NPTEL

NPTEL ONLINE CERTIFICATION COURSE

Introduction to Machine Learning

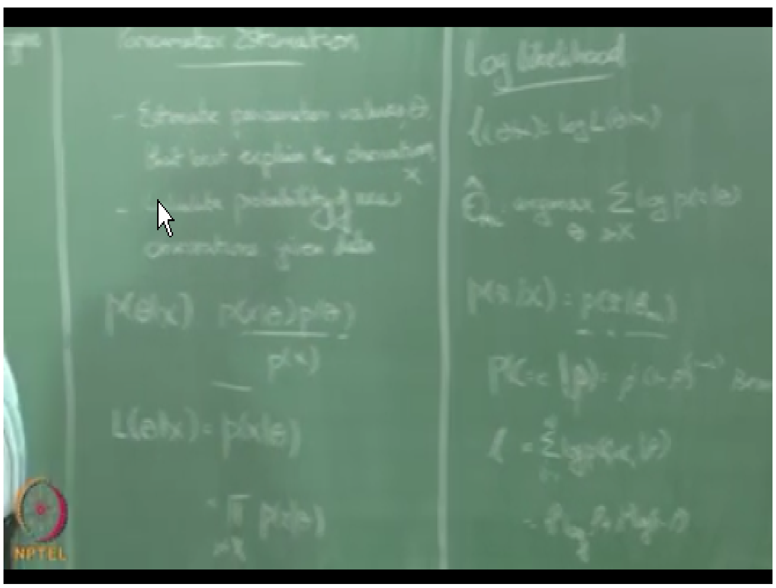
Lecture 36

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

Parameter Estimation I: The Maximum Likelihood Estimate

More of a Bayesian approach to parameter estimation,

(Refer Slide Time: 00:19)



so I would say that there are two goals to us we want to estimate the parameter values right. That best explains some given data to us right. We already looked at this in the context of logistic regression right, so we assume some kind of a model that was generating the data and then we estimated the parameters of the model that somehow best explains the data right.

And we used the notion of a likelihood right we are just going to look at it again little in detail. And then go on and look at a couple of other ways of doing parameter estimation. The second thing we look second problem that we are interested, so essentially calculating the probability of new observations given the old training data right. So I am going to assume that the parameters are given by θ right.

The observation is given by X, so what I am interested in is per probability of θ given X right. That is all the familiar Bays rule; so what is that what is that prior, likelihood right. So I only looked at it right, so I am going to like write the likelihood does likelihood of θ given X right. So if I write it like that people get a little confused, so likelihood of the data right. I already mentioned this is likelihood of θ not data.

Its likelihood of the parameters given the data even though we write it as probability of x given θ. Is it clear? Why is it a function of θ_0 of X? Because X is fixed in our context X is fixed right. I have given the observation X right I am interested in finding the parameter values θ. So the way I am going to set this up for a given X okay, for different θ. What is the probability of that X? Let us say I can consider five different θ for θ 1 what is the probability of x for θ towards the probability of X for θ 3what is the probability of X and so on so for.

So that gives me the likelihood of θ, sometimes people say the likelihood of X with respect to θ. In fact it is so widely used I do not know if it is even right to say it is incorrect anymore. Like pre pone a meeting, but so why are we interested in the likelihood why are we interested in the likelihood? Yeah! So, what I am really interested in is to find that θ that best describes the data right.

So essentially what I am interested in is finding the θ that has the maximum probability here. That given the X which θ has the maximum probability right, so the X is fixed right so this does not matter? And if I really do not have any information about what θ is the best θ to start off with right. This is also irrelevant because it will be the same for all the θ correct, so if I want to maximize P(θ) given X all I need to do is maximize P(x) given θ.

Because this is constant that will also be constant across all θ all right, so it is enough if I look at likelihood right. So if I make an assumption that we make the assumption that my all my data samples are generated independently right as well right my likelihood as the product of the individual probabilities. Then we do not want product or probability so we typically end up using, I agree with that.

So suppose a new data point the X tilde comes what is the probability of X tilde with respect to X. Sorry, I mean given X sorry given that I have already been given some training data X okay. I am asking you what the probability of this new point X tilde is. What will be the probability right. In fact is exactly what we did in the logistic regression case, we found the maximum likelihood parameters for β maximum.

Likelihood estimates for the parameters β and then we plug them back in and say okay. This is how you estimate the parameters right. So let us look at a simple example, it is considered a simple coin tossing experiment right, so there is a random variable C okay. Which has some outcome lowercase C right, so lowercase C if it is one in his heads, if it is 0 it is tails okay. What is the parameter that I have did not coin tossing experiments probability of coming up its okay. Let us change the symbol still looks like P but it is a ρ so the probability of whether you come up with heads or tails right.

Given the parameter ρ okay, so what is the probability that should also look familiar? We already saw that right in the context of class label being 1 and class payable being 0 right. And the probability of coming up class 1 versus probability of coming up class 0 right. It is like heads and tails. Now the probability of coming up heads, well ρ power C is 3 well. If its 1 then it is ρ if it is 0which is tails since 1 -ρ okay.

So this is the expression in simplified form, I always have written I would have to write it as ρ if C=1-ρ if C=0 right. In stuff that I can write this using a selection function so what is this probability density called? Bernoulli yeah! Okay. The Bernoulli is so what is the likelihood when look like for each of the I^{th} toss you would have an outcome lowercase C_i. So what is the probability that the random variable can be lowercase C_i given the parameter ρ right.

And some this over all the right, so there are N 1 times head says occurred. Let us say and N_0 times tales has occurred right. Then I can simplify the summation as N_1 times log ρ +N_0 times log 1-ρ right simple enough. Now take the derivative of the likelihood equated to 0, and tell me what ρ is right. So our common sense way of estimating probabilities from experiments is what toss coin N times find out number of times it turned up heads okay. Divided by N that gives you the probability of it coming up heads. And a turn out that is the maximum likelihood estimate assuming. That your coin is obeying a Bernoulli distribution.

IIT Madras Production

Funded by
Department of Higher Education
Ministry of Human Resource Development
Government of India

www.nptel.ac.in