NPTEL

NPTEL ONLINE CERTIFICATION COURSE

Introduction to Machine Learning

Lecture 37

Prof. Balaraman Ravibdran
Computer Science and Engineering
Indian Institute of Technology Madras

Parameters Estimation II: Prior and the MAP estimate

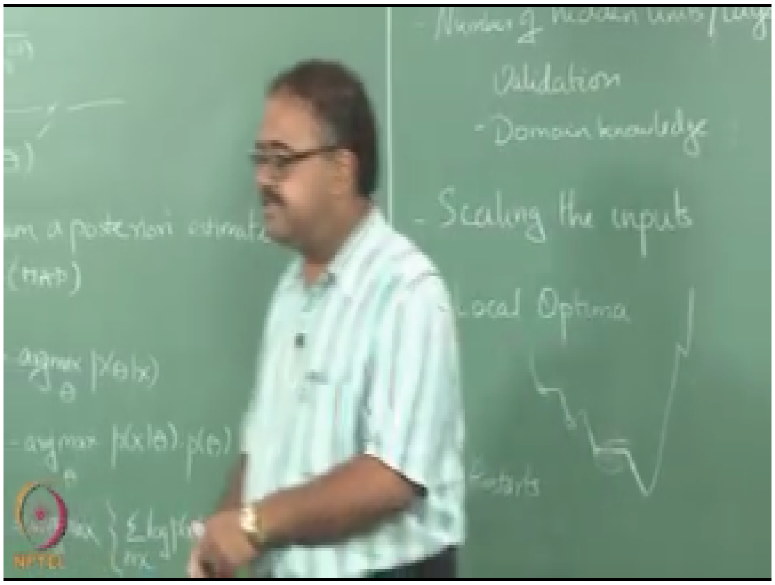
Ok so far we assume that the whole motivation for doing maximum like hood was hey I did not know anything about the parameters .Before I started the experiment right before I gather the data I did not know anything about the parameters suppose I did know something about the parameters. So what could you know what the parameters just stick with the coin tossing experiment give me something from the coin toss case yeah well yeah I know the high probability it is fair.

If you know it is fair or not that is not a prior information that is insider trading okay so with the very high probability and think it is it is fan right so he hands mean coin we look at his face I mean obviously he is not going to cheat me right. So I will assume it is a fair coin to begin with right so what I can do is I can have a prior probability of it being far being very high right now I can in fact think of having a Gaussian with the with a peak at point five.

Form row right I can think of having probability of 0.540.5 right 40 being point size there are two probabilities here all right do not get confused as a probability of the coin coming up heads and there is a probability of that being point file right. So that is the prior probability we are talking about here and I am saying that I can think of it as Gaussian and our Gaussian is a good idea why not great probability is not only problem it can be even greater than one also mean either side is a problem right.

So what is a good distribution to do is useful you already seen that in your tutorials probability tutorials. Distribution that is γ distributions limited between 0 and 1 in fact it seems to have been invented for putting priors on probabilities. Right in fact it was so you can think of that as your prior right so I have some information about right I want to use that in my optimization right.

(Refer Slide Time: 02:43)



So these are called alright so we looked at the maximum likelihood or m l here now we are going to look at maximum a posteriori or map right so this is a priori information right is this prior information about as that is the posterior information about theta right. But are we actually computing the posterior here we do not know it right so we will have to see that anyway.

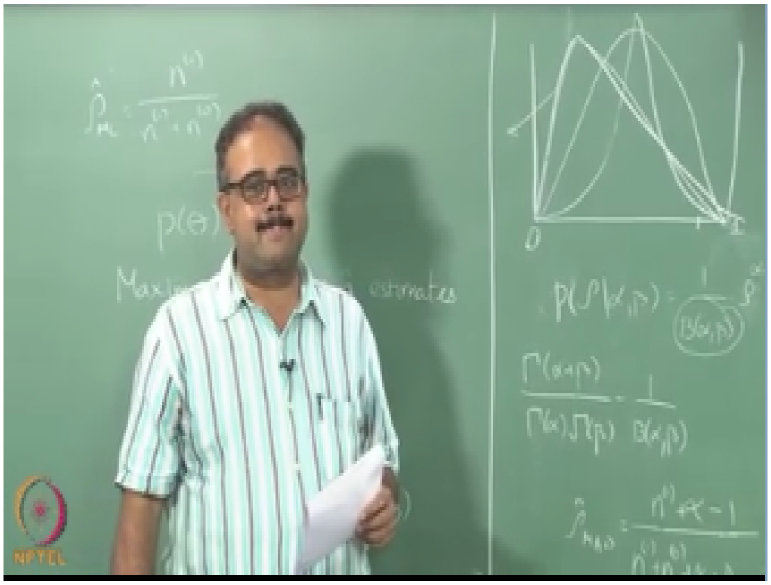
So what we are interested in is finding out that gives me the maximum posterior right so if you think about it I do not have to actually compute the posterior to find out the that gives me the maximum posterior why because X is common I can ignore that I can only down arc max on the numerator I do not have to do the arc max on the denominator right so I do not actually have to compute the posterior.

So as long as I have a convenient form of just dealing with the numerator I am happy right so i can just do the max over the numerator and of course I can take the logarithm because this is a nasty term because it has a product in it so I take the logarithm converted the summation and basically do the max of this.

Right so there are a couple of things which I want to point out here so the one is yeah if you have some prior information you can use that right. So if you believe in the honesty of the person you can use point five right but there are other cases where I do not really have a prior information about how the true solution will be or what the true solution is right but I have some prior over what I want the true solution to be you have cases like that.

We could try to do this when we did rigid aggression or when we did lasso right we wanted the parameters to be small the way I achieve that was by putting a quadratic penalty right instead of that what I can do is I can say that hey I have this prior which is very low probability to high values of the parameter right I can I can make this prayer you know is all of you know but γ distribution I can have all kinds of weird shapes to the γ distribution.

(Refer Slide Time: 07:01)



Right let us see nice prayer right the honest prior right so is a really honest prior and that is something like the l1prior l2 prior sorry wait so the probability of the higher value is γ being high is small a row probability of row being high is small probability of row being small is high right so this is one way of thinking about enforcing regularize right does it make sense so I can use the priors for enforcing my regular the second say no do not give me things

That have very high parameter value of interested in smaller parameter values and if you this is just about single parameters if you want to talk about multi dimensional case you can also say that hey I will give you a low probability half having a solution which has more than thirty percent of the parameters nonzero so what will that enforce for me varsity right that will enforce varsity so then the probability.

So if I need to have a hence equation suppose I really need to have my row here my row is actually there right but I start off with a prior that looks like this will I will I reach my current estimate for row somebody said it depends so I'm happy depends on the amount of data I have right so it depends on the amount of data I have so if you have an infirmity prior grade the amount of data that you actually need is actually low.

If you have prior is correct right if you put the maximum probability on the right solution the amount of data unit is low but if you put these the prior the maximum weight on the pin the price on the wrong solutions the amount of data you need is going to go up significant amount of data is mean is going to go up significantly so I said to I made two points about prayers right so this remember what are the two points price can be used for regularization okay so wrong prayers need more data to corrector and a completely bullheaded prayer.

Can never bright so what is this it is actually the γ distribution where I have written the normalize in a slightly different format you are used to seeing thenormalizer as okay so what is this call that is the γ function okay this whole thing is the γ distribution so you actually have three γ is here so the γ distribution okay and the γ parameter lowercase γ as a parameter and then you have a γ function thing okay.

So the thing to note here is that you are and your γ parameters right almost act like as if you have seen heads and tails right. So your α is just increasing the count of your head s right and the γ is increasing the count of your tails the make sense right α is just increasing the count of heads. So if I had done the actually done the experiment I would have seen n1 heads right, but I am assuming I in addition.

I saw α-1 heads also right so if I am going to have a prayer like this right can you imagine what would be the values of α and γ as I will be less γ will be more because α adds to the heads right so Rho would be higher if α is larger row would be higher so if I am going to skew it like this so they should have a γ larger than α you can start reasoning about all of these things just if you understand what is happening so these things are sometimes called pseudo counts.

IIT Madras Production

Funded by
Department of Higher Education
Ministry of Human Resource Development

Government of India

www.nptel.ac.in