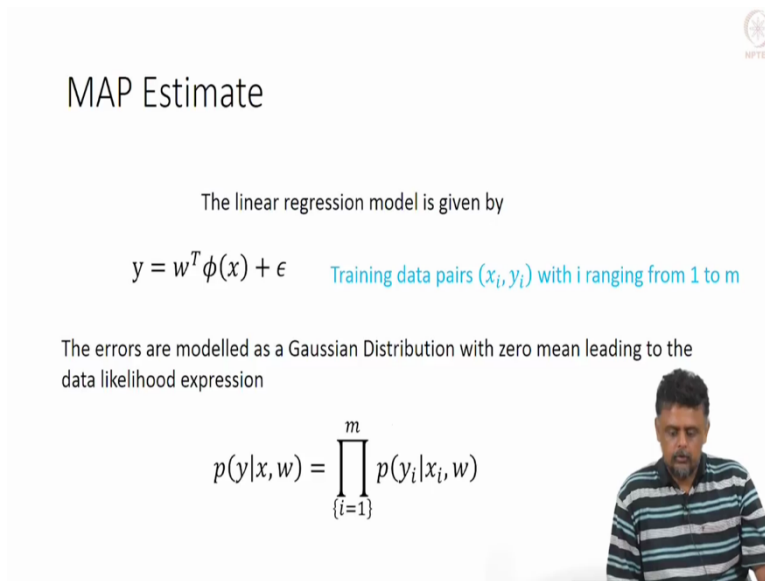


Machine Learning for Engineering and Science Applications
Professor Dr. Ganapathy Krishnamurthy
Department of Engineering Design
Indian Institute of Technology, Madras
MLE, MAP and Bayesian regression

Hello and welcome back, so in this video we will look at map maximum posteriori estimation Bayesian regression. The material is inspired by PRML book by Christopher Bishop and many images are also taken from the book, okay.

(Refer Slide Time: 0:31)



The linear regression model is given by

$$y = w^T \phi(x) + \epsilon$$

Training data pairs (x_i, y_i) with i ranging from 1 to m

The errors are modelled as a Gaussian Distribution with zero mean leading to the data likelihood expression

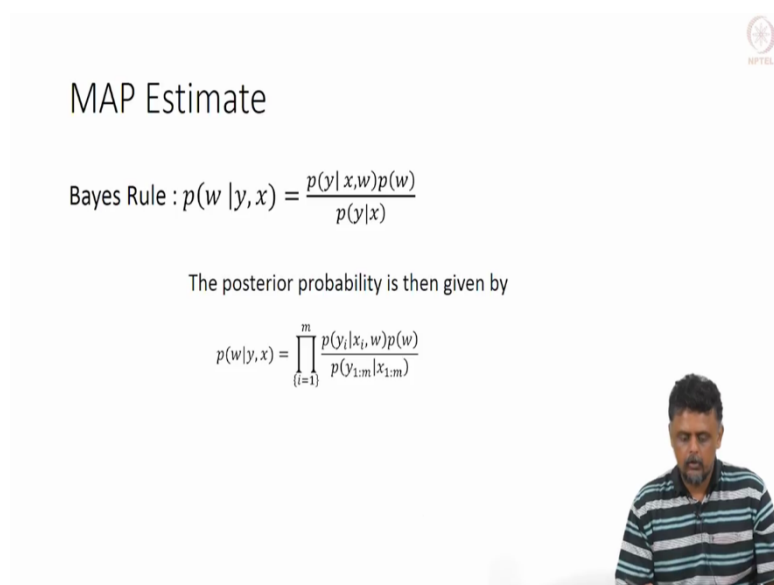
$$p(y|x, w) = \prod_{i=1}^m p(y_i|x_i, w)$$

The slide includes the IIT Madras logo in the top right corner and a small video inset of the professor in the bottom right corner.

So we will just consider it in the context of linear regression, okay. So we have this smaller typically this is how we model in the case of linear regression where phi of x corresponds to some polynomial functions of x and any other functions of x or typically 1x, x square, x cube so on, okay. We are given training data pairs x_i, y_i with i ranging from 1 to m .

We saw earlier if the point x_i, y_i are independent and identically distributed then the probability of observing the dataset is nothing but the probability of the product of the probabilities of the individual data points, okay. So this is the likelihood function, likelihood of data, okay given the model. Okay. So that's how we had earlier modelled it and when we took the log likelihood negative log likelihood. We ended up with the mean square loss function if we model this as if we model each one of these as a Gaussian function, okay.

(Refer Slide Time: 1:43)



MAP Estimate

Bayes Rule : $p(w | y, x) = \frac{p(y|x, w)p(w)}{p(y|x)}$

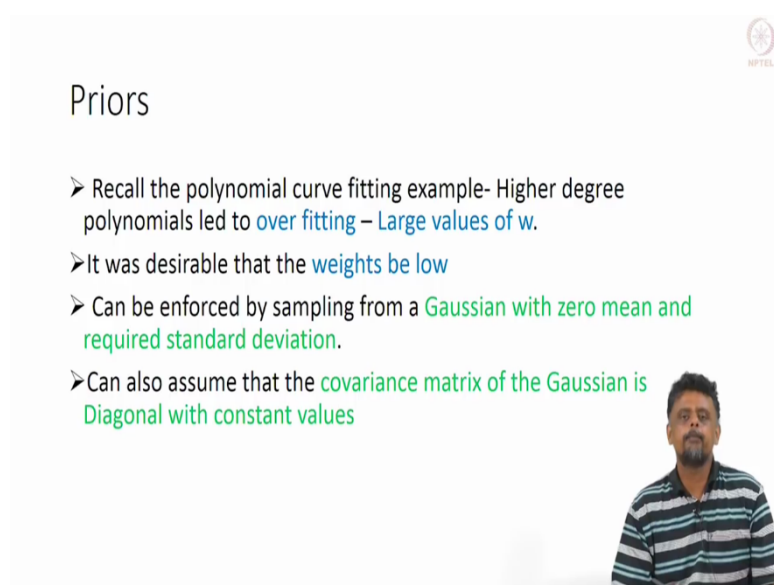
The posterior probability is then given by

$$p(w|y, x) = \prod_{(i=1)}^m \frac{p(y_i|x_i, w)p(w)}{p(y_{1:m}|x_{1:m})}$$

Now if we use Bayes rule, so basically we want to calculate p probability of w given y and x, okay given the data, so y, x training data pair. Then using Bayes rule we can write it in this form, so where this is our likelihood and this is the prior, okay. The denominator we can call it is as observing the probability of data which is a constant, right?

So if we expand out the likelihood comes once again and up with a products of individual probabilities with a normalization factor here which is again evaluates the constant because it depends on the training data set alone and not on w, okay. So it has scaling impact, so we can always observe it as a constant, okay.

(Refer Slide Time: 2:44)



Priors

- Recall the polynomial curve fitting example- Higher degree polynomials led to **over fitting** – Large values of w.
- It was desirable that the **weights be low**
- Can be enforced by sampling from a **Gaussian with zero mean and required standard deviation**.
- Can also assume that the **covariance matrix of the Gaussian is Diagonal with constant values**

So now we take we think about what do we do the prior.

(Refer Slide Time: 2:50)

MAP Estimate

$$\text{Bayes Rule: } p(w | y, x) = \frac{p(y | x, w) p(w)}{p(y | x)}$$

↑ Posterior ↑ Likelihood ↑ Prior
↓ PCF

The posterior probability is then given by

$$p(w | y, x) = \prod_{i=1}^m p(y_i | x_i, w) p(w)$$

$p(y_{1:m} | x_{1:m})$
→

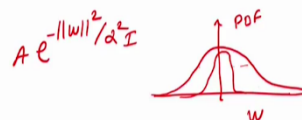
 Evaluates to a constant, depends on the data set and not on the parameters w

So if we go back again, we saw I said this was a prior and exactly mention what it was. It came out of the application to the Bayes rule to the posterior probability. Just to mention again this is the posterior, posterior probability. So the prior same to just come out of the application of Bayes rule. So we have to see what it means, okay.

(Refer Slide Time: 3:12)

Priors

- Recall the polynomial curve fitting example- Higher degree polynomials led to over fitting – Large values of w .
- It was desirable that the weights be low
- Can be enforced by sampling from a Gaussian with zero mean and required standard deviation.
- Can also assume that the covariance matrix of the Gaussian is Diagonal with constant values



So if you recall the polynomial curve fitting example we had as the degree of the polynomial increased there was over fitting which again let or very large values of w , okay. Remember very large values of w . It was desirable at that time that the values below, right? Because w is making very large to compensate for individual data points. So it was over fitting.

So how can we enforce that? We can enforce that by saying that the prior should be a Gaussian. Okay with zero mean and standard deviation which we can rather estimate or we can either set a parameter by cross (3:54), okay. And in this case we can assume that the covariance matrix of the Gaussian is the diagonal. So how does it help avoiding large values of w ?

So if we impose the prior that w is drawn from a Gaussian distribution then we are looking at the functions something like $e^{-\frac{1}{2}w^2}$ to remember zero mean, so some normalizing factor here, okay. And I'm also not writing out let's say some α^2 the variants parameter α^2 I am not writing that out. But you can just look at $e^{-\frac{1}{2}w^2}$.

So if you plot the PDF for w on this axis again w can be positive or negative, we will see that, it can be something like this, right? This is what you have seen. So for very large values of w , so this is the PDF. So for very large values of w the probability of getting that w drops off exponentially.

And we can actually control these width as hyper parameter, so then we can actually make it very sharp also. So that we can only get very small values of w . So by imposing the prior we make sure that there is no over fitting that's an advantage of using a prior.

(Refer Slide Time: 5:15)

MAP Estimate

Bayes Rule : $p(w | y, x) = \frac{p(y|x, w)p(w)}{p(y|x)}$

Handwritten annotations:
 - $p(w | y, x)$ is labeled "posterior"
 - $p(y|x, w)p(w)$ is labeled "Likelihood"
 - $p(w)$ is labeled "prior"
 - $p(y|x)$ is labeled "evidence"

The posterior probability is then given by

$$p(w | y, x) = \prod_{i=1}^m \frac{p(y_i | x_i, w)p(w)}{p(y_{1:m} | x_{1:m})}$$

Handwritten annotations:
 - $p(y_{1:m} | x_{1:m})$ is circled in red
 - An arrow points from the circled term to a box:

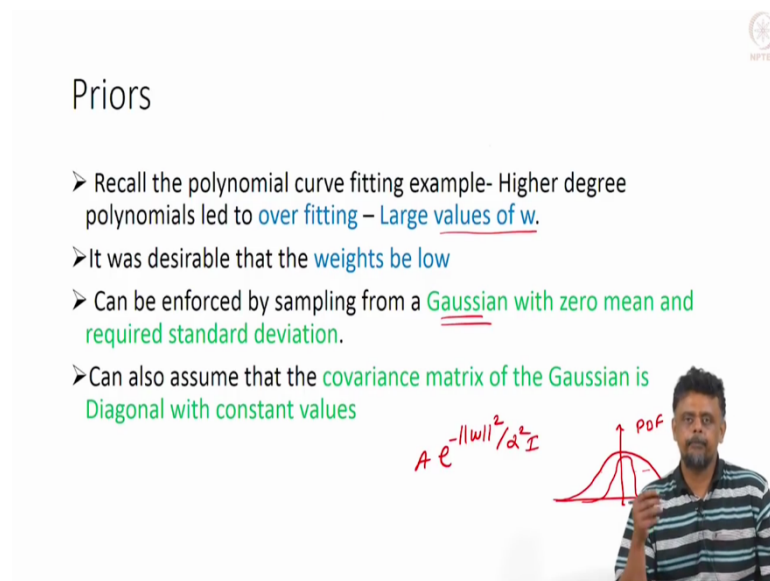
Evaluates to a constant, depends on the data set and not on the parameters w

So to recap we want to estimate the, what we call the posterior distribution like we want to directly estimate w , this is what we do when you do the optimize (5:23), we actually estimate w . So from probabilistic point of view we want to estimate w given the dataset y, x .

Now if you use Bayes rule and rewrite it in this form, so it turns out to be a product of likelihood and prior.

We saw that by taking the log likelihood, we end up if the likelihood function is moral as a Gaussian. The log likelihood leads to the least squares log function. And that what causes over fitting in many of (())(5:48). By looking at the posterior we have this term prior, this is multiplicative.

(Refer Slide Time: 5:58)



Priors

- Recall the polynomial curve fitting example- Higher degree polynomials led to over fitting – Large values of w.
- It was desirable that the weights be low
- Can be enforced by sampling from a Gaussian with zero mean and required standard deviation.
- Can also assume that the covariance matrix of the Gaussian is Diagonal with constant values

$A e^{-\|w\|^2 / \alpha^2 I}$

PDF

The slide features a hand-drawn diagram of a Gaussian distribution curve in red. A man in a striped shirt is pointing at the diagram. The NPTEL logo is visible in the top right corner.

And what we do is you can impose any kind of prior on it, so it has to be a probability distribution. One of the most common used prior is the Gaussian distribution. And if you plot Gaussian distribution you will see that for very large values of w the probability of getting the w kind of goes to 0, okay.

(Refer Slide Time: 6:13)

MAP Estimation

Choose

$$p(w) = N(0 | \alpha^2 I)$$

α^2 is a hyper-parameter that can be learnt or adjusted

$$\rightarrow p(w|y, x) = \prod_{\{i=1\}}^m \frac{p(y_i | x_i, w) p(w)}{p(y_{i:m} | x_{i:m})}$$
$$p(y|x, w) = N(y | w^T \phi(x), \sigma^2 I)$$

Handwritten notes: Red arrows point to α^2 in the first equation, y_i and x_i in the second, and $w^T \phi(x)$ in the third. A red underline is under the second equation.

So now we will just explicitly model this p of w is N is the normal or Gaussian distribution with some Alpha square, a diagonal matrix is the covariance matrix. Again it can be learnt parameter or a hyper parameter then if we write out the posterior distribution of w given the training data, so it can be written in this form once again I pointed out that the denominator is again a constant.

And again recall that you can also model the likelihood here as a, again it is another Gaussian estimate of the mean is given by the model w transports to x , here I should to be more general I should write w transports ϕ of x , okay. You can put that in there. Again we assume a, in this case 2 some Sigma squared I parameter, okay. Some Sigma squared where all the variants are the same, okay.

(Refer Slide Time: 7:06)

MAP Estimation

- Plugging in the distributions for the likelihood and the prior and taking negative log

$$-\ln p(w|x, y) = -\sum_{i=1}^m \ln p(y_i|w, x_i) - \ln p(w) + \text{constants}$$

$$\rightarrow = \sum_{i=1}^m \frac{(y - w^T x)^2}{\sigma^2} + \frac{\|w\|^2}{\alpha^2} + \text{constants}$$

w_{MAP}

Leads to L2 regularization



So again if we take the negative log of a posterior probability then it, you see that it easily simplifies to this because it is a log of 2 exponents, so then the exponent comes down then you see that it comes to this log function which is your least squared log function plus L2 regularization term. So recall that when we did the polynomial curve fitting by adding L2 regularizer we are able to prevent the weight from growing up and also to prevent over fitting, okay.

So which is the same as evaluating the posterior probability or the maximum or the map, okay. So you can actually solve this problem of using (7:46) and you get a value of w which we call the map estimate, okay. That's one way of solving this (7:52).

(Refer Slide Time: 7:54)

MAP Estimation

- The log of the posterior distribution

$$= \sum_{i=1}^m \frac{(y - w^T x)^2}{\sigma^2} + \frac{\|w\|^2}{\alpha^2} + \text{constants}$$

Can be written in the form $(w - \tilde{w})^T K^{-1} (w - \tilde{w})$

$$K = \frac{\phi^T \phi}{\sigma^2} + \frac{I}{\alpha^2}$$

$$\tilde{w} = \frac{K \phi^T y}{\sigma^2}$$

$\hat{y} = W^T \phi(x_{test})$
↳ predictive distribution

$(x - \mu)^T \Sigma^{-1} (x - \mu)$
↳ exp. in the Gaussian

w_{MAP}



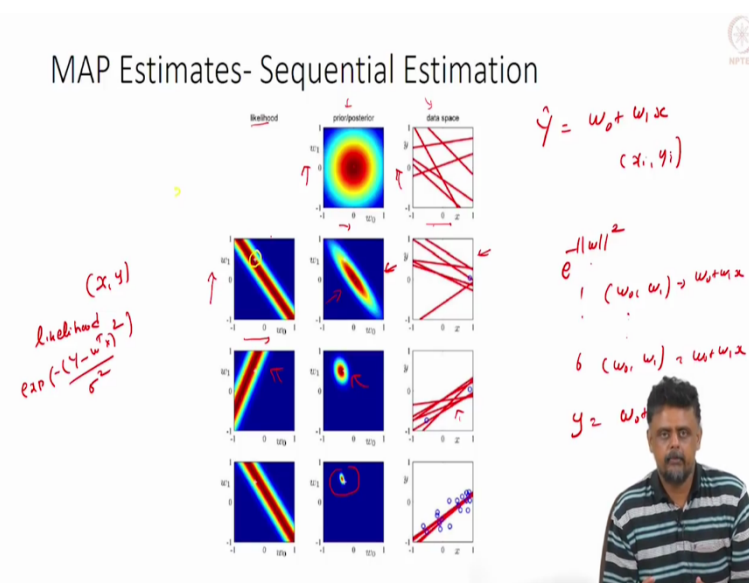
However actually if you look at this log probability. The log probability can be rewritten in this form. So what is this? If you recall we have seen expression of that type x minus μ transports Σ inverse x minus μ , you recall? This is the exponent in a Gaussian. So it turns out that if we use the Gaussian model for the likelihood and a Gaussian prior it turns out at the posterior also is a Gaussian.

And we can actually describe it as a trick called completing the squares then you can get into this form. From which you can directly estimate w tilde this is again the same as w map and the corresponding covariance matrix, okay. Okay, so if you see that w tilde depends on the covariance estimate which contains contribution from both the data and also our constraint that we impose in the form of a prior the α squared and it goes into determining the w tilde or the w map estimate.

So this is the same as w tilde. w tilde is the same as the w map estimate. It is the mean estimate that we have, okay. But once again recall, see that it is still a point estimate of w , so you will get one value of w and you use that w to actually calculate for every new test data you will do ϕ x test data to get a point estimate of your output y , y tilde that is typically what you get, right? That's what you do all the time.

Now that is nice but then we will go one step further and do what is called Bayes regression where we do a predictive distribution on y tilde. So what we want is an error bar on your estimate \hat{y} . \hat{y} is what you predict but it is just one number that is thrown out even if you use map or whether you use MLE just one number that comes out. What would be interesting in many cases, useful in many cases to have an error bar on your estimate and that's where Bayes integration fully Bayes integration comes in.

(Refer Slide Time: 10:14)



But we will look at this map estimate with a slide example, okay. So let's consider this problem where your model is w not plus w_1x , okay. So you are given training data point's x_i , y_i , okay. So in this toy example w not and w_1 were fixed for various values of x you get generated y_i , x_i s would generate y_i s and then some noise was added to the y_i s, okay. So that is a data point.

So then, okay we start off with our prior, right? So the every row the leftmost column is a likelihood, the middle column is the prior which also doubles up, you know it doubles up the posterior also and the right one is your predictions if you can think of it that way, that's the data space. Okay, so first where the prior, as we saw we have a Gaussian prior, right?

We have a Gaussian prior, that's what if you plot this out, this is what this will look like. Where w equal to 0 that's where it is, we will correspond to the maximum value and it will drop-off exponentially as you go away from zero, okay on what the axis. So the axis, this axis w not that axis is w_1 . So we are plotting this e exponential distribution as a function of w not and w_1 , so that's our prior distribution.

From that prior you draw 26 samples and you see the red lines here. The red lines here corresponds to the 6 samples of w not, so you will draw 6 pairs of w not w_1 , okay 1 to 6. And for each one of them you can calculate w not plus w_1x . For your training data, you have training data X , you can do that or for many or the arbitrarily values of x you can calculate.

So if you do that then for every x there is a y , right? x , y so using that you can actually plot line, right? So this is the equation of the line, so y is w not plus w_1x . So if you draw 6 pairs of w not w_1 from here you can get 6 lines in the xy , so this is x and this is y , okay. That's the data space, right? So we actually have the training data, right? Let's say we observe one data at a time.

So in this case we have let's try to highlight it in something that is visible. So if you look this is the data, this is a white cross I don't think it is very clear but that's an observed Datapoint, okay. So you have one data point, okay that you got and if you have the Datapoint then let's say you have one x , right? One xy , you have one xy , there has been observed, okay.

So then what can you do? You can calculate the likelihood, how do you calculate likelihood? You remember it is some exponent, I am remembering all the constant y minus w transports x squared by some Σ squared, okay something like this. So you can actually calculate the given x and y for various values of w not and w_1 you can calculate this likelihood.

And the dark red bands corresponds to regions of maximum likelihood, for this one Data point that we have observed. So now you take this likelihood and multiply it with the prior, okay. So one likelihood (\cdot) (13:49) times prior gives you the posterior distribution which is this, okay. So now you can draw 6 more points from this posterior distribution corresponding to w not and w_1 .

6 pairs and you plot 6 different lines here in the xy plane. So then you go back here you observe one more Data point, you can recalculate, again you can calculate the likelihood for that Data point again it gives you slightly different likelihood distribution. So now this is the new prior. This prior times your likelihood will give you your new posterior distribution from which once again you can draw a 6 pairs of w not and w_1 and plot this line right here, okay.

So you can keep doing that after the 20th Datapoint you will see that the posterior distribution has become very sharp, sharply peaked, right? There is one, everywhere it is zero where it is very maximum value at some point and if you sample once again 6 data points from this distribution you will get these lines here, these red lines here which are actually fitting all these data points if you think about it, okay.

So by sequentially considering data points x,y and each time evaluating your posterior distribution but then using data set prior for the next time that you observe a new Data point, you can actually come to the you know converts to the appropriate posterior distribution.

Now this is how you do map estimate, if you think about it probabilistically. Another way of looking at it if you think about it, you see that by sampling from w , right?

Now that you have this distribution you can keep sampling from w and you can give estimating for every x, y , for every x because estimate y , for every x you can sample from w many many times and you can estimate a y , okay. So that kind of gives you an error bar on your estimate, okay. But those error bars will tend to be very similar across-the-board because w is completely determined by your training data.

So the new x for which you are breaking y there is no impact on the error bar, okay. So you will get a similar error bar for pretty much depending on the training data set that will determine your error bar (16:09), okay. So if you think about it is just a point estimate. How do you get a point estimate? Now that you have very sharply picked distribution for w not and w_1 , you can look at Arg max . So what is the w not and w_1 which corresponds to the maximum probability, okay. You start that space and you will be able to determine the pair w not and w_1 that is still a point estimate, okay.

(Refer Slide Time: 16:35)

Bayesian Prediction

- The MAP estimate is still a point estimate
- It provides a posterior distribution for error in estimate of w
- Prediction is for new values of input data, i.e. y_{test} given x_{test}
- Bayesian regression is all about determining the predictive distribution.


$$P(y_{\text{test}} | x_{\text{test}}, D_{\text{train}})$$



What we want to do next is, since we still only have point estimates. What we want like we said is, we want the predicted distribution, okay. predicted distribution means we want a distribution on our prediction. So that is probability of Y test given x test, right? Crudely speaking and your training data I'm going to call this D and the parameters w , okay that's what we want.

We want an error bar on our prediction, if you think of it in a very simple sense, again we want an error bar on our prediction directly, okay. What is the error bar on the prediction? So the idea in Bayesian regression is to determine the prediction distribution itself, okay. So how do we go about doing that? Okay.

(Refer Slide Time: 17:23)

The Rules of Probability 

Predictive Distribution

$$p(y_{test}|x_{test}, x, y) = \int p(y_{test}, w|x_{test}, x, y) dw$$


$$p(y_{test}|x_{test}, x, y) = \int p(y_{test}|x_{test}, x, y, w) p(w|x, y) dw$$

$$p(y_{test}|x_{test}, x, y) = \int p(y_{test}|x_{test}, w) p(w|x, y) dw$$

$\int f(x) p(x) dx$ $\arg \max$ Variance Gaussian $\int \frac{1}{\sigma^2} p(x) dx$ Posterior

sum rule $p(X) = \sum_Y p(X, Y)$

product rule $p(X, Y) = p(Y|X)p(X)$



So if you pay attention to this, what we actually want? We want this distribution. So we want for a new x test what would be the probability of y test? The output for new Data point given the new Data point and the training data, that's what we want, okay. So that we can write using product rule of probabilities

So this is some rules, so p of x is summation over y, p of x, y. So we introduce this random variable w which is the estimate based on our model and then we marginalized over it. So the summation is just replaced by the integral, okay. The 2nd rule is a product rule where p of X,Y is P of Y given X times P of X. So in this case p of y test, w, we use this expression to decompose into 2 products.

P of y test, w is p of y test given we retain this conditional aspect times p of w, okay given x,y, okay. So this is the short variation for how do get to the predicted distribution. Now why do you want it in this form? So we will go one step further, we have written p of y test give x test xy Yes as p of y test given x test and w, so we have left of this training data.

The reason is because it doesn't really matter because catches all the information in the training data, okay w catches all the data. So we can still estimate w from the training data

might is not a problem. So we can just leave out the dependency on x, y because it is already being taken care of, so we have this, okay. And this we know is the posterior distribution.

This we know, we have modelled is. So can model this as a Gaussian, this also we can model as a Gaussian based on your, this is kind of your likelihood function, okay. This is another Gaussian, okay. So once we have this estimate, so what does this do? What does this $p(y|x)$ test, what is happening here? So we can think of it as, let's say we have a quantity here $f(x)$ and we want to estimate the mean of that quantity, this is what we would do, right? With respect to the probability distribution.

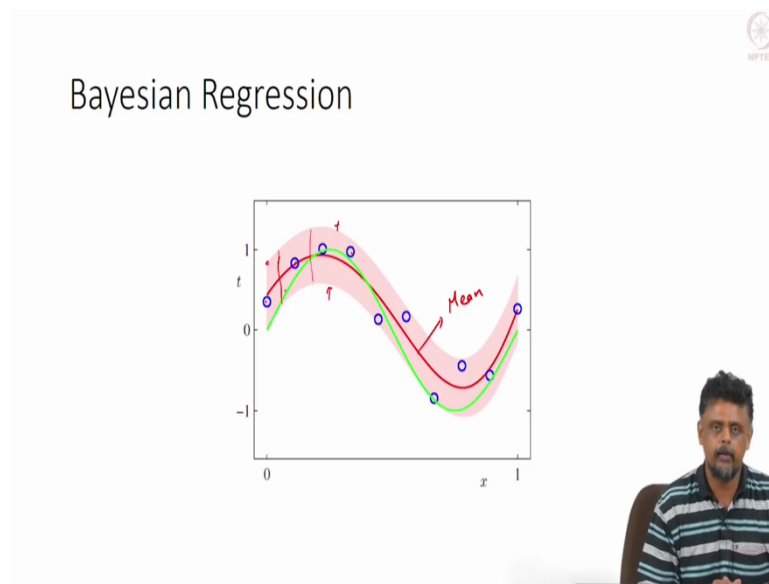
So what we are trying to, we are estimating a mean of this quantity. What is a probability of y test? Given this input x test and a model w , okay. That we are calculating in average based on the probability distribution of w condition on your training data. So that's exactly what we are looking for, okay. So your $f(x)$ here is the $p(y|x)$ test given in x test and w and we are $p(x)$ is nothing but your posterior distribution.

So you can think of it as a way of calculating an average over all possible realizations of y given your training data, okay. So what can we do with it? So we can actually do Arg max_y of this and figure out the y test for which probabilities maximum that is estimate of y . We can also calculate variants, right? Because remember we can calculate $(\int f(x)^2 p(x) dx)$, right?

We can calculate mean then we can calculate $f(x)^2$, remember. $\int f(x)^2 p(x) dx$, right? And if you're zero centered this is a variant, so for every new test data point we cannot only calculate by using Arg Max the mean, we can also calculate the variant, this is possible of course assuming that both of these are Gaussian. It turns out that this integral that you see here generally intractable because you cannot solve it for everything.

But if you assume that these are Gaussian then this can be actually been done and it turns out to be another Gaussian, okay. We want to look at exactly what the forms are because it's little bit confusing but the general idea, the idea is if you estimate this distribution using Arg max we can get the mean assuming that it is Gaussian and you can also calculate the variant. So for every new test data point, okay.

(Refer Slide Time: 21:36)



So then if we look at the output of Bayesian regression, so the Green is a true curve which we have noise added and we have sampled this blue circles we have seen this before. The red is the mean estimate, right? That is the mean estimate for y and this spread for every Data point is available. So variants of the estimate of the mean for every one of your test data points.

So that's a prediction as this error is band which is given here that's an error at every point on your prediction. You see that the band neatly encapsulates pretty much the Green car. So that's the advantage behind doing Bayesian regression.

(Refer Slide Time: 22:21)

The Rules of Probability

sum rule $p(X) = \sum_Y p(X, Y)$

product rule $p(X, Y) = p(Y|X)p(X)$

Predictive Distribution

$$p(y_{test}|x_{test}, x, y) = \int p(y_{test}, w|x_{test}, x, y) dw$$

$$p(y_{test}|x_{test}, x, y) = \int p(y_{test}|x_{test}, x, y, w) p(w|x, y) dw$$

$$p(y_{test}|x_{test}, x, y) = \int p(y_{test}|x_{test}, w) p(w|x, y) dw$$

- Mark Carlo

$\int f(x) p(x) dx$ arg max Variance Gaussian Posterior $\int_{\mathbb{R}^2} f(x) p(x) dx$

So what is the biggest problem in doing Bayesian regression is, except for some forms of p of y given x test w this form is very similar to the likelihood function we can model this as a Gaussian except for this forms wherein these are Gaussian are some very specific distribution this integral is very hard to do, okay. So then is are done with what are called Monte Carlo techniques.

So numerically intensive but the advantage is, once you estimate this distribution you can do Arg max you can do mean of the estimate, right? You can do mean of your estimate, you can also calculate the variants of your estimate, also note that, you see that unlike in the map or MLE estimate your output. The probability of y test depends on your current data point also, right? So that's the interesting thing, right?

So it takes in that it is influenced by the data point. So if you actually write this out for a Gaussian models both the posterior as well as this distribution being Gaussian you can actually see that influence directly, it's in the textbooks I will give you the reference that in the terms, okay you can look that up.

So this and a brief look at you know Bayesian regression. As far as Bayesian regression is concerned the idea is to get a predictive distribution for your output, so you are you are doing \hat{y} that we want an error bar on that \hat{y} that's what we want. The map estimate is basically trying to maximize the posterior distribution, okay. and the advantage of doing that is that you can accomplish regularization okay.

So by imposing a Gaussian prior you got L2 regularization, you can post some for instance a exponential prior you can get L1 regularization wherein most of the coefficients will go to 0 but the probability of getting coefficient is very small. So that way that's possible, okay. So these techniques are applied widely in different scenarios, this is just in the context of linear regression you have seen it.

The treatment in bishop is rather dense but it is actually one of the best treatments available, so I urge you to go through it I will also provide you with the references for the textbook. Bishops textbook by the way is available online for free, so you can look through that, okay thank you.