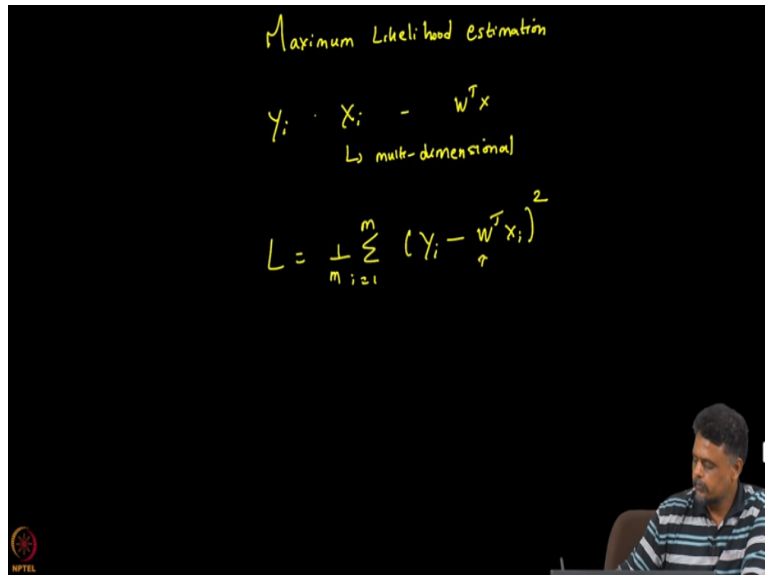


Machine Learning for Engineering and Science Applications
Professor Dr. Ganapathy Krishnamurthi
Department of Engineering Design
Indian Institute of Technology, Madras
Maximum likelihood Estimation Intro

(Refer Slide Time: 00:14)



Maximum Likelihood Estimation

$$y_i = X_i \cdot W^T$$

↳ multi-dimensional

$$L = \frac{1}{2} \sum_{i=1}^m (y_i - W^T X_i)^2$$

Hello and welcome back so in this video we will look at what we mean by maximum likelihood estimation, so you are all familiar with the linear regression model, so we have bunch of data points Y_i and we have the corresponding features X_i , so typically we formulate this model W transpose X , again X_i can be multi-dimensional or we just be single variable but that does not matter in this case, so our model that what are try to fit typically is when you so look at linear regression we looked at the least squares loss function the loss function is nothing but you have summation over M data points square then we have D .

So that was over model and we took the derivative of this model we use gradient descent to estimate W , so where do get this wise in that we use least squares, so why is it power to we have looked at $L1$ and $L2$ norms and all that but still why is this the best way to do it there are many ways of approaching this one way of doing figuring out this least squares loss function is by looking at the probabilistic our perspective.

(Refer Slide Time: 01:53)

$$L = \frac{1}{2} \sum_{i=1}^n (y_i - w^T x_i)^2$$
$$\epsilon_i^2 = (y_i - w^T x_i)^2 \begin{array}{l} \sim \text{error} \\ \text{Error} \\ \text{Noise} \\ \text{Missing data} \end{array}$$
$$p(\epsilon_i^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{\epsilon_i^2}{2\sigma^2}} \rightarrow \text{Zero mean}$$

So one if we will just reduce we will consider individual errors, so for instance we will define this a variable epsilon I as I square the square S that is the error, let me just write, so this error might be due to noise in our measurement and missing data may some features are missing this could be the error due to those concept, so because some X i might be missing for a particular data point and maybe there is an error in measuring X i as well as measuring the Y i, so typically one assumption people about that this is that these are Gaussian distributed, so what does that means, means that the probability of observing a particular epsilon square, we assume that it is given by a Gaussian distributions with zero mean, so once you make this assumption then we can rewrite our problem.

(Refer Slide Time: 03:49)

The image shows a blackboard with handwritten mathematical derivations. At the top, it defines the probability density function of a Gaussian distribution: $P(\epsilon_i^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\epsilon_i^2/2\sigma^2}$, with a note "Zero mean" and "Missing data". Below this, it introduces the likelihood function for independent and identically distributed (iid) data: $P(y_i | x_i, w) = \left\{ \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_i - w^T x_i)^2}{\sigma^2}\right) \right\}^m$, where m is the number of terms. The negative log-likelihood is then derived as $-\log P(y_i | x_i, w) = \sum_{i=1}^m (y_i - w^T x_i)^2$. This is equated to the sum of squared residuals, $\sum_{i=1}^m (x_i - \mu)^2$, which is identified as the variance. The NPTEL logo is visible in the bottom left corner of the blackboard image.

So because we know what sum square is, so we can write this as P of, so we have this probability of P, so we can rewrite this as probability of a P of Y i given X i and W, so then we can just say, so will that side exponent because then it is sassier to minus I might have missed out a negative sign here, so this is our model so the just spray we are just plugged in out model here and then we will just read to pretend this probability S, Y i given X i W so the idea behind likelihood is to maximize this likelihood.

So which means that if you maximize that good which means maximize this expression on the right with respect to X i and W another way of looking at it is we can also maximize any other function of P in this case if you take the negative log of P then I am going to ignore some of the additive factor here you will get if you take the log of the exponent and then the negative sign you will get Y i minus quite again I have just written it for one data point, so that suppose your training data consist of capital or small m data point which we saw then it will be just the summation you know why that happens is because if you want to maximize the probability of observing this data set.

Assuming that they are in divided IID then what we get is the probability of the data set is the product of the probabilities of the individual data, so this will be product of as M such terms M terms so for each term this and why I times Y i minus W transpose X i, so but if you, so if you do actually calculate the probability of the entire data set is a product over this each of the data

points, so and we take a log of this should get a summation you said this is our least squares cost function.

So I have just not done the step where I do the product but that should something you should be able to do in other way of looking at it this you if you assume that our data is Gaussian distributed what we are modeling here is the mean if you remember the form of the Gaussian distribution then the Gaussian distribution form as exponential I am just going to use different variable but that should not throw you off X minus μ by σ squared, so this is the mean and this is variance, so this is what we were what we refer to as the maximum likelihood estimate.

So when you do the least square cost function we are assuming that the errors are Gaussian distributer or basically we are trying to model the mean using this W transpose X by mean in the sense of for every measurement on an average, so you can think of it that way and we are trying to estimate this parameter assuming the Gaussian distribution and when we try to increase the likelihood of observing the data given the parameters, which way what we are trying to estimate then we end up with the least squares loss function of course if you we can also show that for classification problems at least for the two class classification problem if we start off with the Bernoulli distribution.

We can end up with the log loss or the binary cross entropy cost function which is pretty much same way, we can do that are pretty much the same way, so that see an introductory look at the maximum likelihood estimate if time permits either this week or in the subsequent weeks couple of weeks left we will look at some the maximum a posteriori estimate where basically we will be using the Bayes rule and calculating a prior, so we use base we incorporate a prior, so what we calculate here so this we saw that is known as the likelihood of the data

So we incorporate a prior and prior times likely to give you the posterior probability that is what we usually, so if we take that one step further do a full Bayesian analysis it is called base integration time permitting we will address these two topics in the next couple of weeks which basically 11 and 12 weeks will be able to address these topics thank you.