying that $f(\sum_{k=1}^{K} \lambda k\ ak) \geq \sum_{k=1}^{K} \lambda k\ f(ak)$. So, this is just a generalized definition. So, this is what is called as typically should be called as the Jensen's inequality, Jensen's inequality.

(Refer Slide Time: 10:39)



One way to think about this is that, let us say you are I mean in if the function is from real to real, this does not really add any great intuition. But then if you if you imagine a function from, let us say, from are 2 dimensional plane, and if you have a concave function, this is kind of telling you that your concave function will somehow curve like this. And if you take any 3 points, it can take any number of points.

Let me explain using some a1, a2 and a3. Now, if you use λ1, λ2, λ3, and then combine these as λ1, a1+ λ2, a2+ λ3, a3, then you are going to get some point in what is called as the convex hull of these 3 points.

So, all the points here in this region can be obtained as some using some λ, like how, as you vary λ in the original, 1 dimensional case, you moved from f(a) to f(b), seamlessly, or a to b seamlessly.

Now, here you can move around in this region, which is called as a convex hull of these 3 points by varying your λ1, λ2, λ3, and maybe there is a point here, which is 0.1 a1,+ 0.5 a2+ 0.4 a3. Of course, the coefficients should add to 1 they should be between 0 and 1, which is

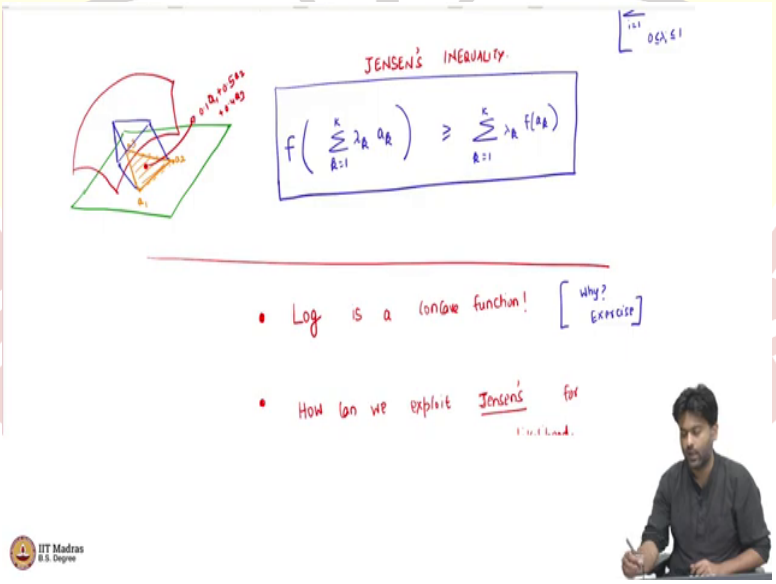
true here. I mean, the representation, I might not have gotten exactly the position correct. But then that is not the main point.

Now, what we are seeing is that, well, if I try to linearize, this curve, it is curving above the linearized version of this. So, if I look at the linearized version, at this point, maybe this is at this point, maybe this value is here, may be this value is here.

Now I look at the linear version of this curve, and the curve actually goes above this, this linear triangle in this case, that is what it means to say. I mean, that is what basically Jensen is saying, so it is saying that this happens, for any set of points you can look at, it is what is called as a convex hull, which is just the set of all points of the form sum over k λk, ak and then the function values about this.

So, the only thing we will need, for our purposes from convexity is this inequality that I put down here. And the reason why this inequality is useful for us, what is the connection to all this to log likelihood of Gaussian mixture model if you are wondering, the connection is the following.

(Refer Slide Time: 13:31)



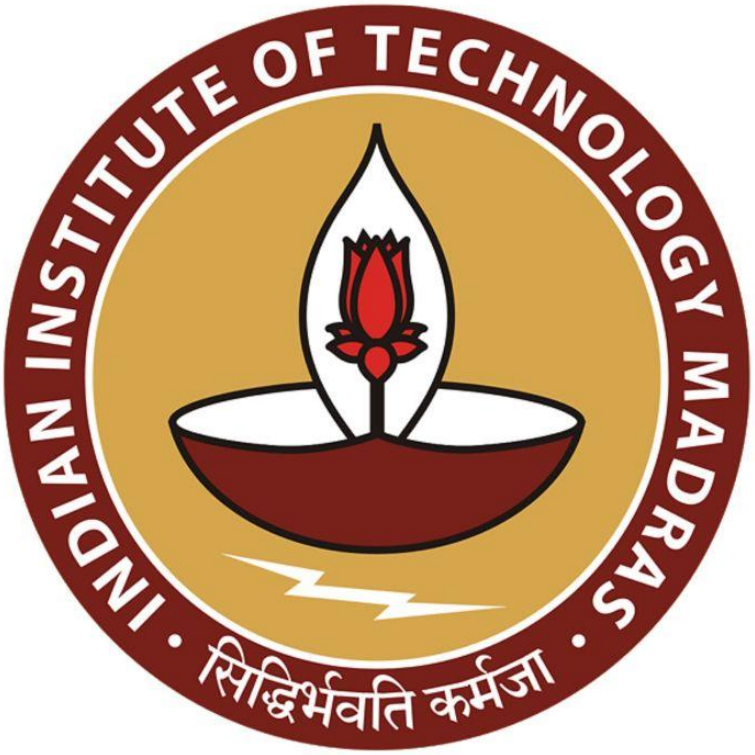
So, the connection is, logarithm which we are using in the likelihood function is a concave function. This is where the connection comes from, which means that it satisfies Jensen's so why is logarithm a concave function? So, take this as an exercise. So, we have put down the definition of concavity, function being concave.

This is one way to define convex concavity, they are equivalent definitions, but you just looking at this definition, can we prove can you prove that logarithm is a concave function take this as a quick exercise as a homework problem.

What we are now interested in let us assume logarithm is a concave function. That is the let us assume that fact. Now how can we exploit this basically, Jensen's, which is what we want to exploit for performing maximum likelihood and we will see why it might make sense to do this.

It should also already somehow hint as to what we are trying to do? We are thinking of a some inside logarithms and that was causing the problem. And now Jensen's kind of tells us that well, you can write it as combination of you know, can remove the sum inside the logarithm inside the function.

That is what Jensen's is telling us. Of course, at a cost, this is not an equality this is greater than or equal to so that has to be somehow dealt with. But somehow this is kind of making perhaps making life easier for us that is the hope and let us see how that actually happens.

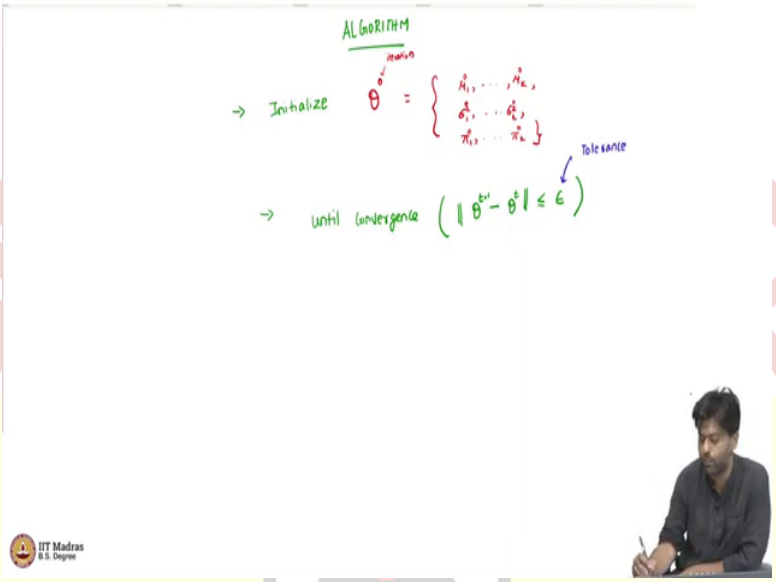


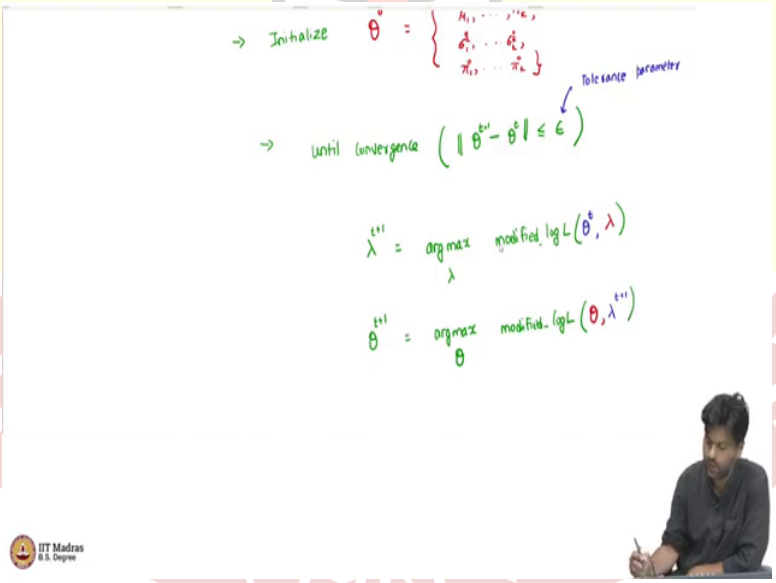
IIT Madras

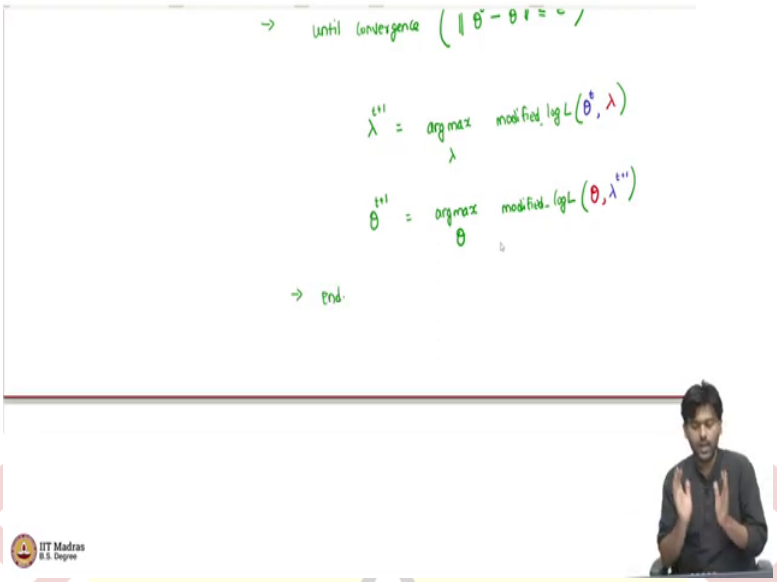
ONLINE DEGREE

Machine Learning Techniques
Professor Arun Rajkumar
Department of Computer Science and Engineering
Indian Institute of Technology Madras
EM Algorithm

(Refer Slide Time: 00:14)







And here is the Algorithm. So, the first thing you do is you initialize some value for θ naught. So, 0 here is iteration basically, it is an iterative algorithm. So, initially, you basically that means that you are initializing some means μ, μ^0 to $\mu^0 k$, some variances, $\sigma 1^2$ to σk^2, and then some π1 to πk.

Now, the algorithm thus as follows, now, until convergence and the way we are going to think of convergence here is that well our parameters θ^{t+1} - θt, so the norm difference is not too much. So, this is some tolerance parameter that we are allowed to tolerate. If your parameter estimates do not change too much, then you stop the algorithm.

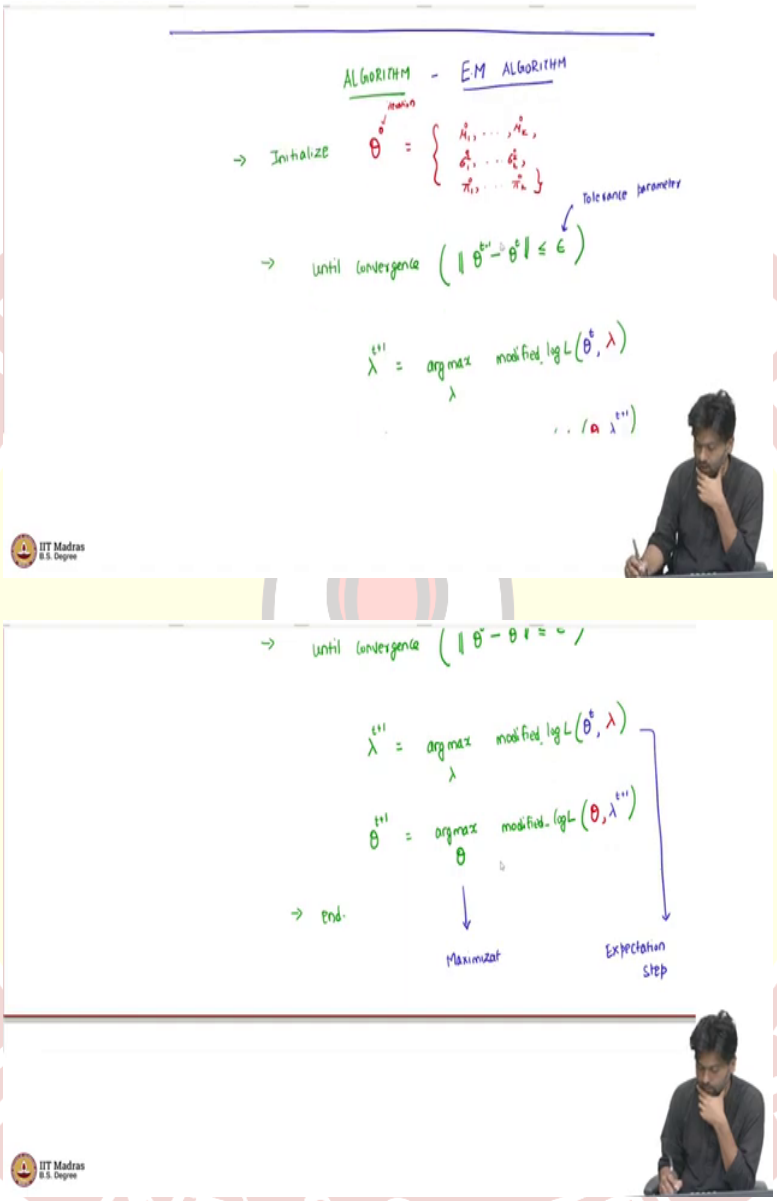
So, this is some tolerance parameter, it could be 10^{-3} or 10^{-2} depending on what you want to run. Now, what we will do is, because we have initialized with some θ, we will solve first for λ, λ^{t+1} is just arg max $_\lambda$ modified log L(θt , λ) and treating λ as the parameter.

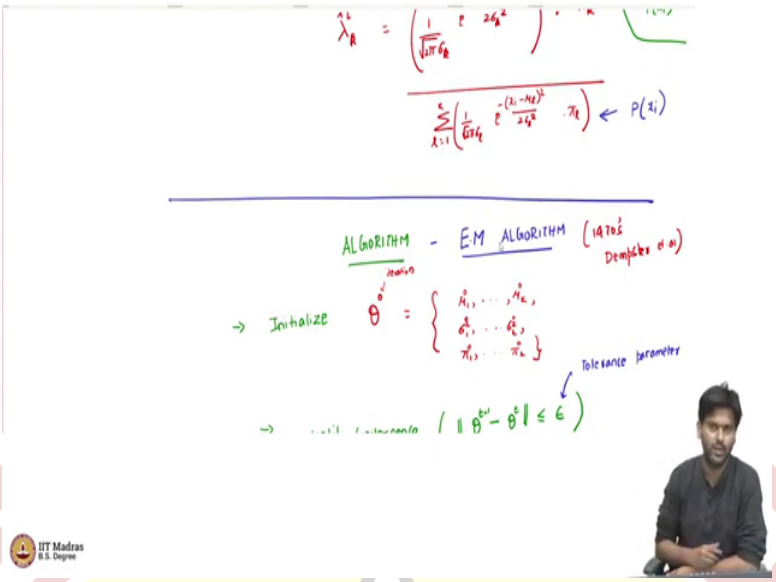
So, you are fixing θt and treating λ as the parameter and then maximizing, we know how to do that. That is a simple problem that we already solved. Now, the second step is once you have λs, you update your θ's, θ^{t+1}, again, using the simple formulas that we have put down, which is arg max over θ, the modified log likelihood. Here, you are going to, treat this as a parameter optimization over θ, where you are fixing λ to be the 1 that you got in the previous one, λ t plus 1. And that is it. That is the algorithm.

So, the key insight is that by introducing this new parameters and using the power of Jensen's you can split the problem into 2 parts, where fixing one solving for the other is easy, fixing

the other solving for other one is easy, and you will now do this, iteratively keep going back and forth.

(Refer Slide Time: 02:55)



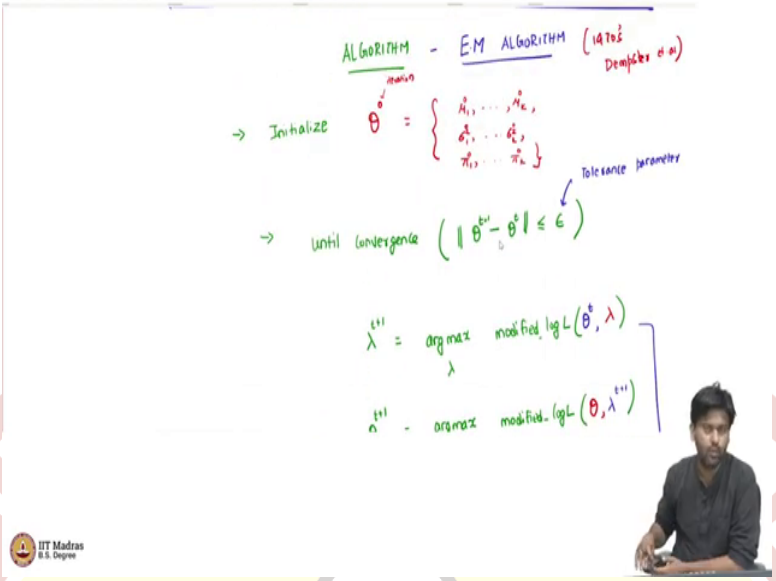


Now, this algorithm has a name. So, this is a, this is an instance of a very famous algorithm called the EM algorithm, where E stands for expectation and M stands for maximization. And basically, it has two steps. And the first step this is called as the expectation step. And this step is called as the maximization step. It is called an expectation step, because you can write this modified maximum likelihood as some kind of an expectation of quantity.

And that is what eventually we end up getting as λ^{t+1}, you can express λ^{t+1} as some kind of an expectation. And so it is called the expectation step. We would not worry about writing the generalized version in this course, but you can solve this in general also, I will make a comment later, but there are only 2 steps and because these 2 steps do this E and M steps alternatively until convergence, this algorithm is called the EM algorithm.

This was developed in the 1970s and still prevalently used I think it it was known in different avatars even before this, but then it was Dempster in his, several paper, put down this algorithm and called it the EM algorithm. And it has been, popular ever since.

(Refer Slide Time: 04:33)



So, all this is fine. But, and we can also argue that this algorithm will converge. So, all this is fine. But how can we understand this algorithm, what is exactly going on in this algorithm. We know the equations are easy to solve and all that, but intuitively. What is it? What is it essentially trying to do? The first thing to understand is that you can somehow try to maybe I should put this here on you so, you can try to relate this to our K-means or the Lloyd's algorithm. You can think of EM as if it is producing soft clustering. Where as Lloyd's produces hard clustering.

In other words, you can interpret lamda ik that comes out of this as the chance that every data point goes to a particular cluster, λ_k^i's is a chance that ith data point goes to the kth cluster, which means that the end of this algorithm, you are going to be not just left with parameters μ's, σ^2's and π's, you are also going to be left with some optimal λs and now, you can use these λs you can interpret these λs as some kind of clustering use this to clustering.

So, remember Lloyd's also had 2 steps, in the first step you would compute the means, and the second step you would do a reassignment you can even try to interpret these 2 steps as analogous to those steps. So, here we are not just computing means. So, once an assignment is given fixed, then what we are computing is well, when I say assignment is fixed here, it means that λs are fixed, we are maximizing over not just means, but then means variances and π's.

So, there are more parameters that we are maximizing over nevertheless, it can be treated as analogous to the finding the means in K-means, now, once the means are fixed in K Means we were doing a reassignment step, which means that we were changing the cluster indicators Zi's in a hard sense.

Now, here, it is done in a soft sense, in the sense that we are trying to see what is the how does the probabilities of points going to clusters change, once the means and variances and π's have changed. And that is your λ expectation step where the λs change. So, it is exactly the analogous algorithm to what we already have seen, but then in a more, full fledged probabilistic setting where you have more parameters.

So, it is not just the means you also have variances. And that is another point. So, EM also takes variances into account. Whereas, well, variances in the sense that in higher dimensions, especially when variances will become co-variances, your Gaussian will have a co-variance. And then you will estimate the covariance matrices, you might be able to estimate structures, which are slightly in a slightly better way than the Lloyd's algorithm.

For example, in 2 dimension, well, if you have 2 means here, Lloyd's kind of tries to assume that, there is a variance that you are trying to measure as the goodness of the cluster itself. But then it could be so that the data points in the first cluster might have a variance in a different direction, whereas the data points in the other cluster might have variants in a slightly different direction.

Now, Lloyd's will perhaps not be able to do so well in clustering points in this region, where these clusters overlap. Whereas, EM algorithm might be able to better understand that when there is a shape variance in a certain direction for one cluster, and a certain direction for other clusters simply because it is estimating the variance, along with the means, so there are more parameters you are estimating.

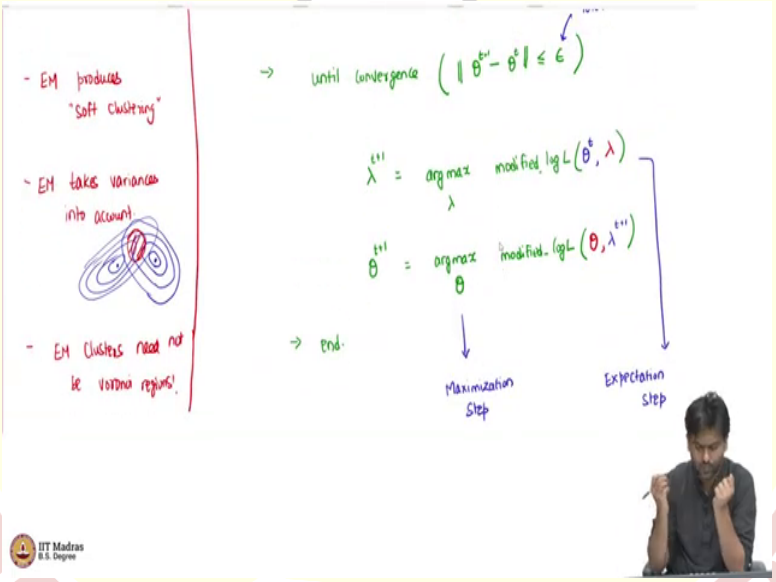
So, that is one advantage, you can think of this. And of course, if you want to do a clustering, using this, you can do a clustering. So, how would you do that? Well, you will start with your standard, you will run this algorithm, you will get your λs. Now, for every data point, you see what is the chance that this point belongs to every cluster.

And you if you want you can convert this into a hard clustering by assigning the data point to the cluster, which has the maximum chance of this point being, so for every i, you look at λ^1_k,

λ^2_k ,.., λ^i_k, and then see which of these is highest. And then you put your point in that particular cluster, that particular box, that way you can get hard clustering out of EM soft clustering that it produces, if you wish to cluster them in a hard sense.

And now this clusters need not necessarily be voronoi regions, so because you are calculating variances and so on, EM clusters need not be voronoi regions, so well, even if you do not have voronoi regions, especially places where clusters overlap. EM does a much better job of assigning points to clusters even if you do the hard clustering, than your perhaps your Lloyd's algorithm. So, these are some ways to understand the EM algorithm itself. So, this is one point I wanted to talk about. And we will talk about one more aspect of EM and then finish this discussion.

(Refer Slide Time: 10:06)

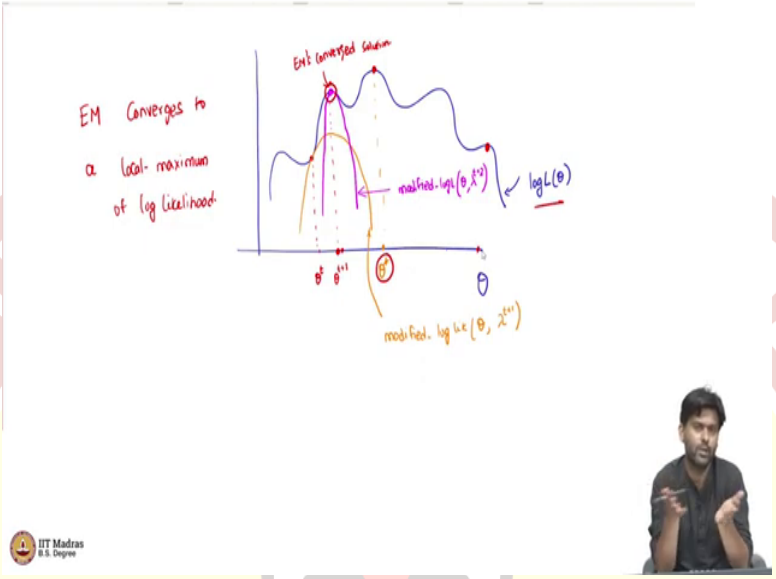


And that point is, well, all this is fine. So, we have put down this algorithm which has these nice two steps. And it resembles our K-means a Lloyd's algorithm and all that is fine. But how does this, compare to the log likelihood which we want to maximize at the first place.

Because we started with a parameteric probabilistic model, which had means variances and π's, we wrote down the log likelihood, and then we use Jensen's to completely avoid the log likelihood, but then use a different function, which is a modified log likelihood. And we are trying to solve for the modified log likelihood.

Now, of course, we are arguing that this algorithm will converge. We would not prove that, but then we can argue that this algorithm will converge and all that. But how does this relate to the original problem, which we wanted to solve, which was to maximize the log likelihood.

(Refer Slide Time: 11:04)



Now, if you think about that, the following picture, remember this? If I, loosely, I mean, try to explain this picture. So, let us say this is our parameter space, so, θ, which is the parameters over which we want to maximize the log likelihood. Now, if I drew, if it simply tried to plot the log likelihood go, it might look something like that. While for simple models, like Gaussian models, or Bernoulli, models, this likelihood would be a nice concave function, and then maximizing it would be easier.

In the Gaussian mixture model, the log likelihood is much more complicated function. So, maximizing this is hard. That is the biggest problem. So, in technical terms, we are saying there is a sum sitting inside log and all that but then end of the day, this is a complicated landscape where you are trying to find that θ* that maximizes this in fact, what you really want is this guy. So, you want this θ* where this likelihood function is maximized.

Now, what is exactly happening in in EM algorithm is the following. We are starting at some, let us say θt. So, we are in some θt, which has a certain likelihood value here. But now we are not working with the log likelihood, we are working with the modified log likelihood.

But then Jensen's is telling us that the modified log likelihood for any choice of λ is going to lower bound the log likelihood function. That is something that we already saw. Which

means, if I plotted the modified log likelihood at θt as a function of λ, then that might look something like this.

This is a nice function. That is the biggest advantage. This is a nice function. And it always stays below the original log likelihood for no matter what choice of λ. So, now, what I can do is what is this function? Well, this function is the modified log likelihood, θ and some λ.

So, I find λ^{t+1}, given a θ^t I found λ^{t+1}, the first step, which is the best λ^{t+1}, and then if I plot this as a function of θ, because I am plotting everything as a function of θ, now, it looks like a nice function. So, for any fixed value of λ^t, this will be a nice function, but then we are fixing that best value of λ^t, and then trying to see how this looks like. And now maximizing this is easier, which means that I can maximize this modified likelihood and get my θ^{t+1}.

Now, what would I what would happen? Well, the modified log likelihood, again, I will try to find the best λ^t, and then try to find the modified log likelihood of with respect to θ at λ^{t+2}, which might look something like this, maybe I will use a different color. So, maybe it will look something like this.

Now, at this point, what might happen is that maximizing this will make me converge and this function is just modified log likelihood of θ at λ^{t+2}. So, once I have found θ^{t+1}, I find the corresponding λ^{t+2}, and then I write this as a function, draw this as a function of θ and try to maximize this at this point.

And then I am kind of stuck here. So, what might be the solution that EM would give me is this, EM's converge, converged solution. Basically, you are trying to, go make better and better guesses in this complicated landscape. Now, you might, and in practice, you typically will, converge only to a local maxima of this, of the original likelihood function, that is a guarantee that you can do.

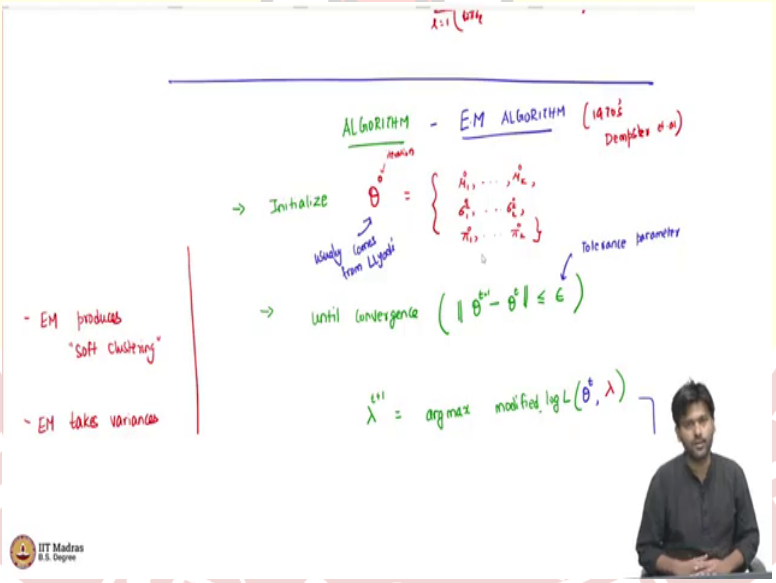
So, while you may not be able to maximize the original log likelihood function using the E M algorithm, in fact, you cannot do it with any known algorithm, what you can do is what you can guarantee using EM algorithm is EM algorithm. And this is the conclusion with which we will end this discussion is EM converges to a local maxima of log likelihood. While you really wanted to maximize the log likelihood function, which means that you wanted θ*, what you will eventually get is only a local maximum, you may not get here, it will converge to a local maxima.

So, in practice, it now kind of clearly says that how well you initialize this algorithm will kind of take you to which maximum you end up in. So, if I start with a very bad initialization here, maybe I will only reach this maximum. So, which might be much worse than the actual maximum that they want to reach. So, initialization will become an important thing.

And typically, what people do to initialize EM is that, while you are given a bunch of data points, you have a K, which is the number of mixtures, run your Lloyd's, get a hard clustering using Lloyd's, and use that hard clustering the means and the variances and the π's that you can derive from this hard clustering as your initialization for EM.

How do you get that well, once you have the hard clustering, you can look at each cluster compute the sample mean sample variance that will give you the π's μ's and the σ^2 for each of the cluster and then to get the π's you just look at the fraction of data points that are there in each of these clusters. So, that would be your θ^0.

(Refer Slide Time: 17:18)



So, this θ naught usually comes from Lloyd's. So, to summarize everything, what we are saying is, we have put down, a latent variable model, where we wanted to understand slightly complicated data, which has some cluster structure where the latent variable is a cluster indicator, and maximizing the log likelihood was a hard problem.

And so, we came up with a different algorithm, which exploits the structure of this log likelihood function using Jensen's inequality, to introduce new variables, such that you can do what is called as alternate maximization by fixing one set of variables maximizing rather and

the other way around. And if you want to initialize this new algorithm, which is called the EM algorithm, you start with the Lloyd's, get hot clustering, initialize it and then do this alternate maximization, typically in practice, you might get better estimates.

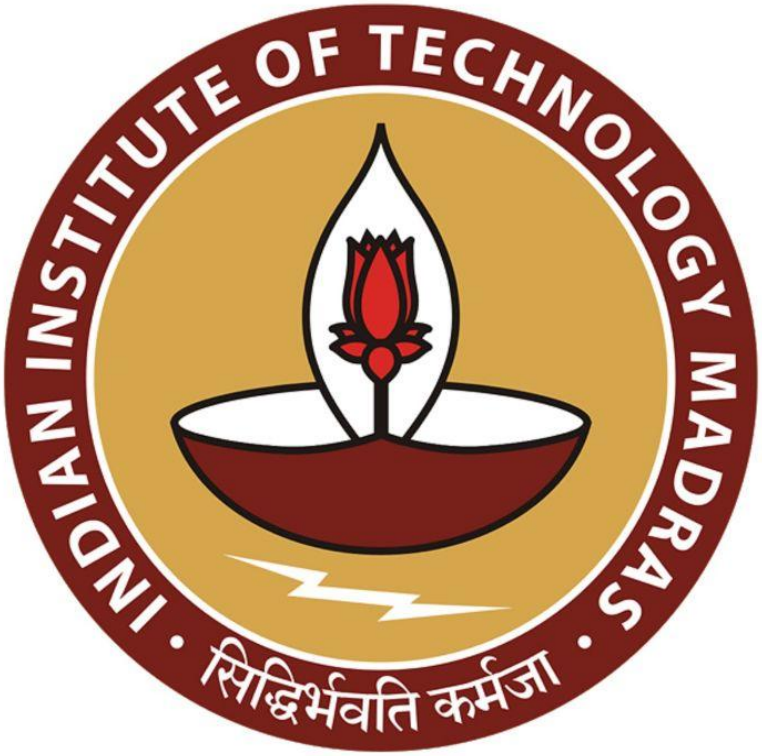
Now, you can do later again, use this to convert it into a clustering or you can use these estimates in whatever way you want to use. So, that is up to what is the task that you do after unsupervised learning. But for doing estimation, this is a very, very powerful technique, it converges really fast in practice, and typically produces very good parameter estimates.

So, this is whatever we have seen is in the context of Gaussian mixtures. But the general principle of EM algorithm is very broad. We will not do that in this course, but it can work for any reasonable latent variable model.

As long as you have some kind of nice structural separation that you have in the log likelihood, you can apply this EM algorithm, not just for the Gaussian model for the Gaussian model, we actually wrote down the algorithm, but this can be applied in a variety of other situations as well.

So, with this, we come to an end of our discussions about unsupervised learning problems, we looked at various methods of unsupervised learning, including representation learning, clustering, and now estimation.

And specifically in estimation, we have looked in detail, maximum likelihood Bayesian models. And now one very interesting practical application of maximum likelihood to the problem of estimating mixture models, specifically the Gaussian mixture model. From next time, we will start looking at different paradigm of machine learning, which is a very popular paradigm of machine learning, called supervised learning. Till then, take care goodbye and I will see you soon, thanks.

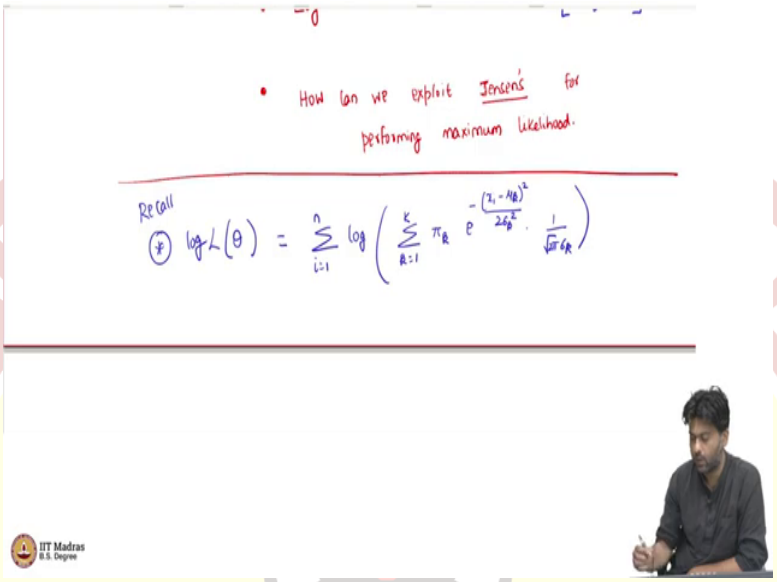


IIT Madras

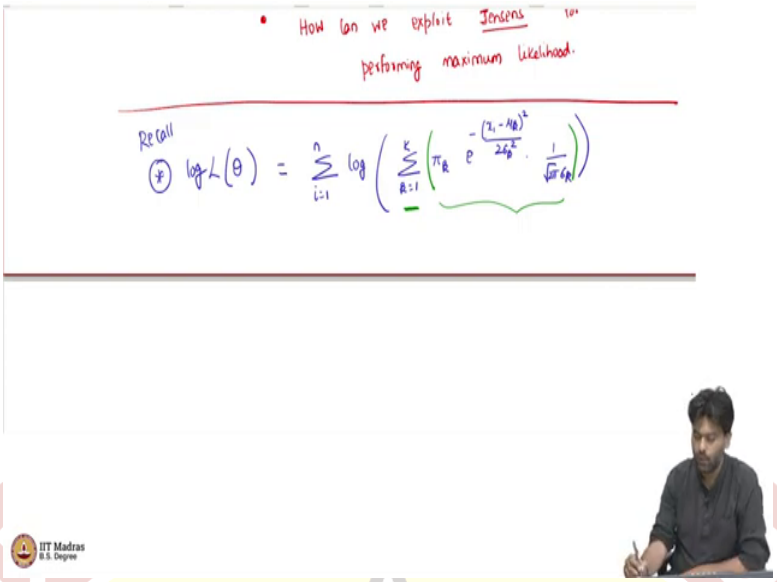
ONLINE DEGREE

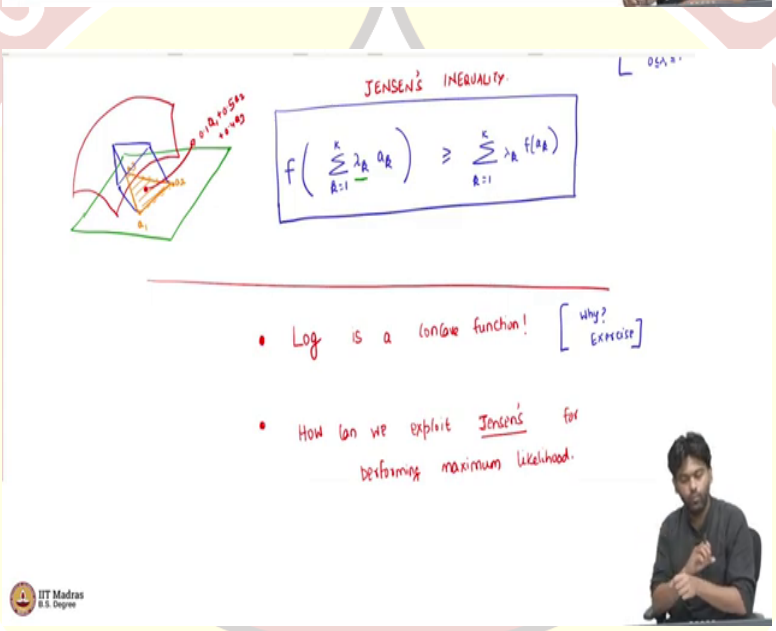
Machine Learning Techniques
Professor Arun Rajkumar
Department of Computer Science and Engineering
Indian Institute of Technology Madras
Estimating the parameters

(Refer Slide Time: 00:13)



So, let's let's try that. So, let's go back to our maximum likelihood function that we had written earlier. Let me recall that, recall this was I noted down this a star earlier. So, the likelihood function looked like this, the log likelihood function looks like this, $\sum_{i=1}^{n} \log(\sum_{k=1}^{K} \pi k$ $e^{-(xi - \mu k)^2 / 2\sigma k^2}$. $1 / \sqrt{2\pi}\, \sigma k$. This was a complicated looking log likelihood function.



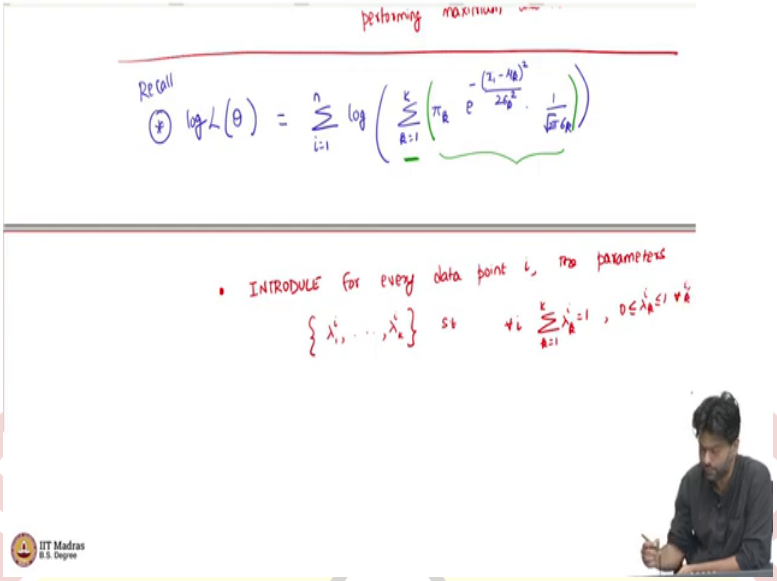


And the problem that we had was that there is a summation sitting inside the logarithm. Now, what we can do is, well, I want to think of this as a sum of a bunch of things. Now, Jensen's is telling us that well, it is not just a sum of a bunch of things, it is sum of a combination of a bunch of things that can be written in a form where you can remove the sum outside, pull the sum outside and then you will get an inequality.

So, we need some combination of the sums, so it is just a bunch of numbers sitting here, I am going to think of these as a bunch of numbers. And then there is a sum here. Now, if now I do not have this combination set.

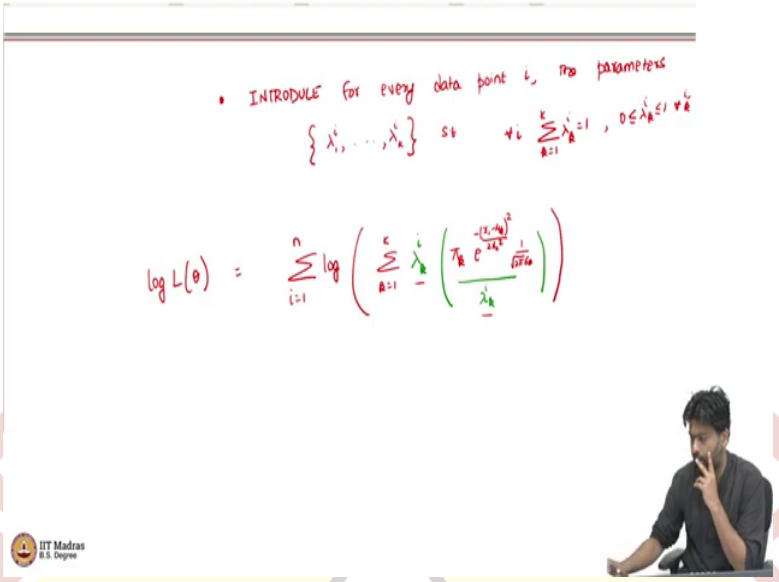
So, what I am going to do is, I am going to introduce this combinations artificially and make it a problem with more parameters. I mean, it looks counterintuitive at first glance, but then we will see the power of this method.

(Refer Slide Time: 01:57)



Now, what we are going to do is we are going to introduce for every data point i what we are going to do is we are going to introduce some parameters. And let us call them λ^i_1, to λ^i_k. So, every data point gets K μ parameters. And what are these well, such that, what should happen is for every data point i that the λs that we are introducing should be, should have this notion of probability. So, they have to sum to 1. So, $\sum_{k=1}^{K} \lambda^i_k$ should be 1, that is $0 \leq \lambda^i_k \leq 1$, for all k, all i and k, actually.

(Refer Slide Time: 02:53)

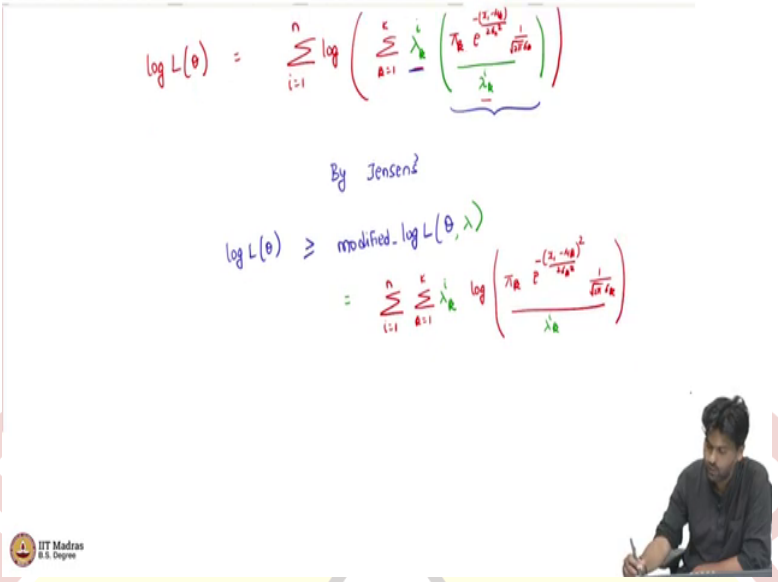


So, what is this? What does it mean to say introduce suddenly new parameters? Let us, see what that means. So, we have log L(θ).And this is $\sum_{i=1}^{n} \log \sum_{k=1}^{K}$. And this is where we want to introduce and then let me just bluntly, put this λ^i_k , which is the parameter that I am introducing, there is 1 for every data point i and for every data point i there are k of them, capital K of them, and the small kth parameter is λ^i_k .

Now I am just multiplying it, which means that the sum that I am thinking of cannot change, so then I am looking at a different function. So, to do that, what I am going to do is let me write down what the sum was earlier. So, the sum earlier was πk. $e^{-(xi - \mu k)^2 / 2\sigma k^2}$. $1/\sqrt{2\pi}\,\sigma k$, which is exactly the density of the Gaussian, but because I have multiplied it by λ^i_k , I will divided it by λ^i_k , now it is just multiplication and division.

So, it is still exactly the same log likelihood but then I am just artificially introducing these parameters I multiplying and dividing by parameter remember, for every i and every k there is a λ^i_k, what does that mean? That means that for every data point, now I am saying there is a distribution over all K's, all clusters will try to interpret these λ^i_ks later on for now in to think of them has some artificial parameters that I am introducing into the picture.

(Refer Slide Time: 04:32)



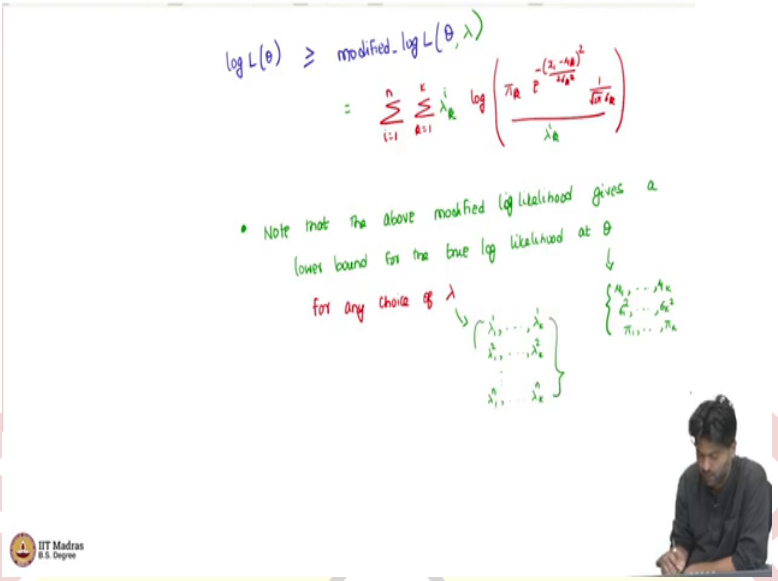
The moment I do this, now, I observed that, well this is a log of a sum of, a weighted combination of a bunch of points. So, the logarithm is evaluated at the sum of λ^i_k times a bunch of things. So, which means I can write use my power of Jensen's, so now by Jensen's I can write this log L(θ), which is exactly this quantity, as greater than a different function, it is not the same function.

So, because the function changes, and then that is why you have an inequality. And let me just call that as a modified log likelihood function, log likelihood of θ. Now this θ were the earlier parameters. Now you have this extra parameters λ, which have been introduced.

Now I am creating a new function, which is a modified log likelihood function, which not only has θ and then λs. But then what is this modified likelihood function? Well, it is basically the log likelihood function with Jensen's applied to it, the moment I apply Jensen's the summation comes out.

So, this becomes $\sum_{k=1}^{K} \lambda^i_k$, let me retain the color for λ^i_k. So, that we see that those are parameters we have introduced into the method. Now, you have log πk $e^{-(xi - \mu k)^2 / 2\sigma k^2}$. $1/\sqrt{2\pi}\,\sigma k / \lambda^i_k$.

(Refer Slide Time: 06:33)



The equations might look complicated, but then what is exactly happening is we had a log likelihood function, which was hard to maximize, but then we observed that there is a log of sum sitting inside it. Now we are introducing these extra λ parameters and then writing it as a sum of logs. And that is just by Jensen's by noting that log is a concave function. That is all has happened so far.

Now, why is this any easier? So, why should this be any easier to solve? First of all, I mean, is it easier to solve this model maximizing this modified logarithm log likelihood function, because it first of all, it is not the thing that you wanted to solve, so it is not the log likelihood function, it is a different function.

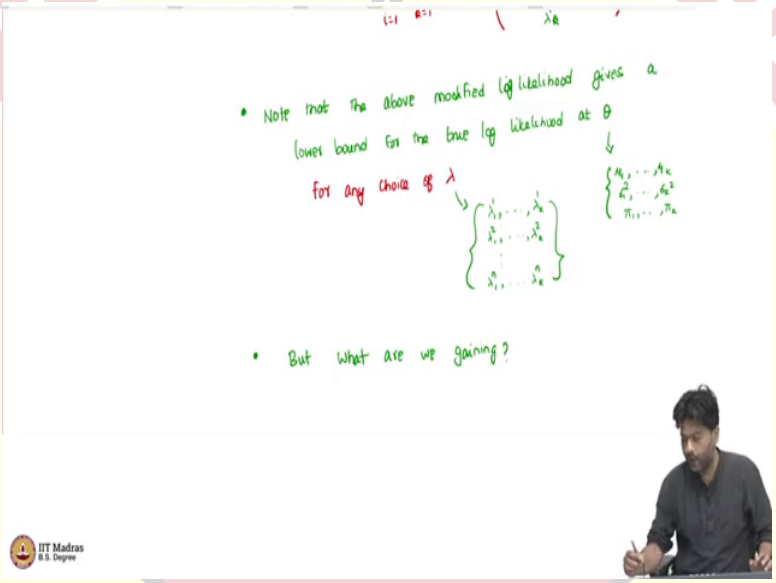
The second thing is that it has more parameters, we have introduced all these λ^i_k's for each i and each k, which means we have introducing n into k parameters essentially extra into the system into the equation. So, it should be worth it.

So, that something should be super simple, something become should become really simple. Otherwise doing all this is not worth it. Now, we will see why this is worth the effort. Now, first thing we note is that note that the above modified log likelihood gives a lower bound. For the true log likelihood at θ, I wanted to maximize the true log likelihood at θ, the function that I care about is to log likelihood at θ. Now if you give me a set of parameters θ, then I can evaluate the true likely log likelihood at θ.

But now if I evaluated the modified log likelihood, now it is going to give me a lower bound. That is what Jensen's tells me and this lower bound. And the interesting part is that this low this is a lower bound for any choice of λ.

So, Jensen's holds no matter what your λ is, as long as λs sum to 1 λs are between 0 and 1. So, for any choice of λ, when I say λ, when I say again, just to be clear, when I say θ, it means that I mean, μ1 to μk , $\sigma 1^2$ to σk^2, π1 to πk. When I say λ, I mean λ1 to λ1k, λ2k, to λ2k,...,λn1 to λnk. And that is what I mean by saying we are introducing this extra parameter λ.

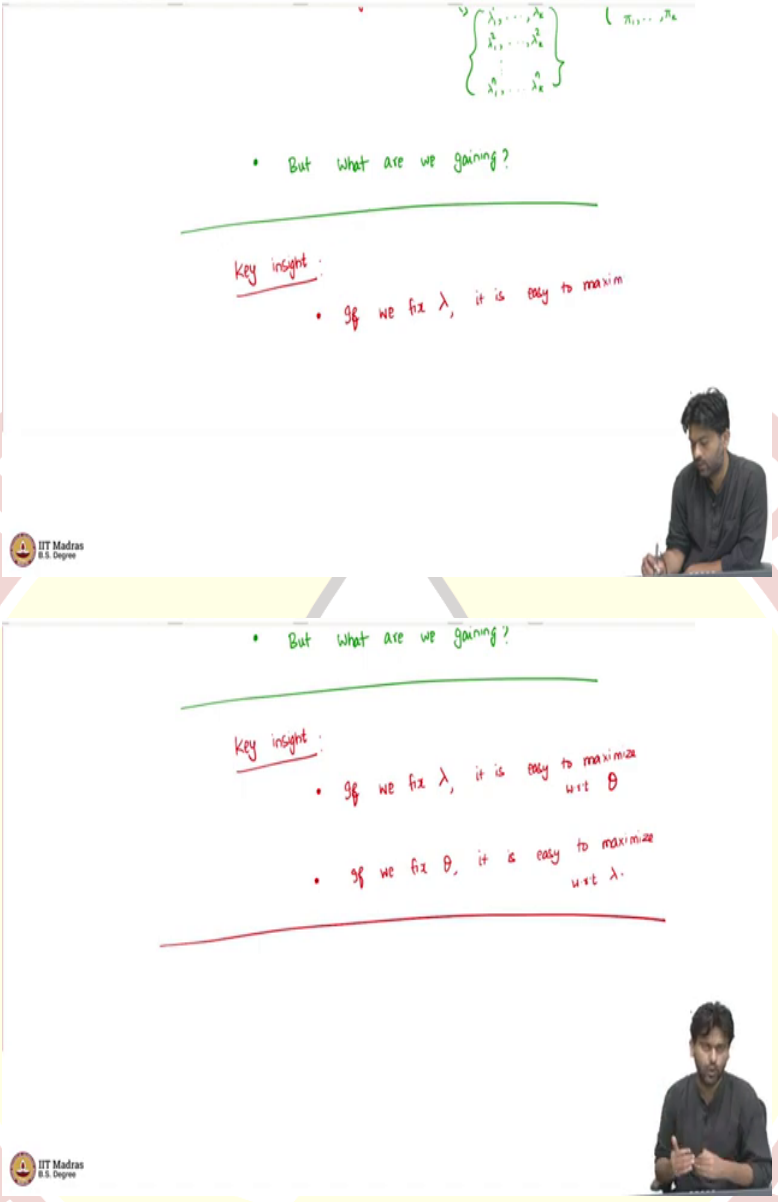
(Refer Slide Time: 09:26)



So, no matter so you give me a θ, which is a bunch of μ σ s and π 's. Now, I do not I mean, I do not want to evaluate the log likelihood. Instead, let me say I want to evaluate the modified log likelihood. Now I can put any value of λ, I will get a number and then give it to you. And that number will be a lower bound for the actual log likelihood at the θ that you gave me.

So, I will use your θ, I will use my own λs. And then I will compute the modified log likelihood it will be a lower bound. That is simply by Jensen's but the question is what are we gaining. So, this is a lower bound. So, all that is good, but what are we really gaining? By introducing this, we need to understand that.

(Refer Slide Time: 10:10)



And here is the key insight why this is such a beautiful method, so the key insight is the following. And that is why it works, we will see later. Now, I originally wanted to maximize my likelihood function with respect to θ. Now I have a new method new function, which is a modified log likelihood, which is a function of 2 different parameters, one is θ, and one is the artificial introduced parameter λ.

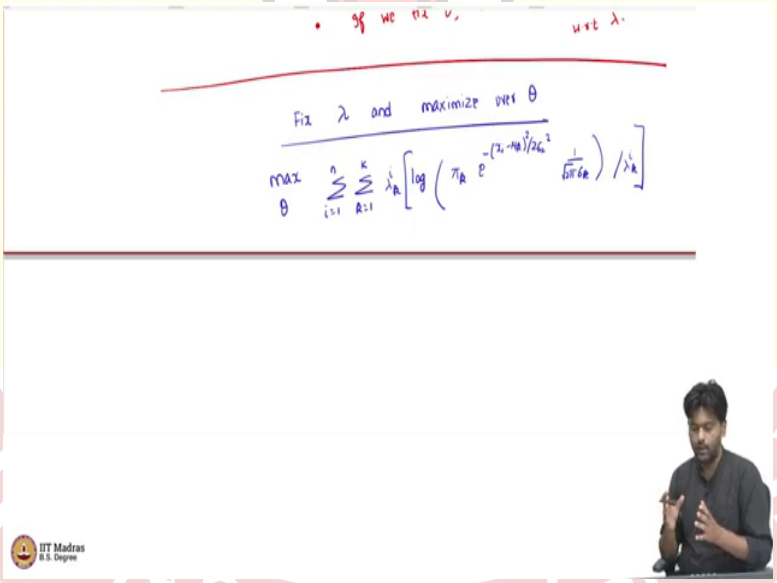
Now, the great advantage we gain by looking at this modified likelihood is the following. And we will justify this in a minute. Now, the advantage is that if we fix some λ, it is easy to maximize with respect to θ. So, for a given value of λ, it is super easy to maximize with respect to θ.

Similarly, we will see if we fix θ, it is easy to maximize with respect to λ. This is the perhaps the most important thing we should take away from here. So, now, originally we had a problem which had only one set of parameters θ, and we do not know how to maximize that in a nice efficient way.

Now, we are saying we are introducing another set of parameters. And now we are saying that if we fix some value for that parameter, I can maximize this modified likelihood with respect to θ very easily. And if we fixed θ, then we can maximize with respect to λ very easily.

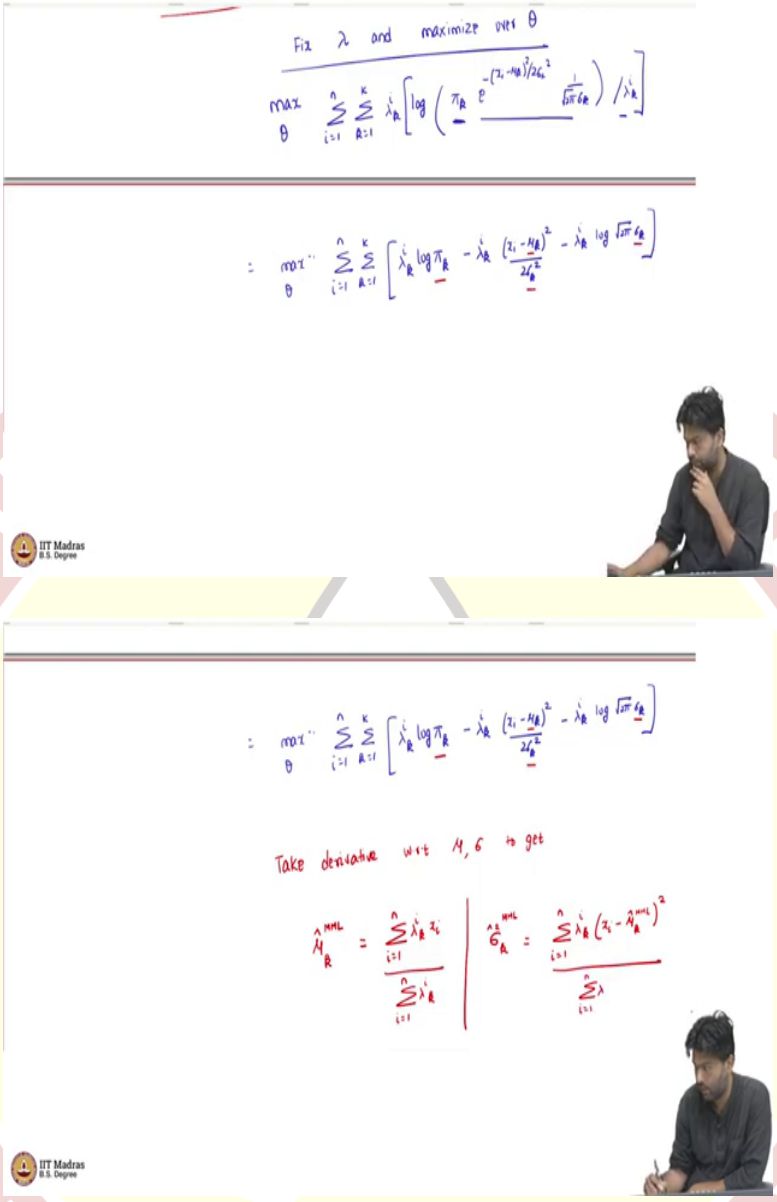
So, how, how is this useful to come up with an algorithm we will see in a bit, but then let us first convince ourselves that this key insight is true. And what does it mean to say it is easy to maximize with respect to θ and λ by fixing the other thing? And then we will actually put down an algorithm.

(Refer Slide Time: 12:13)



So, now what does it mean to say? We will fix λ and maximize over θ? Let us first do that. So, fix λ and maximize over θ. So, what is the actual function? Well, this is maximize over θ, when I say θ, again, the set of parameters that we are looking at $\sum_{i=1}^{n} \sum_{k=1}^{K} [\lambda_k^i \log (\pi k \ e^{-(xi - \mu k)^2 / 2\sigma k^2}. \ 1 / \sqrt{2\pi}\ \sigma k) / \lambda_k^i]$. Now, we are going to when you say we fix λ, we are going to treat λs as constants.

(Refer Slide Time: 13:08)



So, what does that help? How does that help us? Well, this guy is same as $\max_\theta \sum_{i=1}^{n} \sum_{k=1}^{K} \lambda_k^i \log$, Now, I can think of this as there is a product sitting inside the log.

So, that will become a sum over log. So, this will become log over πk. Now, the second term will, things will cancel out with respect to the E, so this will become $-\lambda_k^i (xi - \mu k)^2 / 2\ \sigma k^2 - \lambda_k^i \log \sqrt{2\pi}\sigma k$.

The next term I can ignored because this is $\lambda_k^i \log$, I mean $-\lambda_k^i \log \lambda_k^i$. But then this I am kind of treating it as a constant λs constant, because that is what saying fixed λ means. So, this is

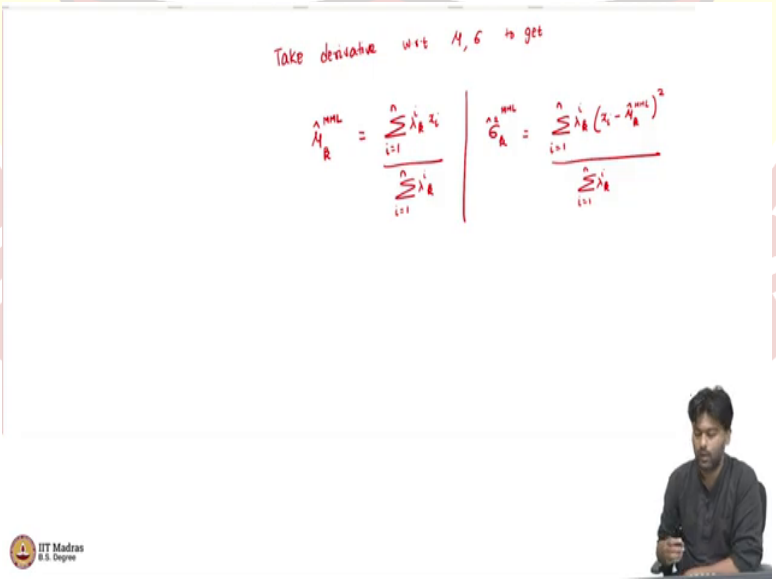
equal into maximizing only with respect to these 3 terms, because the real parameters here are πk,μ, σ , and so on. So, only these 3 terms matter.

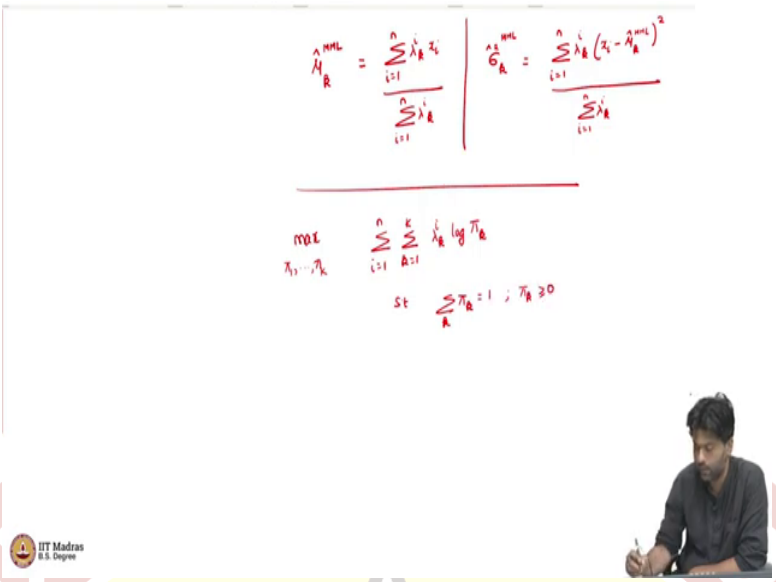
Now, the advantage of looking at this is that, if you now take the derivative of this function, where you are treating λ as constant with respect to the parameters θ, that is with respect to μ, σ^2, and π 's, all of them have a closed form. And let me put down those closed forms.

So, now if you can take derivatives, I would not do the derivation here. And please try that. And that is a very insightful exercise to try that take derivatives with respect to μ's, and σ to get the following. $\hat{\mu}$, I am gonna call this a modified maximum likelihood, it is not the mean maximum modified likelihood.

So, it is not, we are not maximizing the original likelihood, we are maximizing a modified function, the kth value is going to look like $\sum_{i=1}^{n} \lambda^i_k.xi \; / \; \sum_{i=1}^{n} \lambda^i_k$. We have a closed form for μ's. So, if you fix λ and try to maximize the modified likelihood function with respect to θ, similarly, $\hat{\sigma}^2_{k^{mml}}$, will just be $\sum_{i=1}^{n} \lambda^i_k.(xi - \hat{\mu}_{k^{mml}})^2 \; / \; \sum_{i=1}^{n} \lambda^i_k$.

(Refer Slide Time: 16:02)





We will talk about the π 's in a minute. But let us let us look at this and understand what this means. This means that there is some λs which I am fixing. So, I can arbitrarily fixed this λs. And then if for instance, if I had fixed λ^i_k. The way to think of λ^i_k is as follows. So, you can treat λ^i_k as the probability that the, ith point grows into the kth cluster, what do you think is the probability because we are fixing it arbitrarily at this point.

So, let us say I put each point into one cluster. So, I want to assume that every point comes from the same single cluster. So, which means λ^i_k will be 1 for a particular cluster indicator value k, and then it will be 0 every where else. If that happens, then what is this essentially telling us is $\hat{\mu}_k^{mml}$, is simply the mean of the data points assigned to a particular cluster. And the $\hat{\sigma}^2_k{}^{mml}$ is just the sample variance of data points assigned to a particular cluster if our λ^i_k's are 0 for all case, except 1, for which it is value is 1, but then we are not constrained to put λ^i_k's like that.

Now, then you can think of λ^i_k's as somehow, starting with a soft clustering of the data points, which means what do we mean by soft clustering? For every point, I am kind of telling, what is the chance that this point comes from each of the clusters? Now, we do not know earlier how to get this chance.

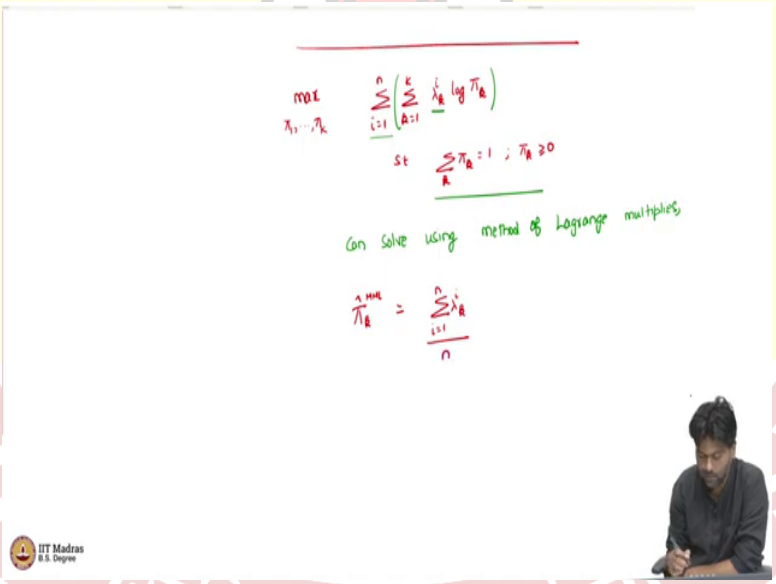
But then let us say we initialize it with some values, then what is these estimators are telling us is that, well, it is going to give you a weighted mean and weighted sample variance where the weights are given by these chances that we are fixing.

That is all this is. So, basically, once I fixed λs, then maximizing the modified likelihood is very easy. It is like saying, I am telling how important is every data point for each cluster. So, I am just going to simply, weigh these points by their importance for each cluster. And, and that will give me the clusters best mean and the clusters best variance. That is exactly in equations, these things that you are seeing.

Now you can, you have to be slightly careful when you are maximizing with respect to π's because π's have this constraint that they have to sum to 1, nevertheless, we can do the maximization you can do maximization of π1, to πK.

Now, if you again, go back to your likelihood and see which are the terms which depend on π , it will simply be some, there will be just only one term and that is also not too complicated term $\lambda_k^i \log \pi k$ such that π 's are not free variables such that $\sum_k \pi k = 1$, such that $\pi k \geq 0$.

(Refer Slide Time: 19:13)



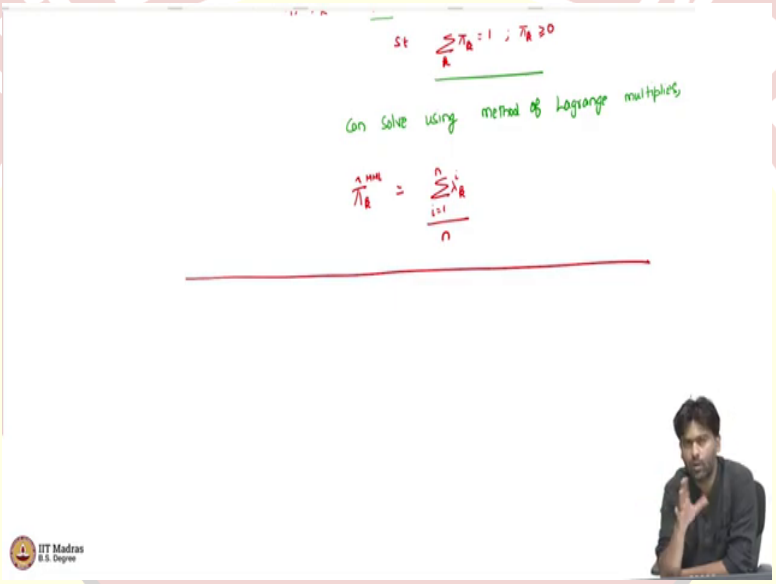
Now, you see that you observe that well, there is a λ_k^i for each i. For every i, we are deciding on a distribution over the data points, sorry distribution over the clusters? λ_k^i tells me what is the chance that the ith point comes from the kth cluster? Intuitively, that is what it is meant to mean.

So, now, here, this is a sum over i equals 1 to n and something which sum over K. Now, we are trying to maximize with respect to πk's. Now what we can do is we can solve I mean of

course, there is a constraint here, which is what will typically cause an issue, but you can solve this by using standard constrained optimization techniques.

If you have seen some constrained optimization techniques, you may be familiar with the method of Lagrange multipliers otherwise, take it at this point that this can be solved in closed form easily. So, this can be solved using the method of Lagrange multipliers to give us the following quantity $\hat{\pi}_k^{mml}$ is going to be simply $\sum_{i=1}^{n} \lambda_k^i / n$.

(Refer Slide Time: 20:34)



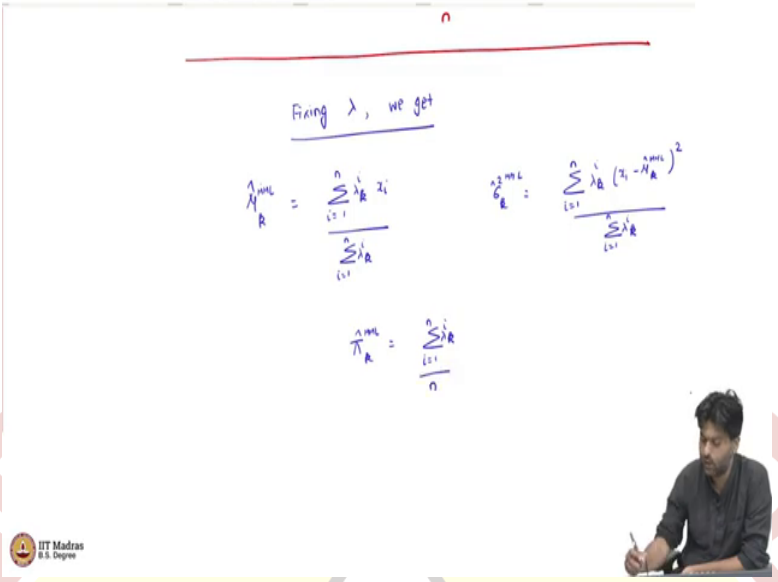
Now, what does this intuitively mean? Well, this intuitively means that well, remember λ_k^i kind of tells us what is the chance that we think ith point goes to the kth cluster? And now, we are asking what is the chance that a, some point will come from the kth cluster? Well, it is the, you are basically how to average the chance of each point going to the kth cluster.

Now, if each point was hard clustered that it will only go to one cluster, which means λ_k^i's where we are taking values either 1 or 0. Then this simply means that $\hat{\pi}$, our best estimate is just the average of the number of points or the fraction of points that went into that particular cluster. So, if λ_k^i's were 0's or 1's, then well, for each data point, you are counting λ_k^i.

So, then it will be 1 only for those points which have been assigned to cluster k. So, you are just looking at the fraction of points assigned to cluster k. But then if you do a soft clustering if λ_k^i's or between 0's and 1's, then this is kind of telling you on an average, what is the

chance that a point belongs to a cluster K, that is what this is. So, basically, putting everything together, so, what we have is that we have the following.

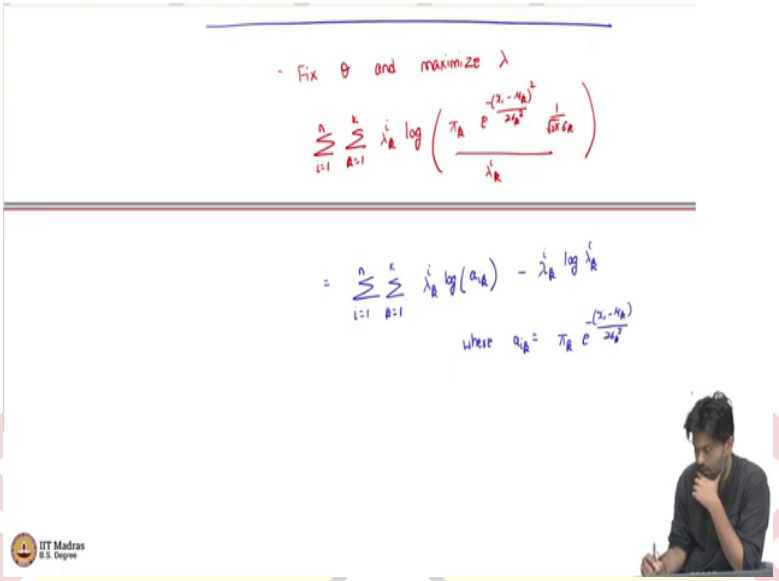
(Refer Slide Time: 21:53)



So, fixing λ we get the following if we maximize with respect to θ, we get the following. I will just summarize this as $\hat{\mu}_k^{mml}$ is just the weighted mean, where the weights are given by these λs that we are assuming $\lambda_{k.}^i$ $\hat{\sigma}^{2\,k}_{mml}$ is the weighted sample is the weighted sample variance.

But again, the weights are given by λ^i's that λ_k^i's that we are assuming $\sum_{k=1}^{K} \lambda_k^i$, and π 's are again, the weighted version of what you would standard expecting, if it is just the weighting of each points chance that belongs to cluster k and then averaged. So, this is good. So, this kind of tells us that Well, I have a problem with 2 sets of parameters θ and λ. I fixed λ. I can maximize easily with respect to θ.

(Refer Slide Time: 22:52)

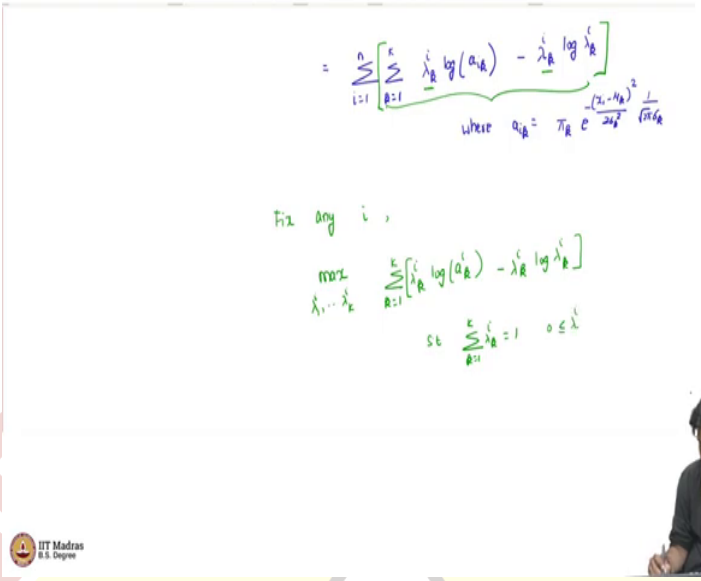


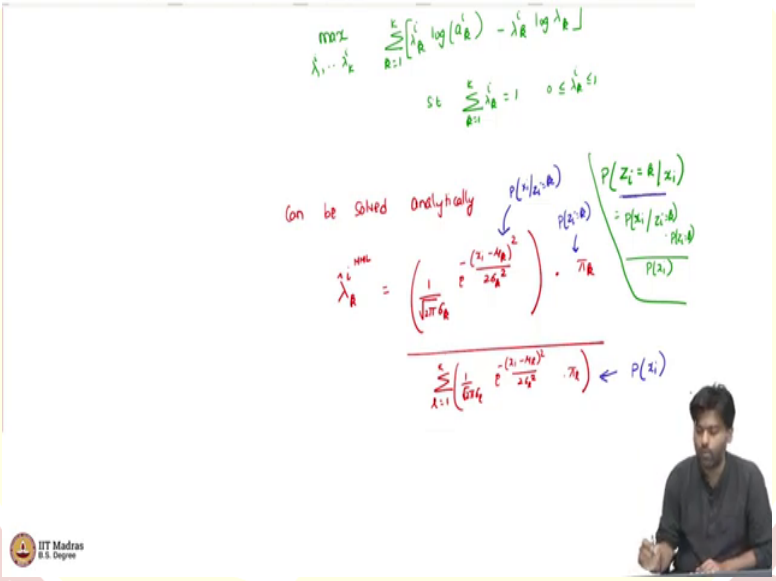
Now, the other way should also happen easily. So, you need to fix θ and maximize with respect to λ. And let us see if that is easy also. And once both of these we convince ourselves that these are easy by looking at the actual closed form solutions, then we can put down an algorithm that can be efficiently used to solve this problem.

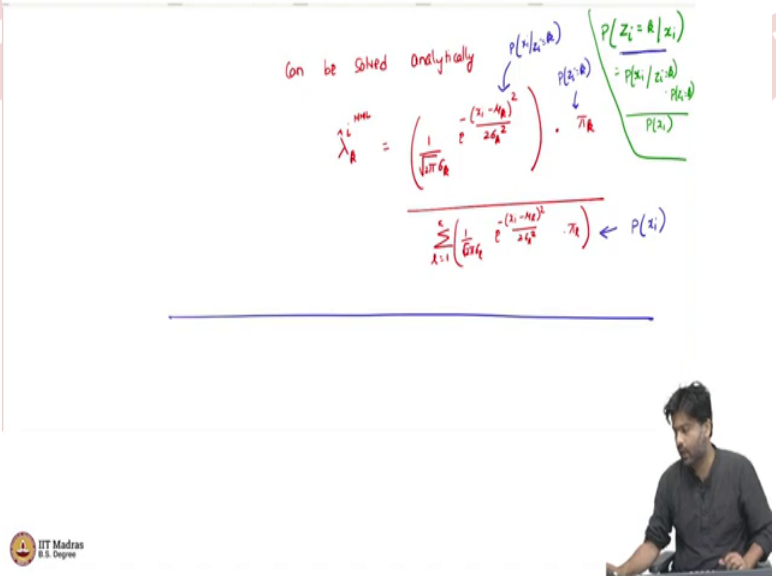
So, how would this look like? So, now, we are fixing θ and then maximizing with respect to λ, which means again, let me recall the likelihood function every time I would have to write this, but it is worth doing that because it will reinforce what we are trying to do in a better way, ik divided by λ^i_k this was the likelihood function. Now, I want to maximize this with respect to λ treating all the other parameters μ's, π 's, σ 's as constant.

Now, I can, I mean, do some simplification and then only pull out the terms which are non constants. And that will look like this λ^i_k. And you can try this this one step I am skipping, but then you can try this out for yourself. So, λ^i_k some constant and they will write what this constant is $-\lambda^i_k \log \lambda^i_k$, where this constant that I am thinking of is ik e power minus basically the density of Gaussian. So, 2 σk^2 1 / $\sqrt{2\pi}$ σk.

(Refer Slide Time: 25:03)







So, now remember, we want to maximize this with respect to λs. So, and then there are k λs for each data point i. Now, if you look at this, equation itself, it is a sum over data points, and then for each data point I have a bunch of K parameters. So, then I am adding these things up. So, because the parameter, so the function is not, does not have any cross terms, so, the λ^i_k and λ^j_k do not appear together at all in the function.

So, I can actually maximize these separately for each data point i, and then that would actually the separate maximize that should actually also, maximize the entire function because it is just a sum of a bunch of functions, which depend on i separately, but then there are no variables shared between these the inner summation.

So, I can optimize this separately, which means that we can fix any i and then maximize over $\lambda^1_k,..,\lambda^i_k$ $\sum_{k=1}^{K} \lambda^i_k \log (a^i_k) - \lambda^i_k \log (\lambda^i_k)$. Of course, λs cannot be arbitrary such that we know that λs have to satisfy $\sum_{k=1}^{K} \lambda^i_k = 1$ and $0 \leq \lambda^i_k \leq 1$, it is a constrained optimization problem, nevertheless, not a hard one to solve.

So, you can still maximize this in closed form using the method of Lagrange multiplier can be solved to enclose form analytically. When I say analytically, all I mean that can write down a formula for the answer to get the following formula $\hat{\lambda}^i_k$ of the modified maximum likelihood looks like the following.

So,$(1 /\sqrt{2\pi}\ \sigma k.\ e^{-(xi - \mu k)^2 / 2\sigma k^2}) . \pi k / \sum_{l=1}^{K} (1 /\sqrt{2\pi}\ \sigma l\ e^{-(xi - \mu l)^2 / 2\sigma l^2}) . \pi l$. So, this looks like a complicated formula, but it is not.

So, this is all this is saying is that, well, what is this. So, I am I am trying to ask, well, if I fixed the parameters, if I tell you what are the means, what are the variances what are the π 's? Then what is the, what is the best guess $\hat{\lambda}^i_k$?

What is $\hat{\lambda}^i_k$ representing it is representing the probability that the ith point goes to the kth cluster, I tell you what the ith point is and then ask what is the chance that this goes to the kth cluster. Now, well, if I know all the parameters, then basically what I am asking is, I am asking something like what is the probability that zi, which is remember from our step one that this is the cluster indicator is k given xi.

So, given a point xi, say I am asking what is the chance that this goes to the K cluster? And that is what I am going to think of as $\hat{\lambda}^{i}_{k}$, as to represent. Well, I know by Bayes theorem, this is just P(xi | zi = k) . P(zi = k) / P (xi).

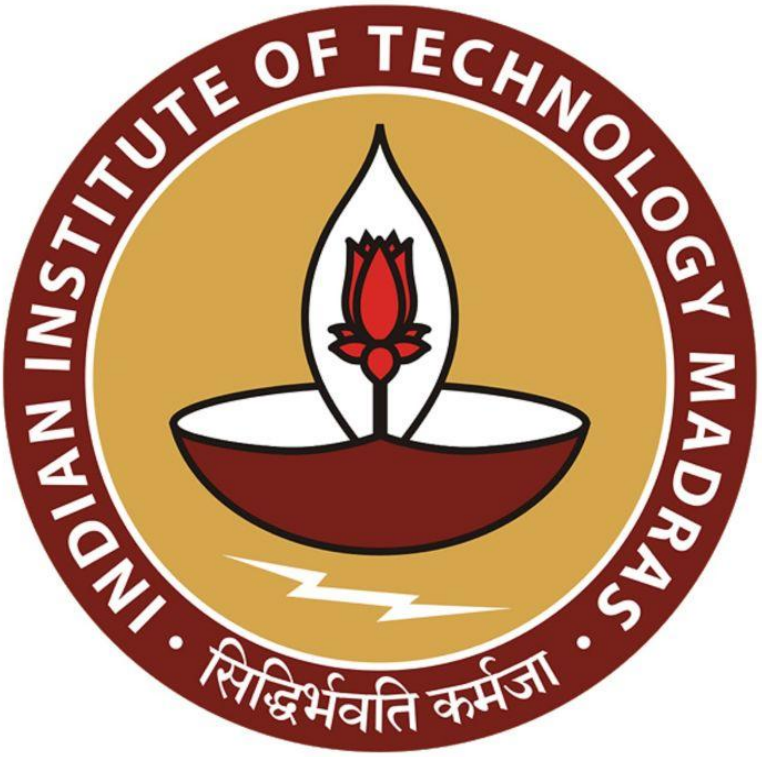
This is my base theorem. So, this is what my Bayes theorem tells me. And now if I have the parameters μ's, σ^2 and π 's, then you can verify that this is simply probability of xi given zi equals k, because this is the chance that xi comes from the kth cluster.

This is simply P (zi = k) this is the step 1 chance that the kth cluster is chosen. And this is summation over like all possible ways of generating xi. P(xi | zi) into P(zi) which is just P(xi) itself. This is just base theorem.

So, it is so, nice that this turns out to be exactly what you get, if you had applied Bayes theorem and try to estimate P (zi = k | xi), that is what is the chance that the xi point goes to the kth cluster? And that once the parameters are given, you will simply use the Bayes theorem to estimate that, that is exactly what comes out as the as your maximum likelihood estimator also.

So, now what we now have is we have solved for λs if the parameters are given similarly, we have solved for the parameters if the λ are given an all of these are just simple formulas. So, given once you fix λs you get a weighted mean, weighted variance and then a weighted fraction for μ, σ^2 and π respectively and if you fix μ, π, σ^2 respectively and then see what is the best λ then that is simply by your Bayes theorem then estimate for P (zi | xi).

So, all of these are simple now, so, all of these are closed form solution. So, which means we can actually write down an algorithm for this for solving for θ. So, remember our original goal was to maximize the log likelihood now, by adding these extra parameters λ we are saying that you can fix λ solve for θ you can fix θ solve for λ efficiently that we have seen how. Now we will use this to come up with an iterated algorithm to solve the problem.

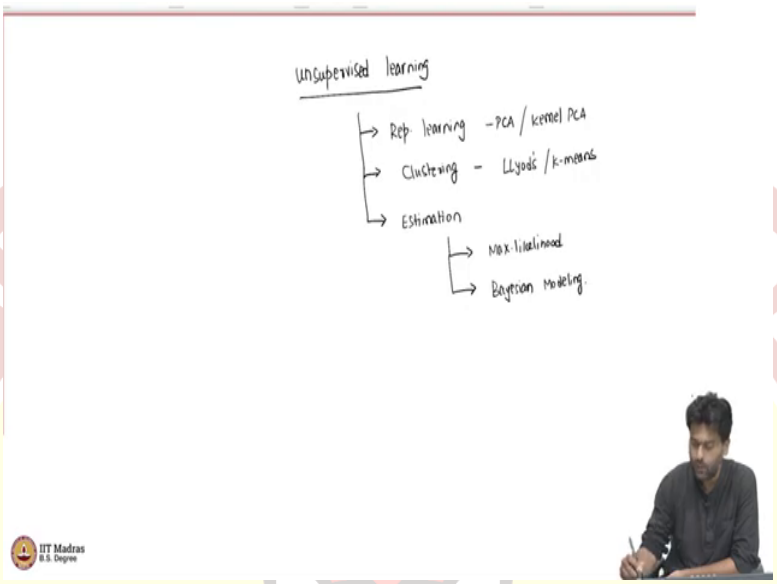


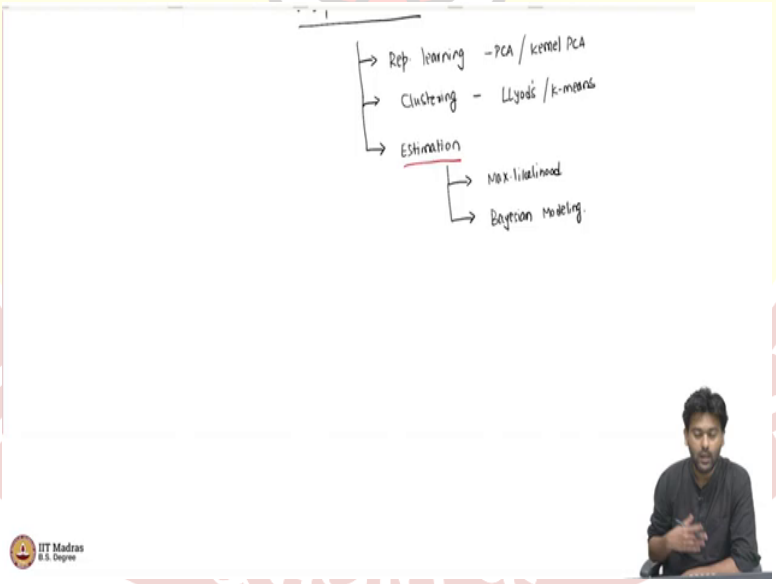
IIT Madras

ONLINE DEGREE

Machine Learning Techniques
Professor Arun RajKumar
Department of Computer Science and Engineering
Indian Institute of Technology Madras
Gaussian Mixture Models

(Refer Slide Time: 00:15)



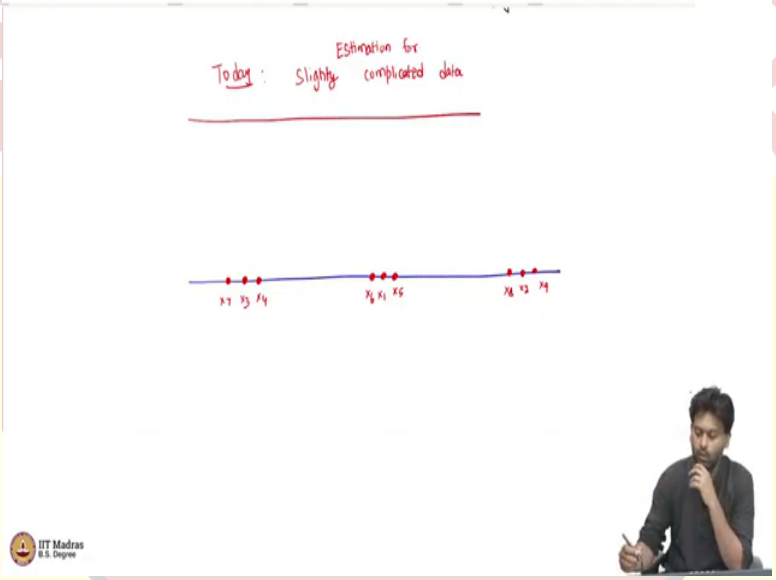


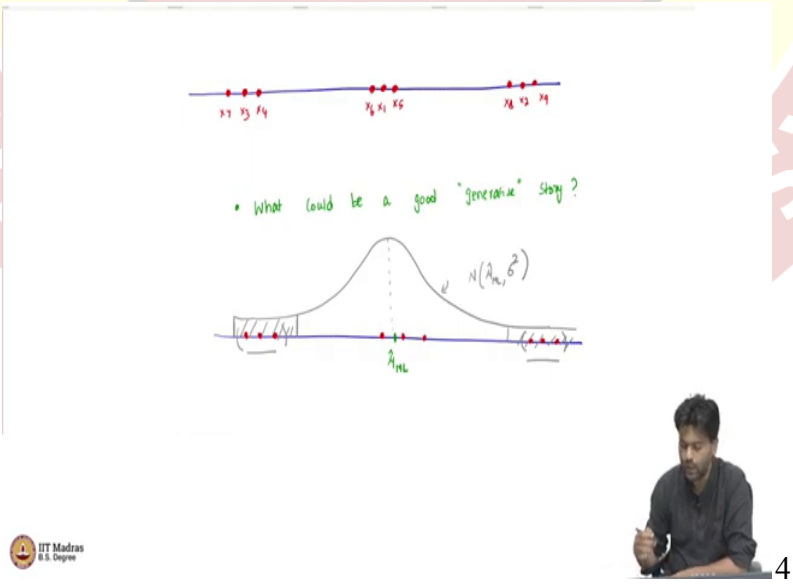
Welcome back, everybody. So, we have been looking at unsupervised learning so far in this course. One we will continue to look at unsupervised learning a little bit more. And just to recap, where we are in unsupervised learning, we looked at representation learning via the means of the PCA algorithm and also the kernel version of it, which we called as the kernel PCA. And then we also looked at clustering methods after that, which gave us the Lloyd's algorithm or the K-means algorithm.

And then, last time, we looked at estimation, as a probabilistic way of doing unsupervised learning. And in estimation, we looked at the maximum likelihood based ideas. And we also looked at a way to incorporate prior beliefs into our estimator, which is using the method of Bayesian modelling where you start with a prior and then you convert it into a posterior.

So, what we want to do today is continue our discussion about estimation in general. And in estimation, we want to look at a slightly more realistic type of data, which we would like to model in an unsupervised way.

(Refer Slide Time: 01:39)





So, today, the goal is to look at slightly complicated data. And we look at estimation for this. So, what is this slightly complicated data that I am talking about? Again, what we will do is for illustrative purposes, we look at one dimensional data and try to understand the ideas but whatever I am going to talk today will carry forward for higher dimensions. In fact, whatever we discussed for estimation and maximum likelihood, Bayesian, everything works for high dimensional data also, but it is easiest to explain using one dimensional data.

So, what is the data that we are talking about here? So, let us say we are on the real line, simple one dimensional data and we have a bunch of data points, let us say like this. And the way the numbers are, the way these data points are ordered, let us say this is x1, this is x2, maybe x3 here, x4, I am just arbitrarily labeling these x8, x9. Let us say this is the data that we have. Now, in the world of estimation, the goal is you have to come up with a model that explains the data. So, the way we are thinking about this is, we will come up with a generative story that explains our data.

So, the question that will first ask is what could be a good generative story for this data? So, for this data, you see this data and then you ask the question, what could be a good generative data, a good generative story? First thing is this data is not just zeros and ones. So, we cannot use like a Bernoulli type of a modelling for here for this. This data has a bunch of real numbers.

So, one immediate thing that you could do is you can try measuring or modelling this using a Gaussian distribution, like how we did last time. So, because we have a bunch of real numbers, we can always use a Gaussian distribution to model it. So, I mean, we could model it. So, it is good or bad is we will talk about it, but then there is nothing stopping us from modelling it. And then you can apply your estimation methods and so on.

So, let us say if you did that. So, and then let us say we did a principle of maximum likelihood to get the best Gaussian, which has the mean, which is best in the sense that it explains the data in the best possible way in terms of maximizing the likelihood that we know that the mean of the maximum likelihood for Gaussian estimation is just the sample mean, which means in this case, the sample mean may be somewhere here.

So, this may be our $\hat{\mu}_{ML}$. So, the sample mean is the, this the maximum likelihood estimator, if you assume a Gaussian model. So, now if I want to generate new data from this model that I have learned, which has sample mean, $\hat{\mu}_{ML}$, let us say the variance is known in this case and

the variance is some 1. Now, how would the density of the distribution look like with this particular sample mean that we have learned? Well, that is going to look something like this.

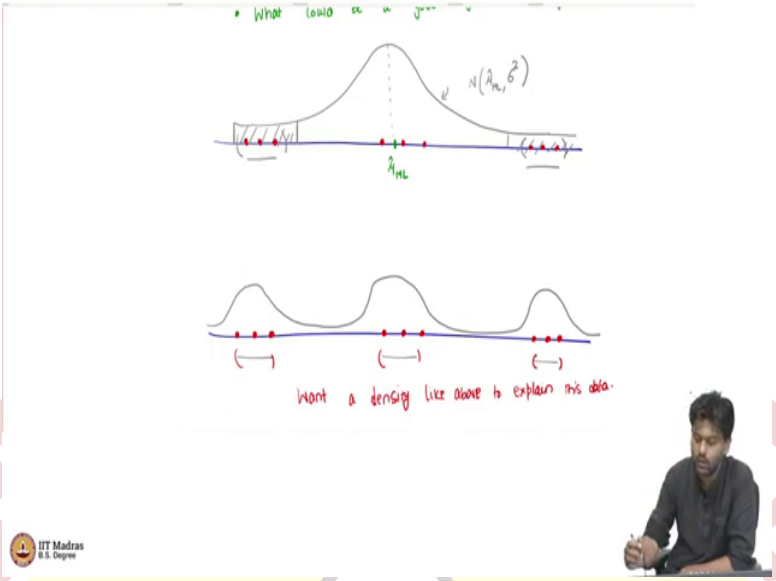
This is the Gaussian. Of course it the mean and the mode of the Gaussian match, it is going to be at the sample mean, because we are positing that the actual Gaussian that generated this data has to have a mean $\hat{\mu}_{ML}$. So, that is what we are estimating. And so if you look at the density, it is going to look like this.

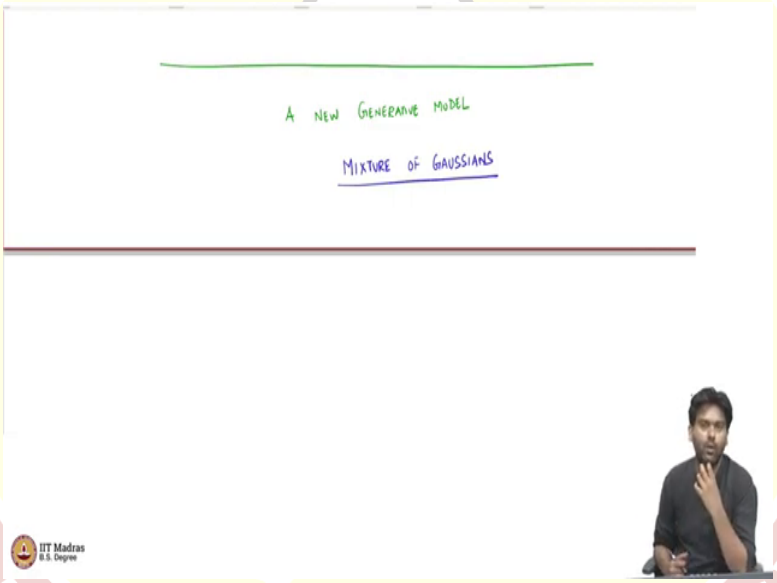
Now the question is, well, what is this? This is the PDF of Gaussian with mean $\hat{\mu}_{ML}$ and variance some σ^2. Now, the question is, the more important question is, is this a good model for this data? Now, the problem with this model is, yes. So, the mean is the sample mean and which is somewhere in the middle of the data points, all that is fine. But then if you stare at this model for a while, you understand that there are these data points which have occurred in our data, but which has very less, which comes from very less dense regions.

So, if I try to find the probability of data coming in this region on the real line and in this region on the real line. Now, according to our hypothesis, the best Gaussian also has very less probability in these regions. However, we have seen data points in this region. So, now what is the problem? The problem is we have made an assumption that the model is Gaussian. And then maximum likelihood assumes that that is the gospel truth and then it is going to try to find the best mean, which explains the data under the model that we have assumed.

So, but if the model that we have assumed is not powerful enough to generate data of the form that we actually see, then maximum likelihood cannot do anything better. So, because it is searching in a space of Gaussians and then finding the best Gaussian. So, clearly we have seen a lot of data in these two regions, but then our explanation by via Gaussian is not very satisfactory, because these are low density regions for the Gaussian.

(Refer Slide Time: 07:29)





So, then how can we what do we want? Well, we want something like this. So, we do not want a Gaussian project to explain this data. But we want what kind of PDF do we want? So, if you think about this, I mean, this is a good place to pause and think, how do you think the shape of the PDF should be that explains this kind of data? What do you want? We want three different modes or three places where the density should peak.

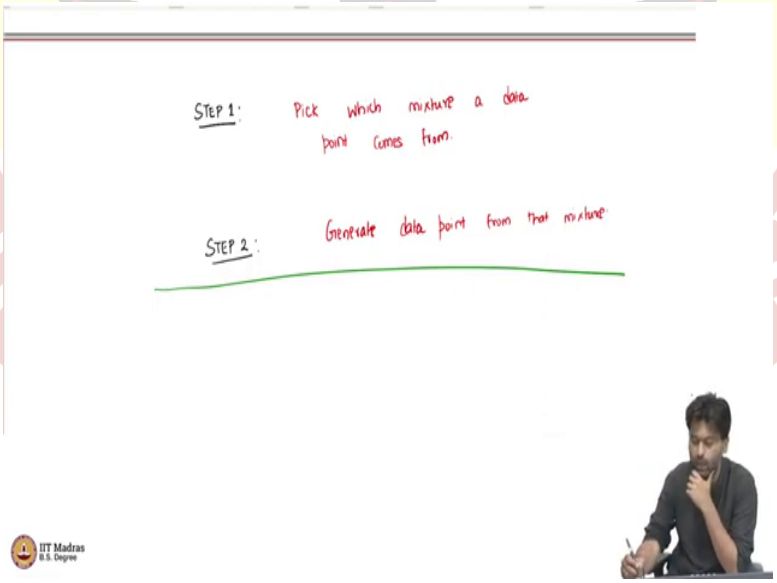
So, there are some data points here, there are some data points here and there are some here. So, your PDF itself should have three modes. So, which means you need something like this perhaps. You need a density like this. So, let me write that down.

So, we want a density like above, to explain this data, why then a density like this has the property that it is it your data that you are actually seeing are coming from high density regions. So, perhaps this is a better model for this data. But then this is not a Gaussian. Clearly, this is not a Gaussian, we know Gaussians are unique model that is only one peak for a Gaussian, but then this needs three peaks. So, which means we need to come up with a different type of model to explain this type of data.

So, as you can see, there is some kind of clustering behavior in the dataset. So, that is what we will eventually get to. So, and then we want to somehow develop a probabilistic model, which can do this, which can model this kind of cluster data points. So, how can we do this? Well, each of these high dense region looks like a Gaussian in itself. But then overall, it is not a single Gaussian.

So, basically, what we then want is a new model. And what we want is a new generative model to explain this data, or the generative story and the name of this new model that we are going to come up is called as a mixture of Gaussians. It is not a single Gaussian. It is a mixture of Gaussians. In this case, it is a mixture of three Gaussians in the example that we just saw. So, what is the story? So, now what is the underlying story that generates this data?

(Refer Slide Time: 10:00)



Whenever I say story, it means that what is the probabilistic mechanism that generates this data? So, I see the data. I need to understand. I need to put down a mechanism that generates the data. So, far in the estimation things that we have seen. The story is simple, you either sample from a Gaussian or you toss a coin into super simple. Now, such simple stories are not

enough to explain this slightly complicated data. So, we need to come up with a new story. And the story that we will develop now has two steps.

I supposed to just tossing a coin, which is a single step, which will generate our data. And now we have a two step story. So, what is step 1 of the story? Well, we will look at both steps carefully, step 1 is the following.

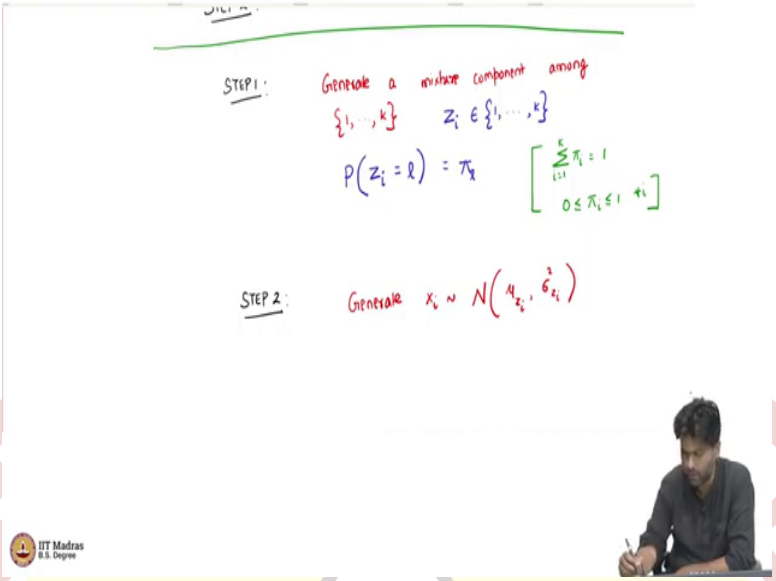
So, pick which mixture, a data point comes from. So, the way we are going to model is. So, I have seen the data, I want to understand how each point is generated. I look at the first point. And now I want to explain how it was this point generated, x1, how was this generated? Well, before generating x1, something has happened in a probabilistic scenario. So, and that is what we are trying to explain now.

And that something has two steps to it. And the first step is when the mixture was first picked. So, somehow, we first decided, well, which of these mixtures there are three mixtures in the example that we showed. And let us for now assume that we know the number of mixtures. So, I give you a number of mixtures. And then you decide first somehow which mixture the data comes from, which you can think of it as mixture, or if you want to think of it as cluster, that is also fine. But which means the common parlance is mixture. So, I am going to stick to that.

So, which mixture the data comes from? So, once you have decided the mixture, then what do you do? Then it is pretty straightforward. So, then the second point is, once you have which mixture it comes from, well, you generate data from that mixture, generate data point from that mixture. So, this is a two step process, I supposed to the single step that we have seen so far. First, decide the mixture and then generate the data point from that mixture, super simple.

Now, of course, we need to make this more probabilistic. So, the way that I have just put down step 1 and step 2, there is no probabilities involved. But because it is a probabilistic model whose parameters we are trying to estimate, we need to make it more precise. And let us just do that.

(Refer Slide Time: 12:27)



So, let us make this precise. So, step 1, in a more precise way is going to look like this. When I say we are figuring out which mixture this point comes from? Well, the way it is going to happen is imagine that, if you want to generate from a data point from three mixtures, then assume that the model has a dice, which has three faces. And now I am going to throw this dice. So, I am going to roll this dice and then the dice falls on one of the face. So, on that face, will have a number either 1, 2 or 3.

Now, we are going to think of this number as the mixture from which this data point is being generated. Let me formulate that. So, we are going to say we are going to generate a mixture component among, well, in the example it was 3 but general it can be k mixture. So, let us make it slightly more general, among 1 to k. There are k mixtures. And I am going to call this mixture component as zi. So, this is a number that we have to first decide between 1 to k, which indicates which mixture the ith data point comes from that values zi.

Now how am I deciding zi? I am deciding it in a probabilistic sense by rolling a dice. So, if the dice rolls and falls on the face 2, then it means that zi equals 2. The ith data point goes to the second mixture comes from the second mixture. So, what does that mean? That means that, there is a probability distribution, which is the dice that I am formulating it as probability distribution. The probability that zi equals let us use some other things. So, let us say l equals some πl. It just means that the probability that the ith data point comes from the lth mixture is given by some πl.

So, if there are only three mixtures, then let us say $\pi 1$ is 0.5, $\pi 2$ is 0.25, $\pi 3$ is 0.25, then it means that 50 percent of the time I am going to get a point from first mixture, 25 percent from second mixture and 25 percent from third mixture. So, we are assuming that this is step 1 of what is happening under the hoods. So, of course, π is a probability so which means that sum over i equals 1 to k πi will be 1. Because they are the sum of two probabilities, I need to choose one of the mixtures in a probabilistic way. And also, we know that 0 less than πi less than or equal to 1. Again, these are probabilities for all i.

All I am saying is that when you are selecting one mixture in a probabilistic fashion where the probabilities are given by $\pi 1$ to π capital K and which face the ith data points dice falls on, we are going to call that zi. So, if you remember from K means we also use zi as a cluster indicator for the data point here it is the mixture indicator, if you will. So, this is first step. So, now we have decided which cluster or which mixture the data has to come from.

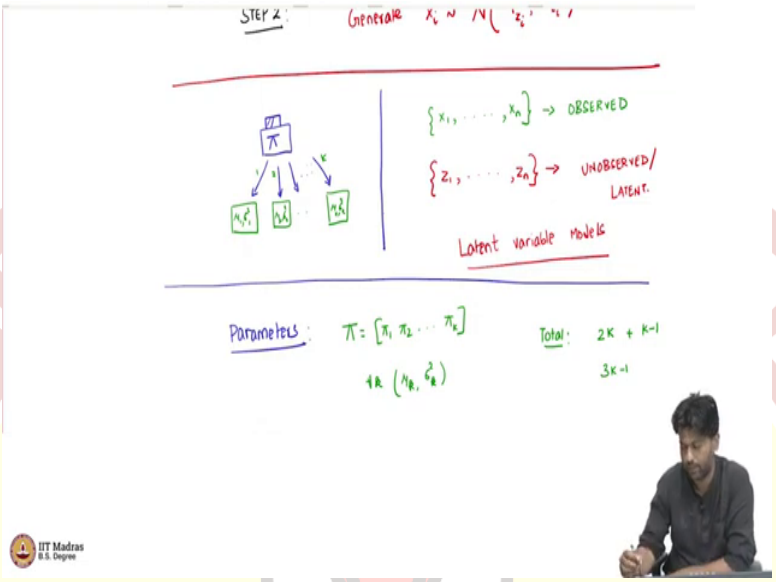
Now the second point, second thing is second step is pretty simple. So, let us say zi was some 5, which means that I need to generate from data from the fifth mixture. Now for each mixture, we have a Gaussian with its own mean and variance. That is the assumption. So, every mixture has its own mean and variance. And now if I roll the dice, it falls on face 5, then it is as if I am going into the fifth door, which tells me there is a mean and variance sitting inside the fifth door and then I will pick a random data point according to that particular Gaussian.

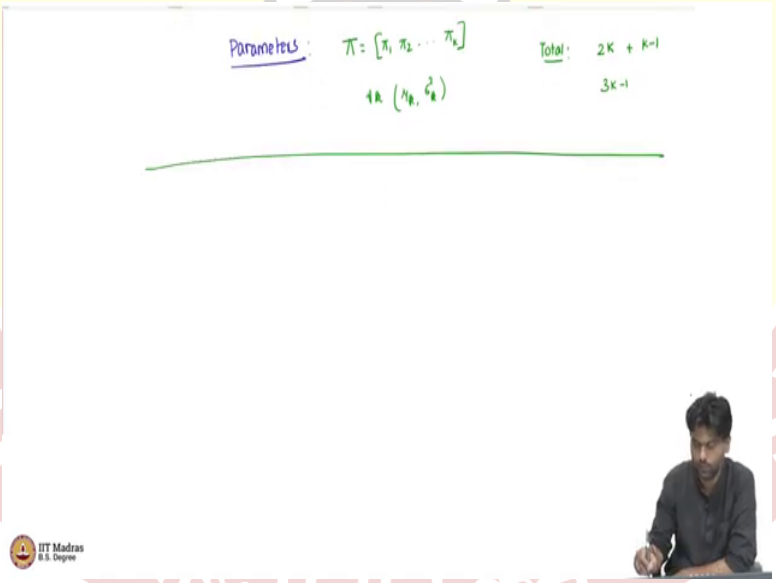
To formalize, this, we will just say generate xi, the ith data point as a Gaussian or normal, which I am using N to denote with what is the mean? Well, the mean is the mean corresponding to μ_{zi}. So, zi is the cluster indicator, which door I am going in to pick a point. So, there is each mixture as a door and then let us say go into the fifth door, then means zi was 5. That is why I went to the fifth floor. And then I am sampling a point according to the mean and $\sigma^2 zi$ according to that particular mixture.

So, this is the generative model, it has two steps. And this is how one data point is generated. To generate a single data point, you go through these two steps. First roll a dice to cut which mixture and then go to that mixture sample Gaussian from that Gaussian data point according to the mean and variance of that mixture.

Now, to generate the second point, you again assume the same story. So, you again, roll the dice, it might fall on a different face, you go to that corresponding mixture and then sample and keep doing this and different times you get a dataset.

(Refer Slide Time: 17:41)





So, if you have to put this in the in pictorial form, like how we have been seen so far, it will look something like this. So, you have the first box, which has π sitting in it. You press this button, what you are going to get is well, one of the values which is 1 to k with different probabilities determined by π. So, once you decide which box, then you go to that particular box and then each of these 1 to k has a corresponding box, which has μ1 $\sigma^2$1, μ2 $\sigma^2$2,..., μk σ^2k.

Now and then, according to each of this, when depending on this, let us say zi was 2, I went to the second box, I get xi from that and so on. So, that is in general. So, this let it be general. So, now, this is a slightly different model from what we have seen so far, because it has two steps, which means of course, we have the data x1 to xn, which we are observing, which are real numbers and these are the observed quantities.

But then in the model that we are put down, there are some unobserved quantities also, that determines x1 to xn. And these in this particular case, what are the unobserved quantities? Well, the unobserved quantities are z1 to zn, the mixture indicator or the cluster indicator or unobserved. So, these are unobserved or latent. So, these are what are called as latent variable models.

So, our Gaussian mixture model is a latent variable model, because the final output that you are observing, you are assuming depends not only on some parameters that you want to estimate, but also on some unobserved latent variables. So, these are latent variable models. It is just an example. There are several latent variable models that people typically use and this is one of the most commonly used ones.

So, now as an estimation procedure, the question is, what are the parameters that we need to estimate in this model? This is a good time to pause and think how many parameters will determine this model completely? I will tell you that now. So, what are the parameters? You only see the data now, what are the parameters that you need to figure out? Well, in the Gaussian case, simple Gaussian case, it was just the mean or mean and variance.

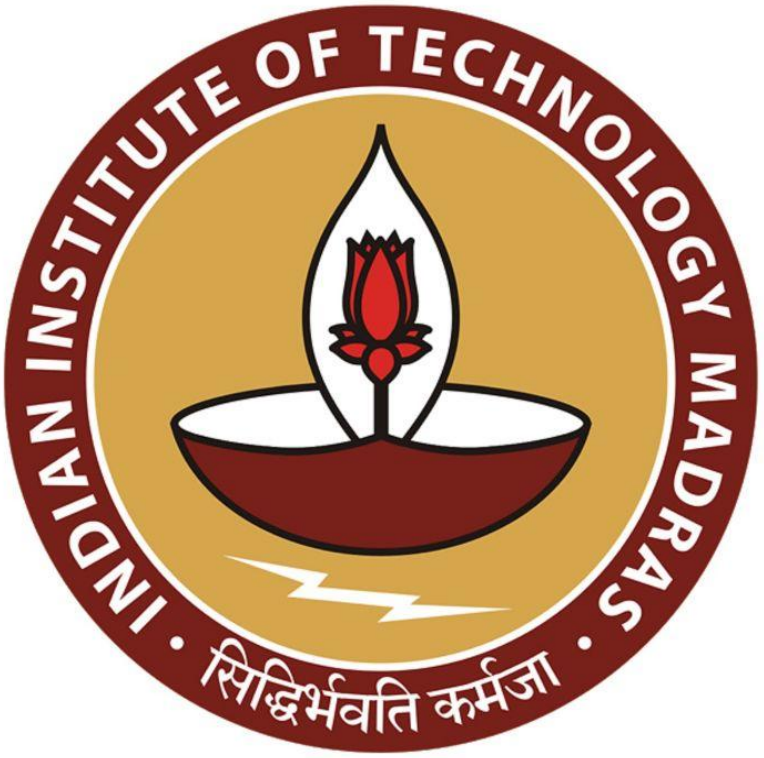
Now we are seeing there are many more parameters. So, I mean, the first thing is there is a π, which is a vector of probabilities, $\pi 1$, $\pi 2$,..., πK. This is something that we do not know, but then it determines the output. So, this is a parameter of interest. And then for each K there is a μK and $\sigma^2 K$. So, there are two parameters. In total, you have 2K mean and variance for each of the mixtures. That is that is 2K parameters, plus there is a common π sitting well, which gives probabilities for each data points.

You can either think of π as having K values, but then because π has this condition that the sum of πi's should be 1. If you know K - 1 of them, the last one comes for free. So, because they have to summed one, there is a there is a restriction there. So, you can think of it as K - 1 free parameters, if you will. So, overall, it is of order of 3K parameters. I mean, if you want

to be pedantic, you can say it is 3K - 1 that is also fine. So, we need to estimate these many parameters from data.

So we are earlier, we were just estimating one parameter, the bias of the coin or the mean of the Gaussian distribution. Now, there are host of other parameters that we want to estimate. Now, it is not obvious what are these estimators? So, just by looking at the data, how can we decide what are πs, what are μs and σ^2? It is not at all clear. So, it is hard to make a educated guess as to what these things would be.

But our method of maximum likelihood is a very principled method that does not that will give you an answer no matter what the model is going to be. So, as long as we have a well specified probabilistic model, we know that you can run the principle of maximum likelihood on it. So, let us try that. So, that is what we are going to do next.

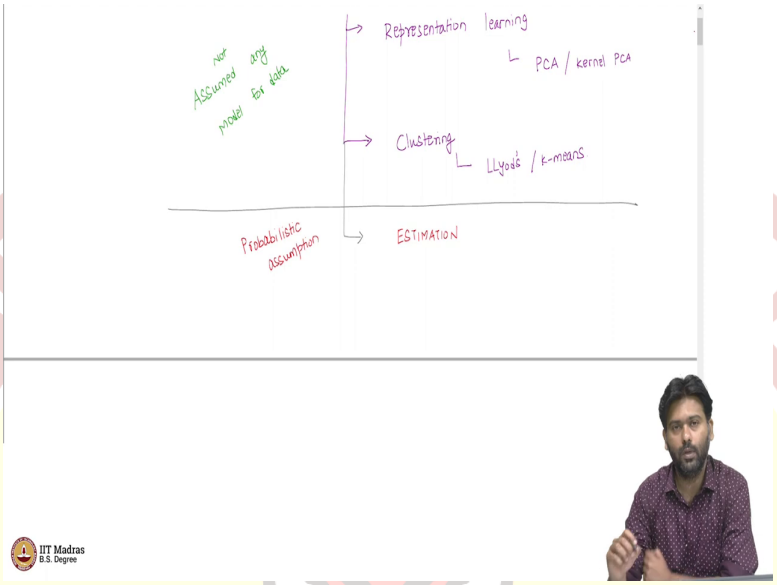


IIT Madras

ONLINE DEGREE

Machine Learning Techniques
Professor Arun RajKumar
Department of Computer Science and Engineering
Indian Institute of Technology Madras
Introduction to Estimation

(Refer Slide Time: 00:14)



Welcome back, everyone. So, far we have been looking at unsupervised learning. And specifically, we have been looking at two main paradigms of unsupervised learning specifically, one is representation learning and the other is cluster. So, we have looked at representation learning. And in representation learning, we have looked at PCA as one way to learn good representations, when the features have some linear relationships among them. We also looked at kernel PCA as a means to learn nonlinear relationships among data points via the use of kernel trick.

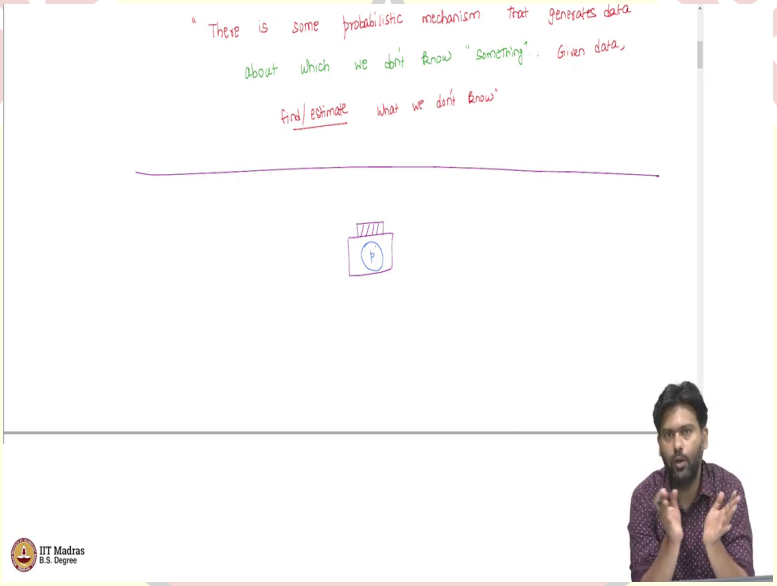
So, another possible way to look at unsupervised learning is via the idea of clustering, which is also something that we have looked at and in clustering, we have looked at the Lloyds or the K-means algorithm. Now, in both these ways of looking at unsupervised learning, one thing that we have not done is that we have not assumed any model for data. By which I mean that we have not assumed any probabilistic model that generates the data. And we will see what that means as we go along.

The goal of this video and a few other videos to follow is to look at a different type of unsupervised learning paradigm, which is not just for unsupervised learning we will see how methods we use are actually useful even for supervised learning later on. But the idea is, we

are going to look at something called estimation where the basic difference between what we have seen so far and what we will see in this video and the few following videos, is that we are going to make some probabilistic assumption about the data.

So, we will see what it means to make some probabilistic assumption about data, it is still going to be in an unsupervised fashion. So, we are still in the unsupervised world. But then we are going to move away from the deterministic methods that we have looked at so far and then look at a more flavor of probability involved when you have some probabilistic model for data.

(Refer Slide Time: 02:55)



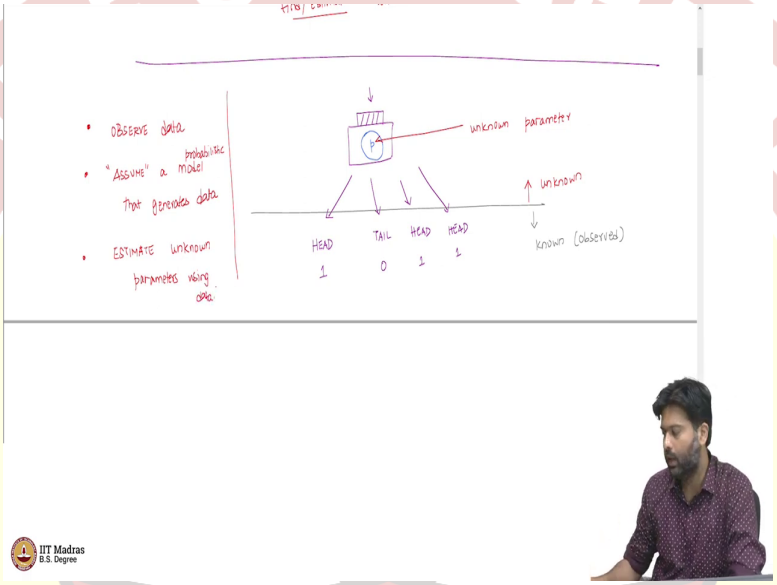
So, what does it mean to say, there is a probabilistic model, basically, our hypothesis is the following. There is some probabilistic mechanism that generates the data. So, this is the assumption that we work with. Now, what else about this probabilistic mechanism that we have to look at? Is that, about this mechanism about which we do not know something, so, we do not know something. So, it is not that we are completely aware of the details of this probabilistic mechanism, there is something that we do not know and I will make it clear what this something is.

And the goal is given data, find or estimate what we do not know. I am trying to put this in a very high level view of what estimation means and we will make it more precise as we go along. So, the first question we ask is, what does it mean to say that there is a probabilistic mechanism that generates the data? For that let us start with a simple example. The way I would like to think of this is as follows. Let us say you have a box and the box is box which

is a black box, let us say which you do not know what is inside it. But then there is a button on the top of this box.

For the moment, we will assume that the inside the box there is a coin and this coin is not necessarily an unbiased coin. In other words, it has a head and tail on either side. But then the chance that if you flip this coin, head will occur is not necessarily 0.5. It could be 0.7, it could be 0.9, it could be some value p, which you do not know. And that is this coin inside this box.

(Refer Slide Time: 05:08)



And what happens is that every time you press this button. This coin gets flipped inside this bias potentially biased coin gets flipped inside. And what you observe is the outcome of this experiment. In this case, let us say its head. Now, let us say head is head means 1, maybe I press this again, I get a tail, tail mean 0 and then I press it again, I get ahead. Let us say head is 1, I press it the fourth time again, I get ahead, let us say head is 1. And I can let us say I have a bunch of such, what I am going to call as observations.

Now, these observations according to us are generated according to this probabilistic mechanism, where the probability is completely specified by this coin and its bias. Now, you can think of it this way. So, there is this side, which is known to us or observed and this is what we see. So, this is what is given to us. Now, once this is given to us, we are assuming and then and let me highlight this word assuming that we are assuming that there is some mechanism that has generated this data. It may or may not be true in real world, but we are going to make some assumptions and then work with such assumptions.

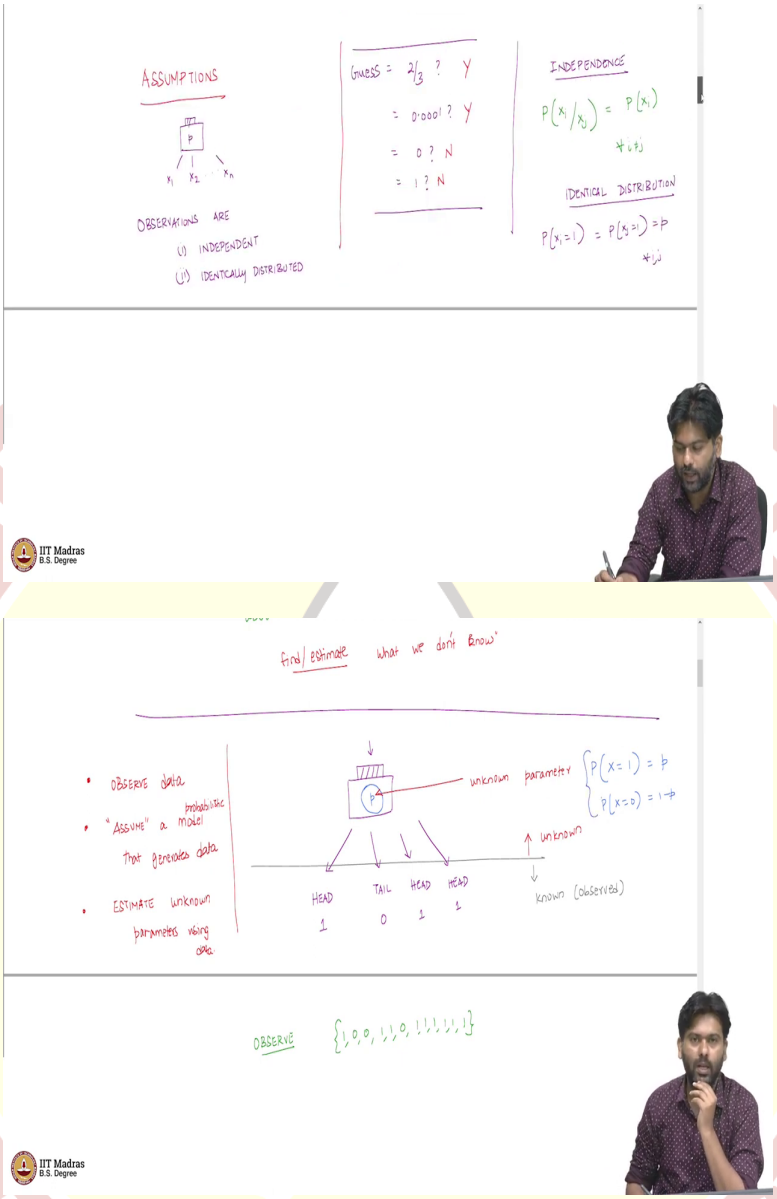
How good are these assumptions and so on, we will talk about later? But for the moment, let us say we make some assumption. In this case, I, you can think of it as I view a bunch of data, which are all zeros and ones. And my assumption about how this data is generated is the story that I build to explain this data is that there is a box with a coin. And then every time I press the box, I get an observation that is how these four observations were generated. So, this is an assumption that I am making.

And remember that this assumption, this part of the picture is what is unknown to us? So, in specifically, so this P is what is unknown. So, this is an unknown parameter. And you can think of this unknown parameter as what is a compressed representation of the data? So, I might have 1000 points, 1000 values, 0, 1 values, but then, to explain these 1000 data points, which are all zeros or ones, there is only a single number, which is the bias of this coin. So, if I know the bias of this coin, I can explain how this data is generated, I have the story. So, it is like I am tossing this coin 100 times, 1000 times and then getting the data.

So, you can view this parameter as a compressed representation of the data observed data that you see. But of course, we do not know what this parameter is. And so the goal of estimation itself is the following. So, you observe some data and you assume and again, I cannot stress this enough, you assume a model, in this case, a probabilistic model that generates data. So, there is some model that generates data. And under this model, you want to estimate what you do not know about this estimate, what you do not know about this model.

So, I have assumed that there is a coin. And that coin has some bias. But if you do not know what this bias is. The proxy for understanding this bias, this value P, this unknown parameter is just my data. So, I see this data, I have made this assumption about the model, I do not know what the value of P is. I want to estimate unknown parameters using data. So, this is the general idea of estimation.

(Refer Slide Time: 09:09)



Now, let us take a simple example, with again, with the coin toss, but then not just four data points. Let us say we have more data points. The first thing is you observe, so you observe, let us say, 12 or let me make it 12. So, 1, 0, 0, 1, 1, 0, 1, 1, 1, 1, 1, 1. Let us say this is the data that you observed.

Now, I want you to the model that I am assuming is, of course, that there is a box and then every time I press a box, a 1 or 0 gets generated and the box has a coin with some bias P when they say bias to just to make this very, very precise. This bias is just the probability of some random variable x taking value one = P and the probability that x = 0 is 1 - P. So, that is what it means to say there is a probabilistic mechanism that generates the data.

So, and from this mechanism, I am assuming this. So, I am observing this data is the assumption. So, now what would be a good estimate for this particular data? So, it is a good exercise to pause and think what would you estimate if you are given this bunch of data points. So, pause and think about this and then I will probably try to guess what you might have guessed. So, basically, if you counted the number of ones in the data, so there are 9 ones if I have written this correctly and there are 3 zeros out of 12 observations, 9 of them are heads and 3 of them are tails.

So, my estimate could be something like 9 over 12, which is three fourth. So, this could be my estimator for the underlying P. Now, what do I do with this estimator? Once I have a once I believe that I know what the true P is or I made a guess for the true P. Now I have my own box. So, now I have this box, where there is a coin and then I am saying the coin has bias three fourths.

Now, once this is there, now, I know I can play god. So, I can create data if I want to, so I do not need the data anymore, because I understood the mechanics of how this data is generated. And that is all there is to this data. This data only contains information about the underlying mechanics, which I am assuming that generates is and once you know the mechanics in this case, which is to know the P, then I do not need the data anymore. So, I have the box and then I have put the values for the parameters. And I am happy.

If we want, I want to generate as many data points as I wish from this box. So, I can do that. So, but remember, what you have, is still a guess. So, this is a guess, this may or may not be the truth. So, the true P may not necessarily be exactly the same as your guess. In fact, in the moment we introduce probabilities, we are never going to be 100 percent sure that we know the true P that generates the data. So why is that? Well, if you think about this, well, what about a guess which is not three fourths, but then let us say two thirds.

In other words, can there be a coin with probability P as two thirds which could have generated this data. Think about this. So, if the true P of this coin is two thirds, could I have gotten this data. Or is there absolutely no way I could have gotten 9 ones out of 12 tosses? The answer is you could have gotten this data. So, even with two thirds probability, there is a chance that you could have gotten this data that is nothing I mean, rejecting the possibility that the truth could be two third.

Well, can the true P 0.0001, can this be the probability of seeing heads? Well, what would you think? So, if you think about it, well, I cannot simply rule out this probability. So, this value of P. There is still a nonzero chance that I could have seen this data, if the true P is 0.0001. Of course, my data is not representative within course, the word representative of this model, where P is 0.0001, which means that I must have had an extremely unlucky day that I see 9 heads out of 12 when the chance of falling heads is 0.0001, 1 and what, 10,000.

So, that is a very rare event, but that does not mean that it cannot happen it could have happened. So, which means, both two thirds and 0.0001 are still possibilities. Now, what about P equals 0? Is this possible case? Well, we are saying this is this could have generated data, this could have generated data. What about P equals 0?

Now, if you think about it, P equal to 0 means what? So, the probability of seeing heads is 0. Which means, if I toss that coin 12 times I should have seen only tails, but then I see both heads and tails in my data which means that P equal to 0 is simply not possible. This cannot happen. So, my true P could not have been 0, I am sure about that.

Similarly, my true P cannot be 1, I am sure about that because with 1 you could not have generated this data you would have got an all heads but then we see a tail as well. So, the only thing that we are sure about is that the true P cannot be either 0 or 1. Everybody else between 0 and 1 is a potential contender for the true P that could be possible. So, there is a chance.

Now in this case, we still somehow guess three fourths, when every P is a possible candidate for the true P. Now, why did we guess that? So, in general, what went in our minds when we guessed this? So, were there some implicit assumptions that we made before making this guess? In fact, we must have made some assumptions in our head before we guess three fourths? And what are these assumptions? So, let us think about what these assumptions are for a moment.

So, the assumptions that one typically works with in when we deal with data are the following. So, let me put it here assumptions. Well, so in this simple case, you have this box with the unknown parameter P and then you get let us say, x1, x2 till x n. So, n data points each of these 0 or 1 and then you have n of them.

Now, there are two basic assumptions we are making about this data. And these are very common, the assumptions are as follows. The observations are first thing is, I will put an

assumption and then explain what they are, are independent. And the second assumption is that they are identically distributed.

So, what do these assumptions mean? Well, independence simply means, in this case, this is a probabilistic independence, which means that the probability that x i equals 1, or rather in general x i take some value given x j takes some value is same as the probability that x i takes that value for all i comma j, so, I naught equals j.

What does that mean? That means that, if I told you that the third toss was a head, then that does not change my uncertainty about my fourth toss. So, the information that the third toss was a head does not affect the probability that the fourth toss is going to be ahead or tail. So, that is when we call these tosses as independent tosses. And this happens for any set of any pair of observations that you see. And you can also argue that n I mean, P of x i, given any set of other random variables here is still same as P of x. So, they are completely independent of each other. So, that is the first point. That is something that we are assuming implicitly.

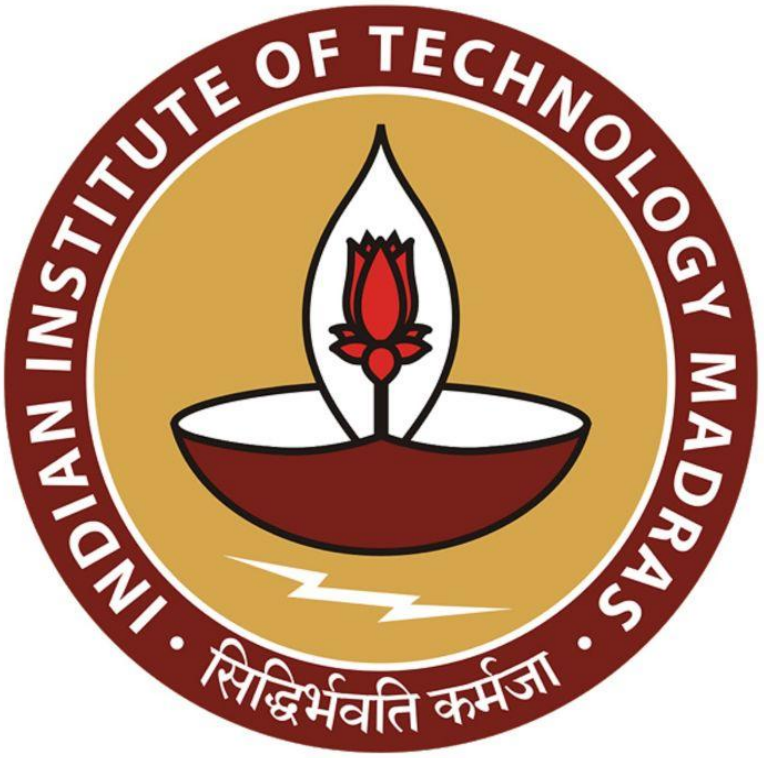
The second assumption is that identical distribution. What does that mean? Well, that means the following. That means that P of x i equals 1 equals P of x j equals 1 equals P for all i comma j. That simply means that I mean, in pictures or in words, it means that we are using the same coin every time. So if I, tossed the first coin, it fell heads. And in the second toss, if we pick a different coin and toss it, well, these two coin tosses are still independent. So, because the first toss being head does not tell me how the second coin toss would be head or tail, but then they are not identically distributed, because these two coins might have different biases.

But here we are saying that well, you are essentially pressing and then getting the same coin gets flipped, all n different times. So, the distribution that determines the outcome, it is the law that determines the outcome that you observe in a probabilistic sense is exactly the same every time. So, that is what it means to say the data points are identically distributed. So, basically, you can think of it as saying there is only one box. So, there is one box and then you press the same box, every time.

So, now, let us come back to our question of there are all these guesses, which are possible, but then we still made a guess three fourths. Now, we are assuming that the data of has these two, satisfies these two assumptions that they are independent and identically distributed. Now under this assumption, now, is there a way to say one guess is better than another guess?

Is there a more principled way to say two thirds is a better guess than let us say 0.0001 or three fourths is a better guess than any other guess? If so, then yes, guessing three fourths might make sense in a certain way.

So, otherwise, these three fourths seems to be arbitrary. So, it is intuitive that you should guess the chances, the number of heads in your data fraction of heads in your data. But it is not at very principled. So, the question is, is there a principled way to get estimates from data?

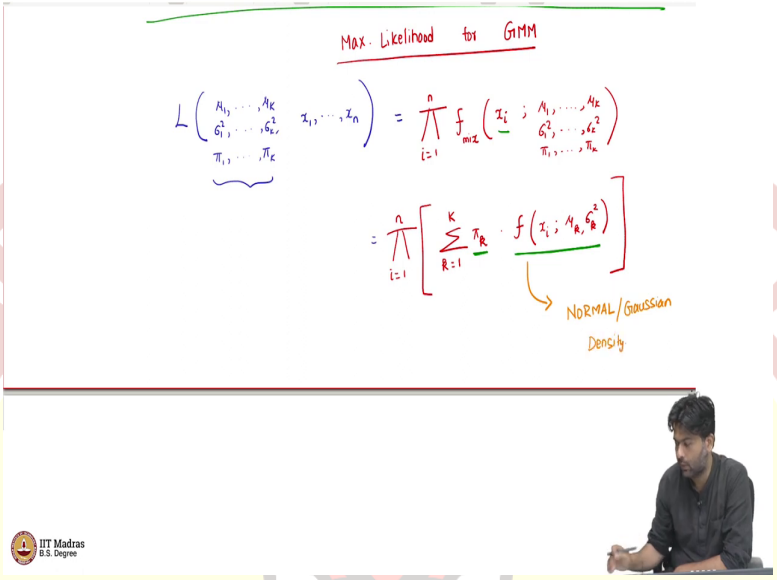


IIT Madras

ONLINE DEGREE

Machine Learning Techniques
Professor Arun RajKumar
Department of Computer Science and Engineering
Indian Institute of Technology Madras
Likelihood of GMM

(Refer Slide Time: 00:14)



We are going to try maximum likelihood for the Gaussian mixture model which is called as GMMs, sometimes Max Gaussian Mixture model. Now, the likelihood function L is a function of a lot of things. So, it is a function of $\mu 1$ to μk, which we do not know, $\sigma 1^2$ to σk^2, which we do not know and $\pi 1$ to πk, which we do not know.

And of course the data x1 to xn, which we have observed. It is a function of all these things. Of course, we are going to treat it as a function of the parameters, the data will act as a constant in this function and then we will maximize only with respect to the parameters like how we have always been doing. Nevertheless, let us put this term.

Now the i.i.d. assumption still holds, so every data point is generated according to the same 2 steps. The second data point, again, you go through the same 2 steps, which means that knowing the outcome of the first data point is not going to affect the probabilities of the second data point taking a certain value, independence still holds. And then it is the same process that generates each of the data points. So, identically distributed also holds, but then the distribution is going to be slightly different.

Now, what is that distribution? Well, because of independence, the first thing is we can write this as a product of i equals 1 to n. Now I am writing this, this pi is a big pi, it is not the pi,

which is a parameter, so it is just a product as usual. Now, there is some mixture distributions density, which is determined by where we have to see what is the density of observing xi when you have parameters μ1 to μk, $\sigma 1^2$ to σk^2 and π1 to πk, small π1 to small πk. Well, what is this mixture density?

Now, because we have these 2 steps, what is the density of a point xi is determined by which mixture it comes from. Now which mixture it comes from is determined by the roll of the dice, which is determined by our probability vector π? So, I can write this density function itself as $\prod_{i=1}^{n} [\sum_{k=1}^{K} \pi k.\ f(xi; \mu k, \sigma k^2)]$ What is happening here? What I am saying is, well, what is the density of xi coming from this mixture distribution?

Now we are seeing, we do not know what was the coin what was the face on which the dice ended up in when the step 1 was performed? Because that is an latent variable. So, we do not get to see that. But then we are assuming that there is some probability πk that it would have come from cluster 1, the same point xi.

There is some probability π1 from cluster 1, π2 from cluster 2 and so on. It could have come from any cluster? So, we do not know that Apriori. So, we have to use the fact that well, it could have come from any cluster. So, I am weighing the chance that it is coming from a cluster by the probability that it comes from that cluster.

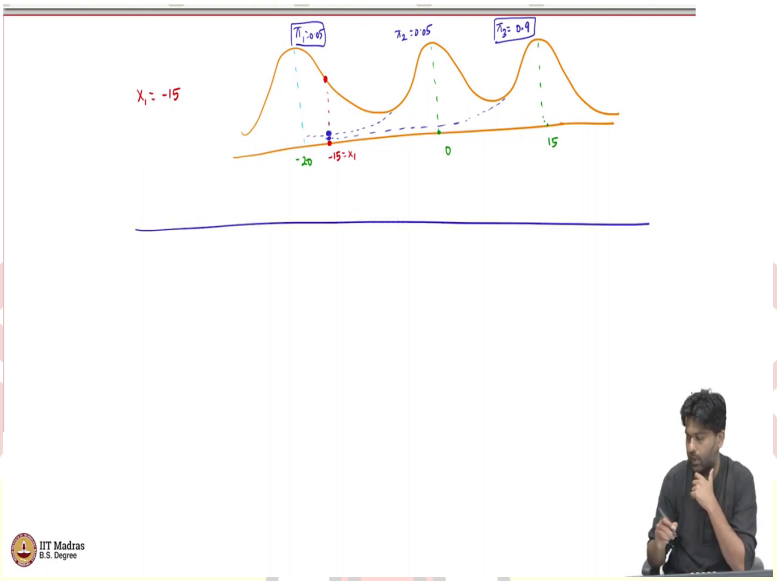
So, but we also know that if it comes from cluster 1, it cannot come from cluster 2. So, these are mutually exclusive events. Coming from cluster 1 is completely exclusive of coming from cluster 2. So, the chance of seeing this data point is a sum of these mutually exclusive events of coming from cluster 1, coming from cluster 2 and coming from cluster k. So, I can add these events chances up, but then what is the chance that it comes from cluster 1? Well, if it has to come from cluster 1, 2 things should have happened.

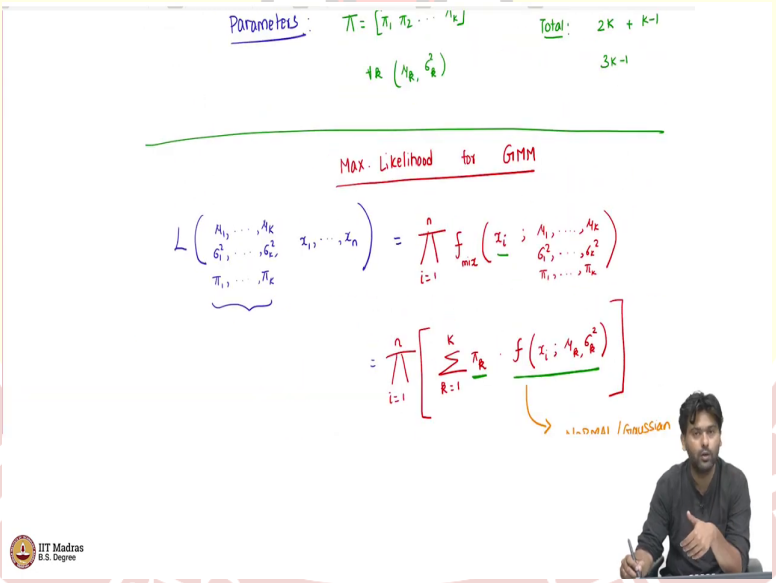
The first thing is that well, the dice should have fallen on face 1, which means the probability of that happening is π1 and this point should have been generated according to Gaussian with mean by μ1 and σ variance $\sigma^2 1$.

So, now that is a product these 2 things have to happen together, that the point was chosen from cluster k and mixture k and then well, mixture k itself gives this point which is this density. So, now what density is this? Now, this is just a Gaussian density. So, this is a normal

density or Gaussian density, because that is our assumption. That is the Gaussian mixture model which we know how it looks like.

(Refer Slide Time: 05:11)





So, just to give some intuition here, so let us say the true density looked like this. This is a the true density, which means that is some mean. Let me put some numbers, maybe - 20. Maybe the mean of the second Gaussian was 0. Maybe the mean of the third Gaussian was 15. Let us say I saw x1 as - 15 which is a point here. Now, this does not mean that immediately that x1 necessarily came from cluster 1. Not necessary.

We follow 2 steps. Well, what could have happened is, of course, it cluster 1 could have been chosen and then this point came from cluster 1 according to this density value. Now, it could have very well been the case that cluster 2 was chosen when I rolled the dice.

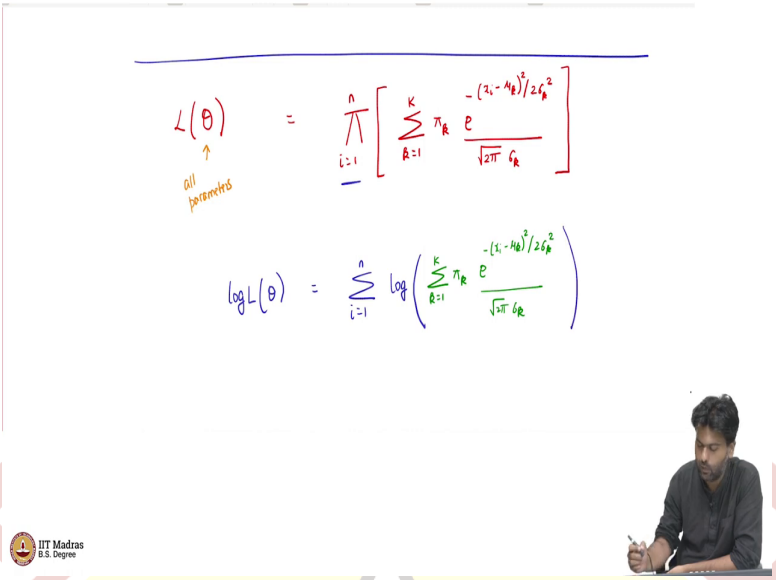
Now, in that case, the density of this would be from the second Gaussian, which would be smaller value. So, because it is closer, more to cluster ones mean, of course, the density that of that cluster 2 explain this point is smaller. It could have come from cluster 3 also where the density is even smaller. So, it is super small, but then it is not 0. Gaussian will not give you 0 values for any point.

So, it could have come from any of these, so we cannot immediately dismiss the others, it could have come from any of these. In fact, it is not just the closeness to the means that determines this. It is also the π's that determine this. So, it could be that $\pi 3$ was 0.9, $\pi 1$ was just 0.05 and $\pi 2$ was just 0.05, in which case that though - 15 is very close to - 20, the chance that the first cluster was picked itself is only 5 percent. Whereas, the chance that the third cluster was picked this much higher.

So, in which case, it is not just how much density that the Gaussian has for this point, which depends on how close you are to the mean, that determines the density of seeing this point. But it is also the chance that such a cluster was picked.

All these are unknown variables at this point. So, we have to factor in all of these. And that is exactly what this equation is here. Let us go ahead now and then see what is the density and how we can do maximum likelihood here.

(Refer Slide Time: 07:48)



So, now, I am going to write the likelihood function. I am going to create, I mean, there are so many parameters, I am not going to keep writing all the parameters every time, let me just call it as θ. So, this is all parameters. So, μ, σs, and our π's, it is all put together. I am just calling it as likelihood or as the parameters.

So, this is product of i equals 1 to n, because of independence. And now the ith point was generated according to k equals 1 to K πk into the Gaussian density that the kth cluster generates this, which we know is $e^{-((xi - \mu k)2/ 2 \sigma k2)} / \sqrt{2\pi}\, \sigma k$.

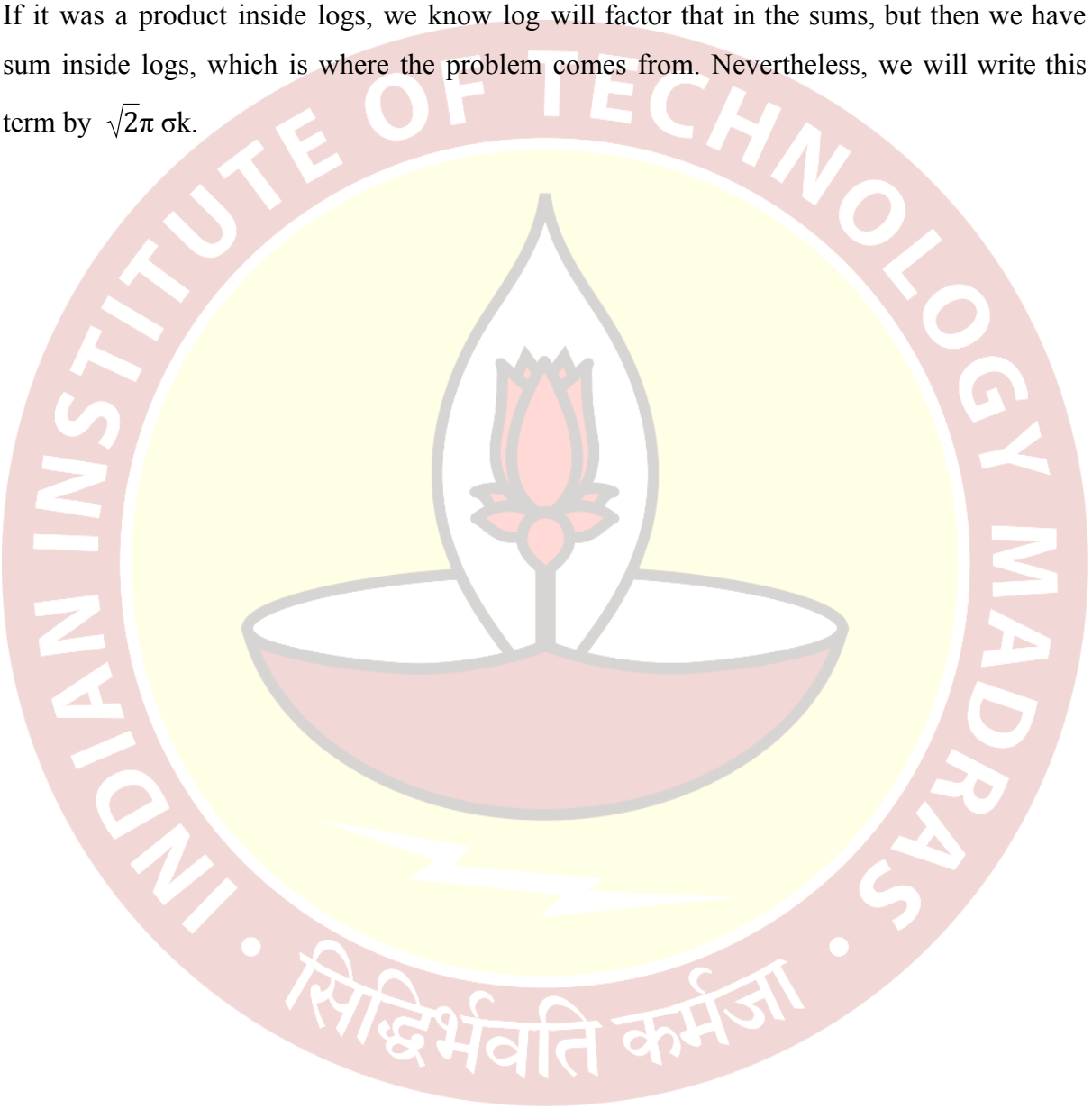
So, this is the actual mixture, the likelihood function, which we are trying to maximize. Well, it is a likelihood function that is nothing stopping us from writing down a complicated likelihood function. That is the whole reason of going to likelihood functions. Because if it was always simple as sample mean and fraction of heads, I mean, why develop a theory? So, we are doing this because we want to solve complicated problems and we better be able to handle such complicated problems.

So, this is the likelihood function. Of course, this is a product of a bunch of things, it is easy to typically handle sums then products. So, our usual method would say that look at the log of the likelihood of theta. And how that looks like? Let us see that. Now that is going to be sum over i equals 1 to n, this product here becomes sum. You have a log. Now, here is where we hit a bottleneck in this very step.

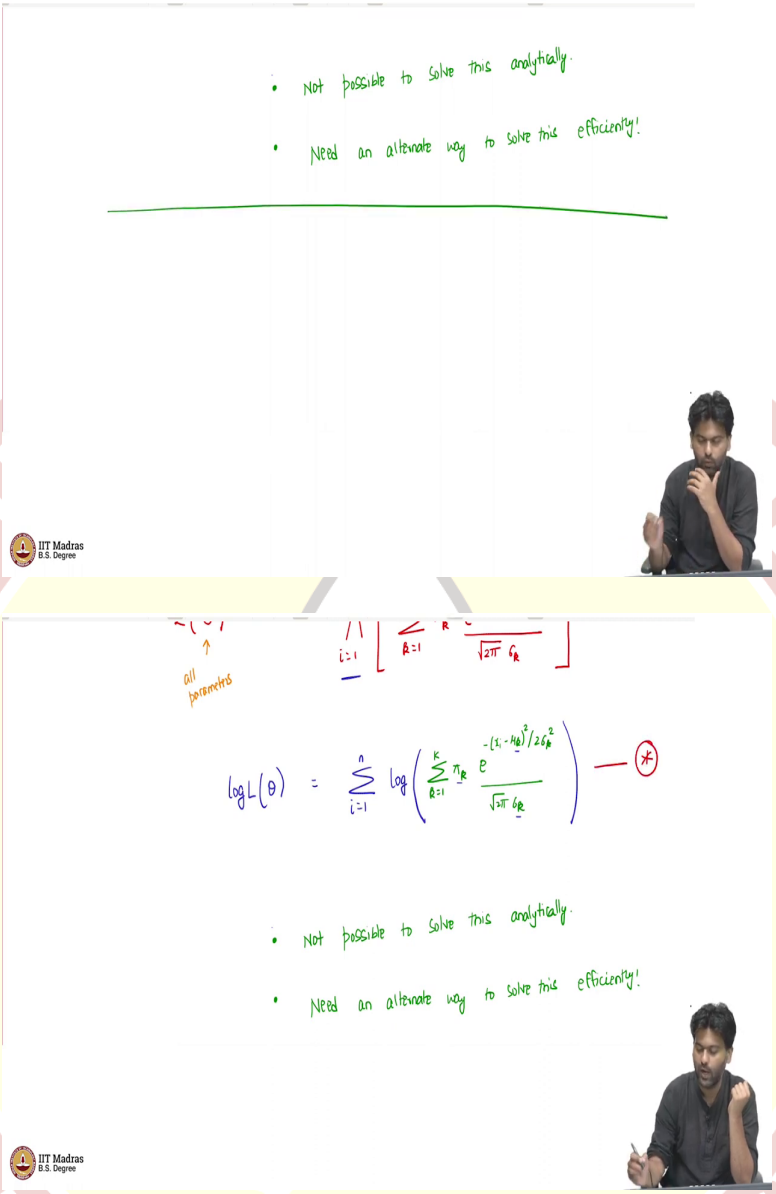
Now what is happening is earlier, this logarithm, serves two useful purposes. One it converted products to sums. The second thing is that it simplified our density really well. So,

if there was a Gaussian, then it had a e power something. You did a log and the logs and the e's cancelled. But now what is how happening is that is a log inside, I mean, there is a sum inside the log. So, this is a sum. And we do not know have nice ways to handle sums inside logs usually. We will somehow get to handling it in a minute. But it is not immediately obvious how to handle sums inside logs.

If it was a product inside logs, we know log will factor that in the sums, but then we have sum inside logs, which is where the problem comes from. Nevertheless, we will write this term by $\sqrt{2\pi}\,\sigma k$.



(Refer Slide Time: 10:41)



So, this is a complicated log likelihood expression. It is a a function of now we are going to treat this as a function of μ's, μk's, σk's , πk's and then try to maximize this. Well, what we can try is go over usual route and say that well, I will try to take the derivative of this with respect to each of the parameters of interest and then try to set it to 9 and see what happens. There are multiple problems with that.

So, the first problem is and you can try doing this, but then it is not possible to solve this analytically. When I say analytically, there is no equation that we will end up with, like $\hat{\mu}_{ML}$, earlier was just the average of the data points. So, that was a nice equation for $\hat{\mu}_{ML}$. It is not possible to solve this analytically. So, if you take the derivative with respect to μ and try to set it to 0, there is no closed form solution for the μ's.

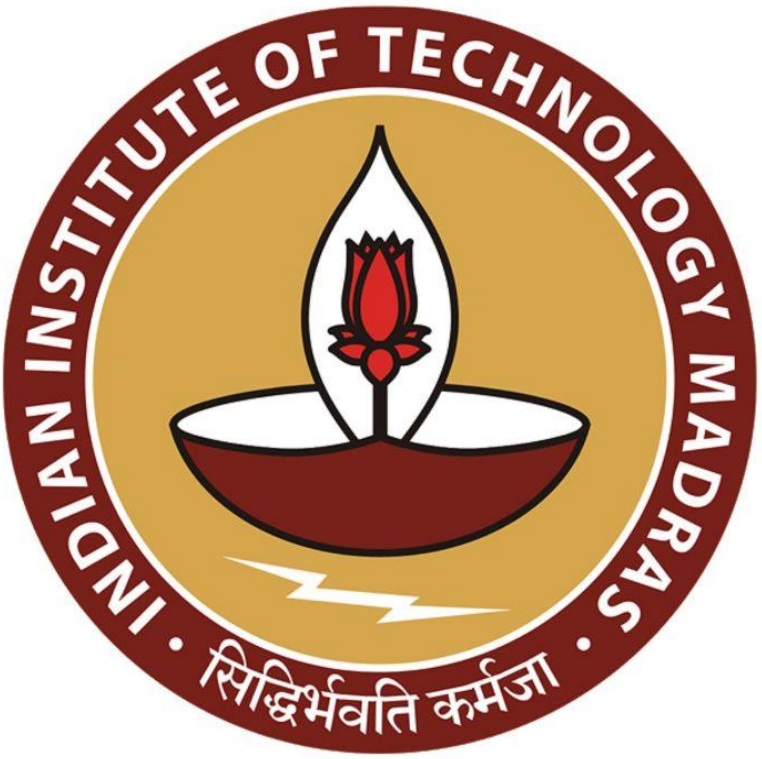
If you take the derivative with respect to π, that is a bigger problem, because π's not just are not free variables. So, they are constrained by the fact that all the π's should sum to 1. So, it is not a unconstrained optimization problem. It is a constrained optimization problem that is an even bigger problem. So, we have to take care of all that, if you are taking the derivatives and trying to set it to 0. And in general, it is not possible to solve this analytically.

Of course, for people who have seen some kind of optimization methods before, there are some gradient based approaches that you can use to write down the gradient and then do a gradient ascent, because this is a maximization problem and then try to solve this. So, that is like, common general purpose method, which works for any function and then we are trying to apply it to this particular log likelihood function, which we want to maximize with respect to some parameters. That is one way you can do it. Nobody is stopping us from doing that. You will get some estimates, μ hats σ^2 and π.

Instead, what we want to do is, we somehow want to use the power of the structure that is there in this problem. The structure that is there in this problem is that there are well defined 2 steps that generate the data. That is the structure. So, first, you have a cluster indicator, and then you generate the data according to this. But if you are using a general purpose optimization method, it does not necessarily exploit the structure variable. So, the question is, can we come up with an alternate way to solve this, which exploits the structure in a better way?

So, what we want is we need an alternate way to solve this efficiently. So, what do we want to solve? We of course want to solve the maximization of log likelihood with respect to the parameters, all the three key parameters that we wrote down earlier. What I want to do is do take a very quick detour now, and then we will discuss a few ingredients, which will be helpful for us to solve this optimization problem.

Once we see the ingredients, then we will try to see how we can apply to this particular likelihood function. So, let me call this likelihood function star. We will come back and revisit star in a while. But what we are going to do now is take a quick detour and talk a bit about convex functions, and we will see how that will be helpful in solving this problem. Once we go over that discussion.

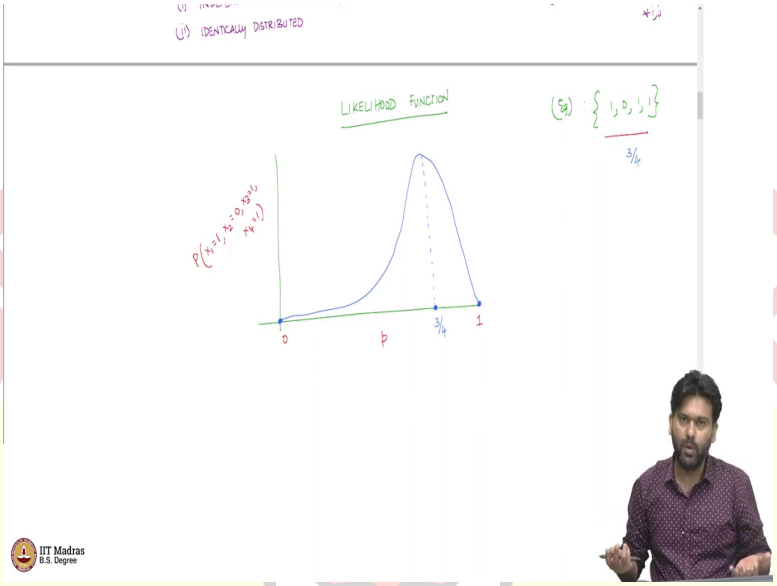


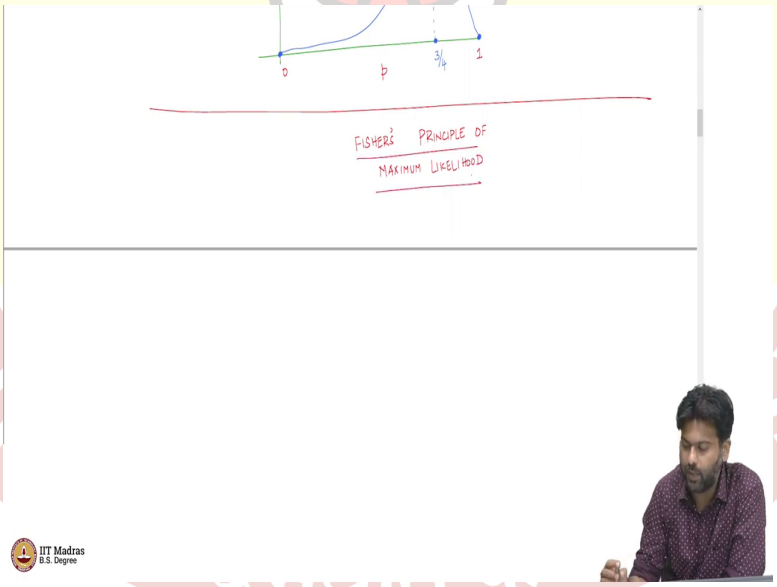
IIT Madras

ONLINE DEGREE

Machine Learning Techniques
Professor Arun Raj Kumar
Department of Computer Science and Engineering
Indian Institute of Technology Madras
Maximum Likelihood Estimation

(Refer Slide Time: 00:13





So, the question is, is there a principled way to get estimators from data and the way one could do this, is by looking at what is called as the likelihood function. The likelihood function looks as follows. So, let us take the example. It is easy to explain a smaller example. Let us say 1, 0, 1, 1 was our data.

Now, now we are going to say the x axis goes from 0 to 1. And it is all choices of P, that could have possibly generated the data that we see, which is 1, 0, 1, 1. Now for every value of

P, we ask the question, well, I see I have seen this data, if the true P was this, what is the chance that I would have seen the state? So and now we can plot that chance. So, this is the probability that your first random variable, which is the first outcome was 1, the second outcome is 0, third outcome is 1 and the fourth outcome is 1.

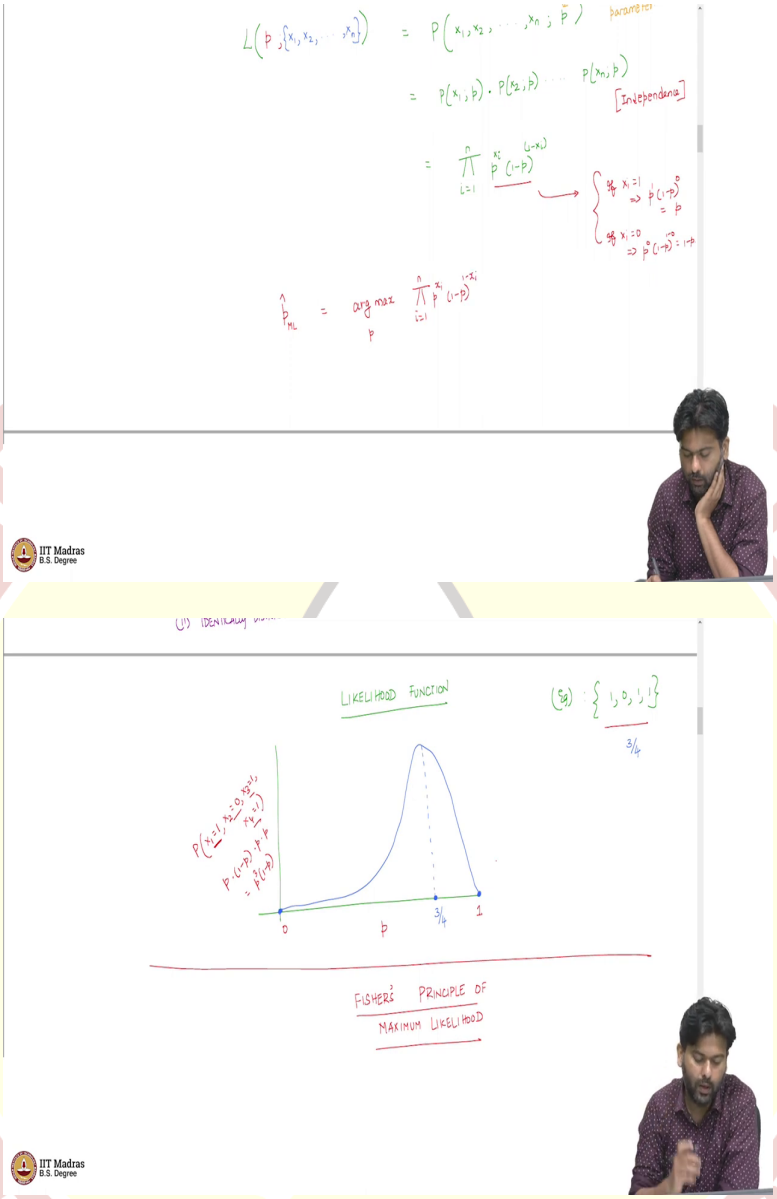
Now you can plot this. So, for every value of P, you can ask, what is the probability that I observed this data and that is what we are calling as likelihood here. And if we plot this and you could try plotting this for this example. And I tried this and then it looks something like this. So, the curve looks something like this. So, basically, it, we know that at t = 0, there is no chance that I could have generated 1, 0, 1, 1, the probability that I see the data is 0 at t = 1 also it is 0, we already know that.

For the remaining we can calculate these probabilities and then I can plot this and what I get here is something of this form. And now I can see, where does this curve peak? In other words, which is the value of P, for which the chance of seeing the data is highest? So, the data is most likely for which parameter of P, which choice of P, so and that choice in this particular case, in this example, if you try it would have been three fourths. So, which is exactly the guess that we made which is while looking at the fraction of ones.

But why should this happen? In general, I mean, what is the method here is a question to ask. So, then we will see about that in a minute. So, what we want to then do is we have a bunch of potential parameter values, which could have generated the data. And then our goal now is to pick one of those values and say, this is what I am going to bet on. This is my guess. And the way we are going to do that is by saying that, I will look at the probability that my data is generated given this the truth is this parameter and whichever P maximizes that would be my guess.

So, and this method was is quite old. So, it has been there since the fifties. And this was proposed by Fisher and it is called Fisher's principle of maximum likelihood. This might be familiar for some of you this think of this as a review, for estimation, if you have seen this before, otherwise, this is still precursor to what is going to come later. This is Fisher's principle of maximum likelihood.

(Refer Slide Time: 04:04)



So, what does that principle say? Well, it says that you write down what is known as a likelihood function L, which is a function of two things. One, the parameter that you are trying to estimate and it also depends on the data, x1, x2,..,xn. So, I see the data and I want think of this as a function of the parameter because for every value of P, I am going to ask what is the chance that I see this data? Now, how do I write this? I will write this as the probability of seeing x1, x2,..., xn. When I write this, it means that specific value that each of these x1 to xn takes.

So in the case, it would be x1 = 1, x2 = 0, x3 = 1 and x4 = 1 and so on. So if the true value is P, so if the underlying parameter is P, this is the underlying parameter. Now, how can so this is a joint distribution. So, this is the probability that all these things happen together, that

the first toss is x1, second toss is 0, third toss is 1 and the fourth toss in 1, so the probability that all four events happen together.

But because we have assumed independence, that one event does not affect the probability of the other. I can write this as P(x1,p). P(x2,p)....P(xn,p). I can do this and what lets me do this is independence. This is independence.

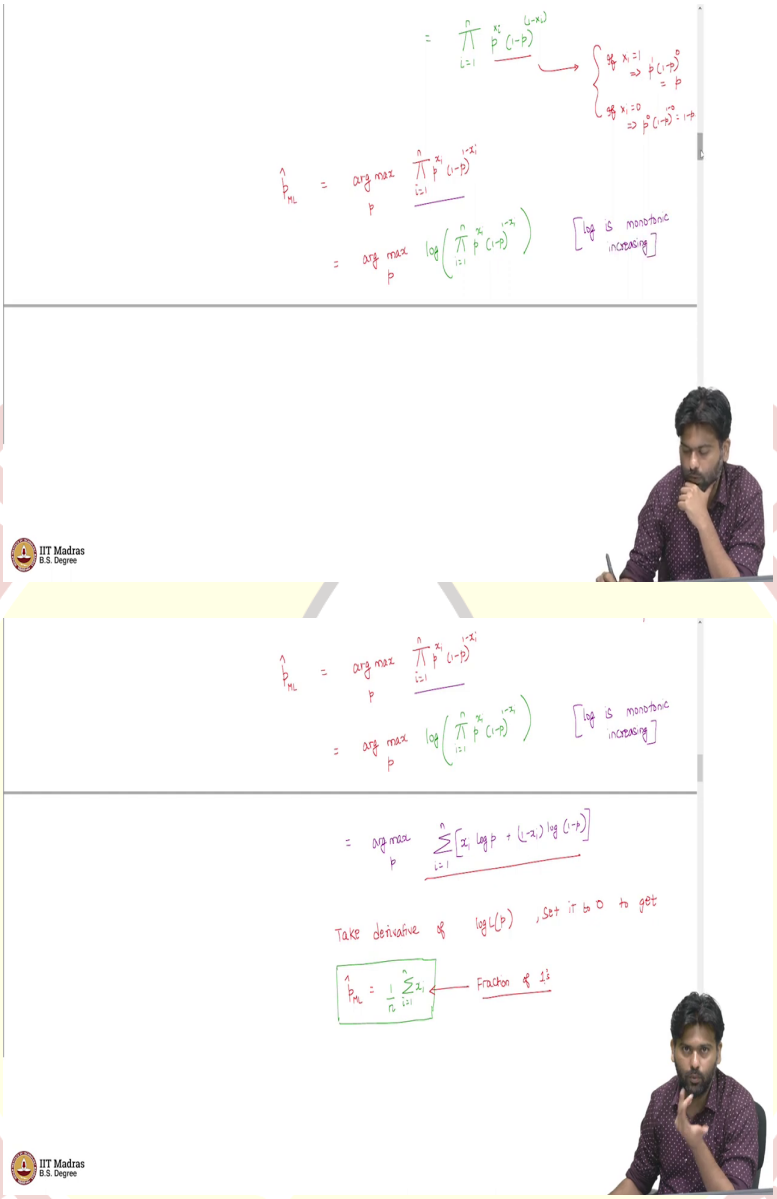
And now I will I also know that each of these coins was from the same P, so I know that. And so I can write this whole thing, as in a simplified way as product and this is the sign for product, $\prod^{n}_{i=1} (p)^{xi} (1 - p)^{(1 - xn)}$. This is just a compact way to write this thing.

So, basically, what does this tell me, if xi is 1, so, this term is what does it, when you observe a value of 1, now, if the true probability of seeing 1 as p, then I should multiply it with P. So which means that this if xi = 1, it's value is $P^1 (1 - P)^{(1-1)}$, 0, which is simply p.

On the other hand, if xi = 0, this implies this is $P^0 .(1 - P)^{(1- 0)}$, which is 1 - P, which is what we want. So, which means in for example, in this particular case, if x1 = 1, x2 = 0, x3 = 1 and x4 = 1, this would be p for this x1 into 1- p into this p into this p, which is p^3 into 1 - p. So, essentially, the plot that I have done here is that of the curve p^3 into 1 - p. So, that is precisely what we have here, in general.

Okay so, now what is the estimator, so this is the likelihood function. And now we want to define our estimator, which is our guess for p and that is where you put a hat on top of p to say to emphasize that it is a guess. And the guess comes from the maximum likelihood principle. And so this is called as an ML, so I am also writing this as ML is that argument of p that maximizes my likelihood function, which is, in this case, just $\prod^{n}_{i=1} (p)^{xi} (1 - p)^{(1 - xn)}$.

(Refer Slide Time: 08:00)



So, this is what I, this is the function that I want to maximize to get my guess for P. Well, this is a function which has a lot of products. Typically, it is easy to deal with sums then product. So, what you could do is you can take the logarithm of this function and in fact, you can look at arg max $_p$, log($\prod^n_{i=1}$ $(p)^{xi}$ $(1 - p)^{(1 - xn)}$)

Now, remember, this is fine, because logarithm is a monotonically increasing function, which means that if there is a $p^$, which maximizes the original function, that means that the functions value at $p^$ is greater than the functions value at any other point.

Now, if I take the logarithm of it, because it is of its monotonicity the log of the functions value at $p^$ will be greater than the log of the functions value at any other point. And so it is

okay if I either maximize the original function or the logarithm of it, the point where the maximum occurs does not change, So, because log is monotonic function, increasing function I can do this.

Now, things become simpler. Now, this becomes arg max $_p \sum^n{}_{i=1}$ and that is the power of log products become sum this becomes xi log p+ (1- xi) log (1- p).

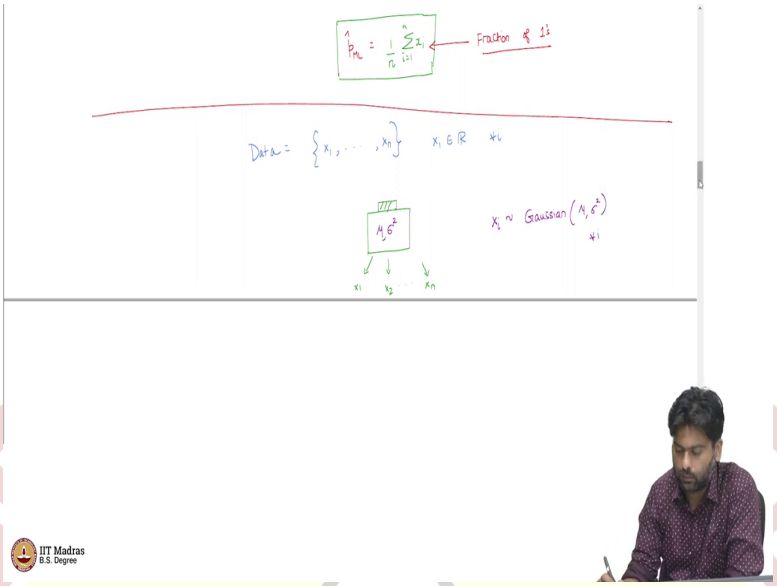
Now, this function happens to be a nice concave function. Do not worry if that you have not seen that term that is fine. So, this is a function which has no you can maximize this by taking the derivative and setting it to 0.

So, now, this is take this as an exercise take derivative of this function, treat this as a function of p, derivative of L (p). In fact, log L (p), treat this treat the likelihood functions is the log likelihood function, set it to 0 to get our guess $\hat{p}_{ML} = 1/n \sum^n{}_{i=1} xi$. If you did that, if you had taken the derivative, set it to 0 and I urge you to try that out, you would get that the answer as simply the average of my data points.

This looks like the average but then it also has a simple interpretation because my xi is just 0 or 1 heads or tails out of the sum, what contributes to the sum, it is just the ones not the zeros. So, basically, the sum counts the number of ones in my data. So, this is simply the fraction of ones.

So basically, here is a method which is called the principle of maximum likelihood, which seems to give us reasonable guesses. So, my guess for the true p, by looking at the data is that value of p which maximizes the chance that I observed this data. And that has led us to guessing the fraction of ones as the, which was also our intuitive guess. But remember, we have implicitly used two important facts, one is independence of the trials. The other is identically distributed nature of the trials. So, now this is one example.

(Refer Slide Time: 11:53)

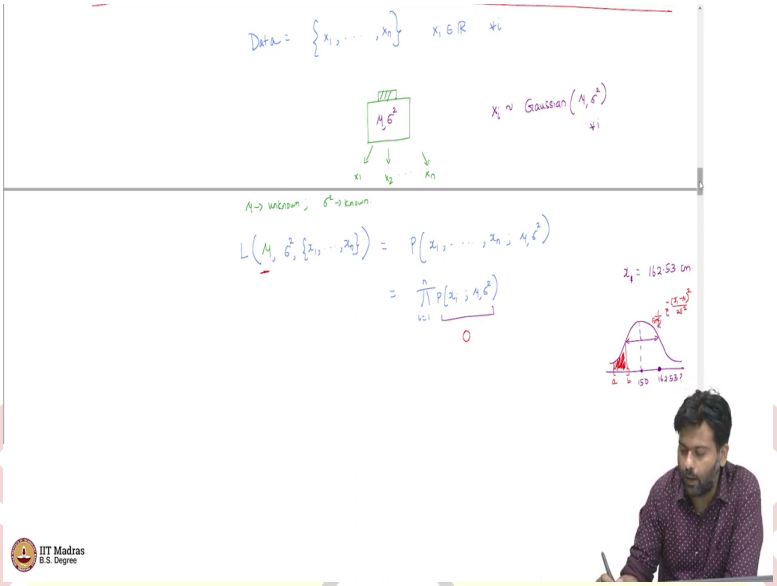


Now, one thing to keep in mind, when we are doing this is the following. So, let us say we now have data for different form. We still have x1 to xn, but x1, xn are no longer zeros or ones, let us say xi belongs to real numbers for all i. Let us say I collect the height of a bunch of 100 people. And then I want to reason about that in a probabilistic sense.

Now, the height cannot necessarily be 0 or 1. So, it can be any value. And so I cannot use the previous model. If I use the previous model the box with a coins inside that, that is not going to be useful for explaining this data, because that box can explain only data which has zeros and ones. So, I need a different box to explain this data, which means I need a different probabilistic model that I need to assume to explain this data.

And the natural model, in this case, would be something like assuming the box is still have a box, you still have a box with a button. And then our data comes from it. But because our data can be any real value, in this case, we want a box that can explain real numbers, when we are going to assume the most simplest thing would be to assume that the box has a Gaussian random variable with some mean μ and some variance σ . So, that is I am assuming xi is Gaussian with mean μ and variance σ^2 for all i. That might be another reasonable assumption to explain this data.

(Refer Slide Time: 13:31)



Now, if I do that and if I now let us say we again, write the likelihood function like how we had written earlier, which is now for simplicity, let us say, we know the variance that generates the data. It is not true in general. So, if you have a bunch of data points, you would not know anything about the data. But let us make it even simpler and say that we know the variance. So, the only thing that we do not know is the mean. So, let us say mean is the parameter that we are trying to estimate. So, mean is unknown. Say variance σ^2 is known.

So, now what is the likelihood function look like? Well, it is a function of μ, σ^2 , the parameters. I mean, one parameter is known, one is unknown, but still, it is a function of these two, these two and of course, the data x1 to xn. I see a bunch of data points. Now, what is this likelihood? Well, the way we have put down likelihood earlier is that this is the probability that the joint distribution of x1 to xn parameter is by μ and σ^2. So, of course, I can write this as a product by independence of i = 1 to n P (xi ; μ_1,σ^2)

Now, what is this probability? So, let us say x1, the first height that I measured was 162.53 centimeters. Now, let us assume some Gaussian distribution with mean 150 let us say. So, let us say this is the Gaussian PDF, here is 150, some variance that I know, now I am trying to get the mean. So, let us say I start with my guess as 150.

Now and I am asking, what is the chance that if I draw a sample, according to Gaussian distribution with mean 150, what is the chance that I am going to get 162.53? That is the question that we are asking here. So, if we define our likelihood like this. But if you remember, Gaussian is a continuous distribution. So, you can, if you ask for the chance that a

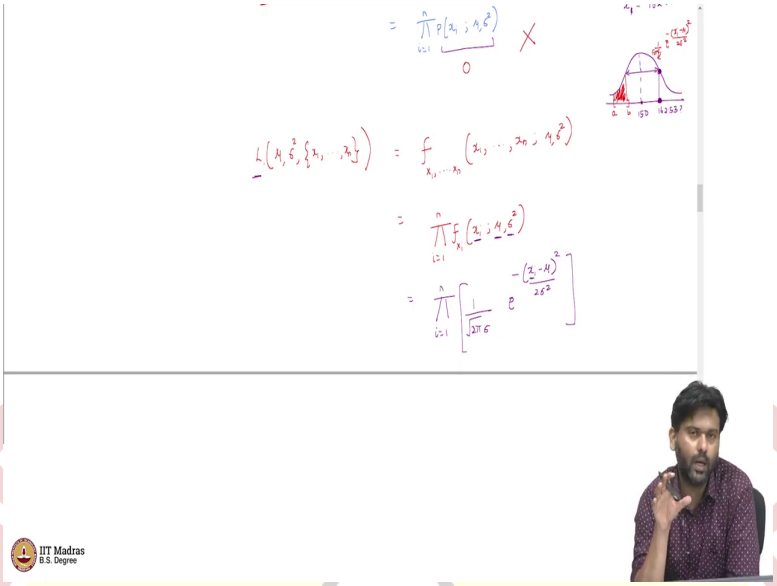
random variable that you sample according to a Gaussian distribution falls in an interval a to b, then we know that that chance is just given by the area under this curve of the PDF.

So this PDF, if you remember, is $e^{-(xi - \mu)2 /2 \sigma 2} .1/\sqrt{2\Pi}\sigma$, all that should we know. So now, this is the chance for an interval, but then we are asking what is the probability of a specific value that we have observed. So, I stopped, the person, asked for his or her height and then I got the value is 162.53.

Now, I want to ask the probability that I see this particular value with some Gaussian with some mean μ. Now, because Gaussian is continuous, this value is going to be 0 for any individual point or a bunch of points. So, only intervals will have non-zero values for continuous distributions. And so no matter what my μ is, this is always going to give me a zero value. Because I am asking for the probability of a bunch of points to be generated according to any μ any Gaussian with any μ however, away, it might be from the data, or close it might be to the data, it is still going to be 0.

In other words, this function, the way that it stands now as a product of probabilities, will give me 0 value for all possible choices of μ, which means this is not able to distinguish one μ from another, so it does not have the capability to distinguish one μ from another. So, it might not be useful in making good guesses. So we cannot really, there is nothing to maximize here, because every μ gets a value of 0. So, all μ's are equally bad. So, you cannot really use this function to make a guess.

(Refer Slide Time: 17:31)



And the problem arises because of the fact that Gaussian is continuous and individual points get 0 probabilities. So, Fisher's proposal was to not use the probabilities. And instead, replace the probabilities not to this and instead define the likelihood of the parameter in this case, as not the product of the probabilities, but then instead the product of or in general, the value that these parameters take these observations take is given by the PDF of the distribution that generates the data. So, you replace probabilities with PDFs.

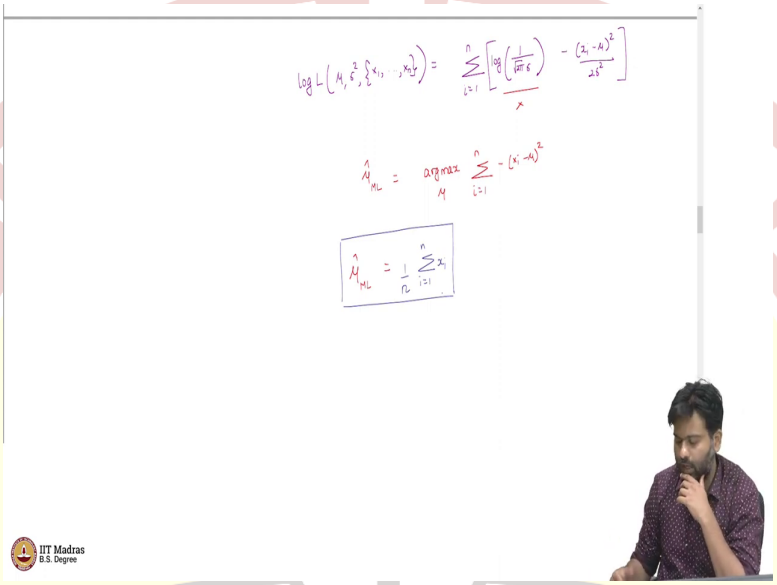
In other words, you want to replace the x probabilities with the joint PDF x1 to xn of x1… xn parameterize by μ_1 , σ^2. Now, PDFs behave similarly with respect to probabilities for a lot of cases. In fact, if the data points are independent PDFs will factorize.

So, this can be written as i = 1 to n, f of xi parameterize by μ and σ^2. What does this mean? This means that I am looking at the Bell curve's value at xi. So, I am not asking for the probability. The probability of seeing this point is 0. But then the value that this point x has given me the PDF is not 0. That is some non-zero value. It may not even be between 0 and 1, but it is some non-zero value that I can use as a proxy in some sense for the probabilities. That is the proposal of Fisher.

And in fact, what is this going to be we know what this is, so, for the Gaussian. So, this is asking if the true mean is μ and the variance σ^2, what is the density at a particular value i and we know that by definition is $.1/\sqrt{2\Pi}\sigma\ e^{-(xi - \mu)2\ /2\ \sigma2}$. So, the same μ and 2 σ^2but then you are multiplying it over different xi's. So, that is what this essentially means. So, this is our new modified likelihood function.

In fact, that is why we call it to the likelihood. So, we do not call it the probability function. If it was always using probabilities, why give it a different name? You giving it a different name because it is not strictly probabilities, so of observing the data, it is the likelihood of observing the data. That is what Fisher called it. And we want to maximize this likelihood function.

(Refer Slide Time: 20:12)



Now you do the same drill. This is a product of a bunch of things. So, you take the log of the likelihood function of μ σ^2 x1 to xn. I will just do this for completion only one example. So, this is $\sum_{i=1}^{n} \log (1/\sqrt{2\Pi}\sigma)$. Now, this is logarithm to the base e. So, you can do this as $-(xi -\mu)^2$, the logs in the exponential cancel. And this looks like this.

Of course, we are trying to maximize this with respect to μ because we want to find that value of μ that maximizes the likelihood. So, this term does not have mu, so I can remove this term. It is not going to contribute to my maximize, in other words, I could might as well maximize $\sum_{i=1}^{n} -(xi -\mu)^2$. I can even remove $2\sigma^2$ because that is I am assuming as a constant rate, so that does not really affect my maximization.

So, I want to find that arg max, over mu, which is $\hat{\mu}_{ML}$. So, that is going to be my guess, which is an arg max μ, the negative of this, which again you can take the derivative, set it to 0. And please do that. Try that out. It is just two steps. This again, happens to be 1/ n $\sum_{i=1}^{n}$ xi.

What does this tell us? This tells us that if you have a bunch of data points and this is the important part that if you assume that this data is generated according to a Gaussian, with certain unknown mean mu, then the best guess, according to the maximum likelihood principle for this unknown mean, mu, is the sample mean.

Now, it might feel that well, what is the big deal. So, we could have any way taken the sample mean and it would have been a good guess for the unknown parameter. But this is a good guess, the sample mean happens to be a good guess only, or not I should not say only, in this case, where we are assuming that the underlying data comes from a Gaussian.

If I change that assumption, to a different distribution, maybe I have some domain knowledge, which makes me believe that the data is not coming from a Gaussian, but then it comes from, let us say, a Laplacian distribution, for example. That is a different probability distribution, which looks like a Gaussian, but then it is more sharp at the peak.

Then if we do a principle of maximum likelihood, it is no longer going to give me the sample mean. It will give me something else. So, in fact, it will give me the sample median which means to say that you are estimator is very closely tied with the probabilistic model that you assume that generates the data.

So, it is always the sample mean, may not be a good guess. It depends on what model you believe generates this data. If you believe in a different model, the guess is going to be different. And our method adjusts to the model, so because the method essentially tries to use the PDF of the model that generates the data. And so it works well for the model that we use. So, that is the point that I just wanted to mention.

And in general, this is a very general purpose method. So, it does not make it is a simple idea. So you put down a model, whatever model it might be, you put down that model, right down this PDF of that model, or the, if the discrete distribution, the probability of that model and write down the likelihood function of the parameters maximize it. So, it is as simple as that. It is not just a simple it is as general as that that is more important. So, it is super general, in the sense that once you have the model, you have a method to get an estimator.

So, these estimators may not necessarily be very intuitive apriori. In this case, perhaps it is intuitive. In the previous case of the coin toss, it was perhaps intuitive. It is good that it matches our intuition when we do have intuition about what might be a good guess. But even if we do not have intuition about what might be a good guess, the method is robust enough

that it can give us some estimates and that is why this is a very super popular method in general, in statistics and machine learning. Now, so this is one side of the story.

So, this is the summary so to say of the principle of maximum likelihood. Now, of course, if you are in a statistics course you will try to understand why is this a good estimator? Maybe somebody smart, tomorrow might come back and say that, hey, I have another estimator, which is a better estimator. And then you have to argue why maximum likelihood is a better estimator than what they might have and so on. And that is what a statistics course will do. So, they will try to argue good properties of the maximum likelihood estimator and in cases where it can be in general it will give you very good estimators.

So, and in some cases, you can argue that you cannot get any better than a maximum likelihood estimator. So, in certain well defined ways, it might, you can argue that it is the best that you can ever get and so on. But because this is not really a statistical course, we are not going to dwell deeper into the specific properties of the maximum likelihood estimator itself, we will agree that it is a good estimator, it typically is estimated that gives you reasonable guesses. And what we want to use is somehow ask the question.

So, this is a good estimator given data, it gives me good estimators. Is there something else that might be typically available in practice? And if so, is the maximum likelihood estimator still the best thing to do or is there something else that one can do? Of course, I am being vague here. So, let me make that a little bit more precise. And then we will try to see how you can potentially come up with different estimators, which might be in some cases better than the maximum likelihood estimators.