

NPTEL

NPTEL ONLINE CERTIFICATION COURSE

Introduction to Machine Learning

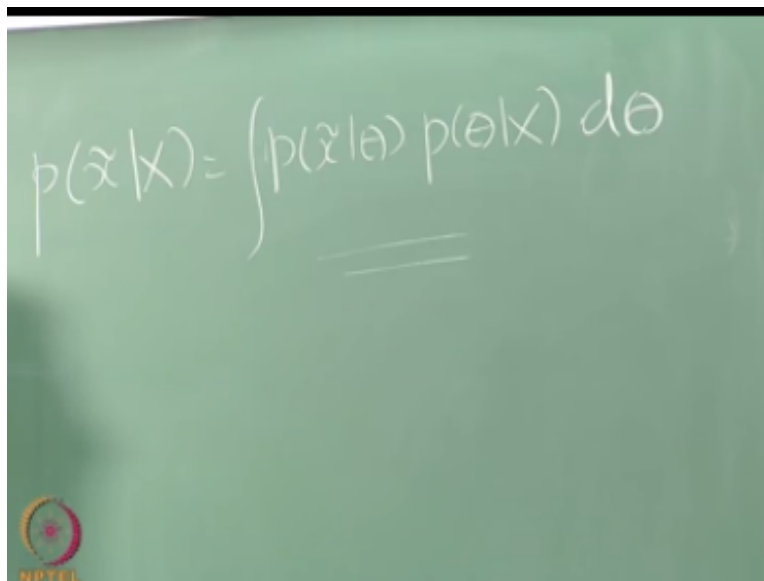
Lecture 38

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

Parameter Estimation III

So there is one thing that we are doing here see if you remember what was our two stated goals which I have erased of the board, what are the two stated goals that we had? One was to find some parameters that best explain the data, what is the second goal? Exactly right I am doing the best I am actually finding the best parameter like one single setting for the parameter that best explains the data that was given to me that is what we have been doing so far. But in terms of finding the best prediction for a new data point am I doing the right thing so far is it the right way to do it right, so if you think about it right.

(Refer Slide Time: 01:01)


$$p(x|X) = \int p(x|\theta) p(\theta|X) d\theta$$

So probability of $X \sim$ given X , it is a probability of $X \sim$ given θ times the probability of θ given X summed over all θ if θ was a discrete probability distribution but since we have

been considering Bernoulli and other things it is going to be an integral over all θ right I am not talking about the outcome I am talking about the parameterization so the parameterization is a continuous parameter right, θ is a continuous parameter.

So I cannot sum over θ it is not like I am only considering $\theta_1 \theta_2 \theta_3$ I am considering θ in the interval 0 to 1, right? So is integral over θ , right. So this is this is the actual outcome right but think about what happens in the case of the prior or any one of these cases right I am only picking one θ it could very well be that there is another θ which also has a high probability of being correct.

But since I am picking only one θ I am sticking with that, there could be two different θ which I could have used them right in fact I should ideally be using all the θ because for a certain parameter setting some $X \sim$ might have a high value right, so even if that probability of that happening is very small and I should still be accounting for that in my prediction, that makes sense why this is a much better predictor than using ML map. But why do not people use this then computationally hard, why?

(Refer Slide Time: 03:21)

$$\begin{aligned}
 p(x|\theta) &= \int p(x|\theta) p(\theta|x) d\theta \\
 p(x) &= \int_{\theta} p(x|\theta) p(\theta) d\theta \\
 p(\theta|\mathcal{G}, \alpha, \beta) &= \frac{\left(\prod_{i=1}^N p(c_i|\theta) \right) p(\theta|\alpha, \beta)}{\int_{\theta} \left(\prod_{i=1}^N p(c_i|\theta) \right) p(\theta|\alpha, \beta) d\theta} \\
 &= \frac{p^{\alpha} (1-p)^{\alpha} \cdot p^{\beta} (1-p)^{\beta}}{B(n^{\alpha}, n^{\beta})}
 \end{aligned}$$

So far I was actually trying to avoid computing probability of θ given X right here I did that by assuming everything else was constant right and I just had to do the likelihood right here I said okay I am just doing a point estimate so I can ignore the denominator I can only do the numerator right but when I go here boom, I have to do the full computation right this essentially means I need to know $p(x)$ right.

And that becomes hard but computing that is actually harder to actually multiplied over all the data points that you have right so it becomes a little tricky right and what is $P(X)$ by the way, no yeah but what is $P(X)$, $p(x)$ is a probability of seeing the data right what does the probability of seeing that I do not know that I have only given you the data right I do not I do not know the distribution from which the data was drawn that is exactly what you are trying to do so the θ gives you the distribution over which the data was drawn right, so what would be $P(X)$ right so that is $P(X)$, how do you compute that? Good point, so whenever we talk about parameter estimation right so you need to have some parameterize form of a function for you to do the estimation of the parameters right.

So if you remember in the logistic regression it was not Bernoulli it was the logic function that we were trying to estimate the parameters for and I also told you in when we looked at LDA I told you we could make a lot of different assumptions about the parameters in the LDA we made an assumption what did we assume it was a Gaussian right, anything else? Covariance was the same right this is for LDA right.

And I at that point I told you could use mixture distributions as well and you could use whole bunch of other things I also said you could use nonparametric techniques right but I told you it is a misnomer it is a misleading name because nonparametric really means that you just keep adding parameters and things like this so there is a very flexible very powerful modeling paradigm so you could do parameter estimation for nonparametric methods also right.

Where you have to actually figure out how many parameters you need as well so then the distributions you consider become more and more complex now we are looking at very simple forms right but the distributions become more complex and it is the parameter estimation consequently becomes harder right, so in fact most of machine learning research nowadays is essentially on parameter estimation for all kinds of different things like non parametric models how do you do the parameter estimation things like that lot of research is going into that. And a lot of powerful models have come out.

Let us go back to our Bernoulli case for a minute right I have Bernoulli and my prior is a β distribution right now I can try to do this so this is this will be what, this $P(x)$ given θ right that is $P(\theta), p(x)$ given θ is $p(\theta)$ divided by make that makes sense right this just the is $p(x)$ right this is $P(\rho)$ given X right so X is your C , C is the set of experiments that we were talking about that right so $P(\rho)$ given X is equal to $P(x)$ given $\rho \times p(\rho) / p(x)$.

Next $p(x)$ is given by integral over the entire row space which is 0 to 1 okay $p(x)$ given $\rho \times p(\rho)$ d ρ and it is just the base rule I have it now, right. Towards one thing you notice here what is gone here or nice convenient logarithms are gone right but it does not matter too much why? We are not doing any maximization here we are not have to take the derivative or anything now right and this actually interested in computing this whole functional form again and I am not interested in taking derivatives and trying to maximize.

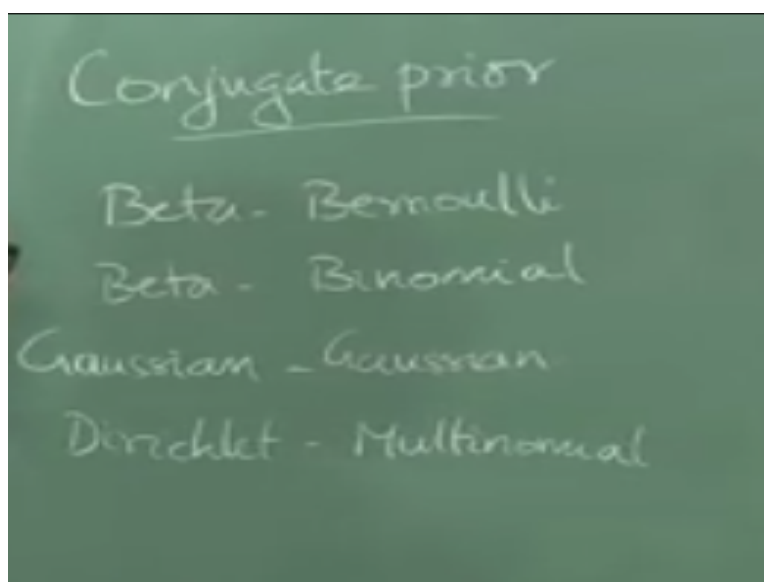
So it is okay if we look like if I do not have logarithms but it just makes the whole thing more of a nasty right, when it turns out this is pretty easy to compute well so I skipped a few steps in between but you can figure that out so I wrote out probability of ρ given $\alpha \beta$ which is essentially this right and this product I can write like this as we did earlier that we have done both of this before so what I have left out here is a normalizing function.

There should be a $1/\beta$, β function of $\alpha \beta$ right and then I have this integral also and it turns out that this whole thing including that normalizing factor is actually equivalent to the β function of $n_0 + \alpha$ I mean $n_1 + \alpha$ and $N_0 + \beta$ okay. These are P's these are ρ 's yeah if you remember the β function right it is $\rho^{\alpha-1}$, $(1-\rho)^{\beta-1}$ and the normalizing factor is a β function of $\alpha \beta$ right, so it is $\rho^{n_1 + \alpha - 1}$, $(1-\rho)^{n_0 + \beta - 1}$ and then there is the β function of $n_1 + \alpha$, $n_0 + \beta$ this actually itself a β distribution right.

So it is exactly the same β distribution so we started off with the β distribution as the prior over the ρ right and then we did this computation and the posterior turned out to be a β distribution as well is very convenient right? Such distribution which allowed us to do this are known as conjugate pairs that are conjugate distributions so what are the two distributors are talking about here β and β and Bernoulli right.

So the data distribution was Bernoulli prior distribution was β right if that is the case then the posterior will also be β right people know the difference between Bernoulli and binomial, what is the difference between Bernoulli and binomial? Single trial is Bernoulli repeated trials is binomial right it turns out that β is also conjugate prior for binomial, right.

(Refer Slide Time: 15:12)



Any the famous conjugate pairs that you guys know? So both the data and the prior can because it so remember what we mean by the prior distribution right so the prior distribution is the distribution over the parameters of the data distribution when I say Gaussian-Gaussian that means that okay I am assuming my data is coming from a Gaussian and I am going to assume that the mean of the Gaussian is coming from another Gaussian right.

The probability of the mean is going to be given by another Gaussian so that is what I mean by a Gaussian-Gaussian prior like the like in the β Bernoulli prior I am assuming that the probability of heads is ρ and the prior distribution ρ is a β distribution, so when I say Gaussian-Gaussian I am assuming that the data is coming from a Gaussian distribution and the mean of the Gaussian is coming from another Gaussian distribution that is what Gaussian – Gaussian. There is also another very famous and so deliciously so people about multinomial is what is multinomial?

It is a distribution that will describe multiple roles of a dye for example, binomial is when you have two outcomes multinomial is when you have multiple outcomes right, so the single experiment single trial version of multinomial is called not too many people know it and multinomial, binomial is called Bernoulli the single trial of a multinomial is called know the unimaginative name is called the discrete distribution okay.

So but so multiple trials is called the multinomial distribution and the prior the conjugate prior for it is an original a distribution which is nothing but the multi-level extension of the β distribution so it is like the β distribution when it is a multi-dimensional extension of β distribution okay and there are a bunch of others okay so but there are several that are known and so typically what you do is you look at your data right look at the data and figure out what distribution is a good distribution for the data right.

So for example coin tossing experiments we figured out that Bernoulli is good right, so die rolls right we will figure out that multinomial is a good distribution what about text, people typically use multinomial distribution so you can think of having a very large dimensional die on one word written on each side of it right so what is the next word you use roll the die that will tell you or the next word to use right so that is that is do not laugh I mean that is the seriously the model that people use for modeling text you know they use multinomial distribution so they assume that each word is a generated independent of the previous word sometimes when okay fine let another yeah.

So that each word is generated independent of the previous word and so you can model that as a multinomial distribution right, so they have actually have different names for it, it is sometimes called the Uni Graham model right also it is called the roughly a bag of words model right where the sequence do not matter and each word is generated in differently so many ways of describing the same idea right but at the end of it is nothing but using a multinomial as comical as it sounded that is what it means.

Having this huge die and rolling it every time I want to add a word to the document okay so that is multinomial right so once you have decided what is the distribution that you think is appropriate for modeling the data then you go and decide on what your prior should be right so sometimes the choice of the distribution for modeling the data is driven by not whether there is a conjugate prior is available for the distribution or not right so maybe there is a different distribution that is perfect for modeling data.

But because there is a very convenient conjugate prior for multinomial right people want to use multi no means so like that there are other instances where even though Gaussian is inappropriate for example people want to model discrete value data right the Gaussian cannot do discrete values right but then coming up with the distributions which have allowed discrete values and have nice conjugate priors is hard right so people just go with Gaussian sometimes you end up operate they use Gaussian.

Because it has a nice conjugate price so θ and y are conjugate priors important, because an easy to do things in iterative fashion because once I run some data through the β Bernoulli pair, okay I am going to end up with a β distribution over the parameters again so if I get more data and I can just happily just go ahead and do it and if I every time I run through it I keep getting a different probability distribution there is no functional form for me to stick these things into and then it becomes very hard for me to do this in any tractable fashion then I cannot come up with some parameter update equations or anything like that so it becomes very hard so, so the conjugacy is very important.

IIT Madras Production

Funded by
Department of Higher Education

Ministry of Human Resource Development
Government of India

www.nptel.ac.in

Copyrights Reserved