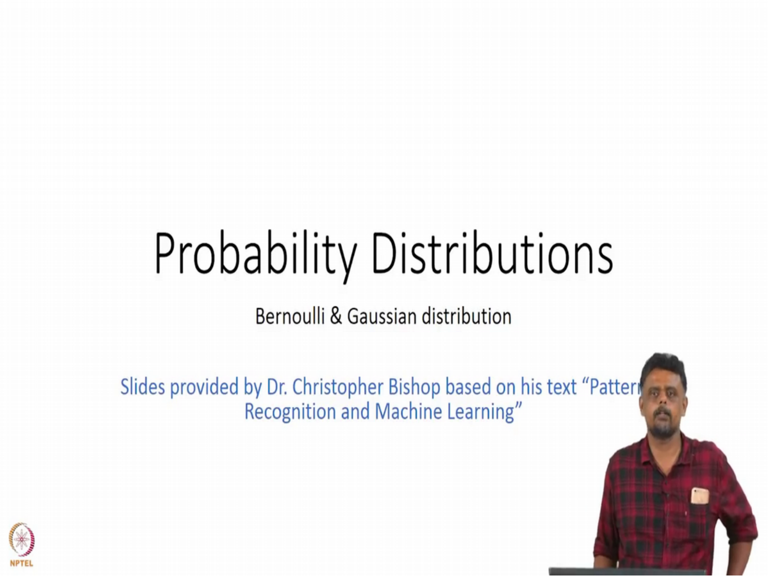
Machine Learning for Engineering and Science Applications
Professor Dr. Ganapathy Krishnamurthy
Department of Engineering Design
Indian Institute of Technology, Madras
Probability Distribution Gaussian, Bernoulli

(Refer Slide Time: 0:15)



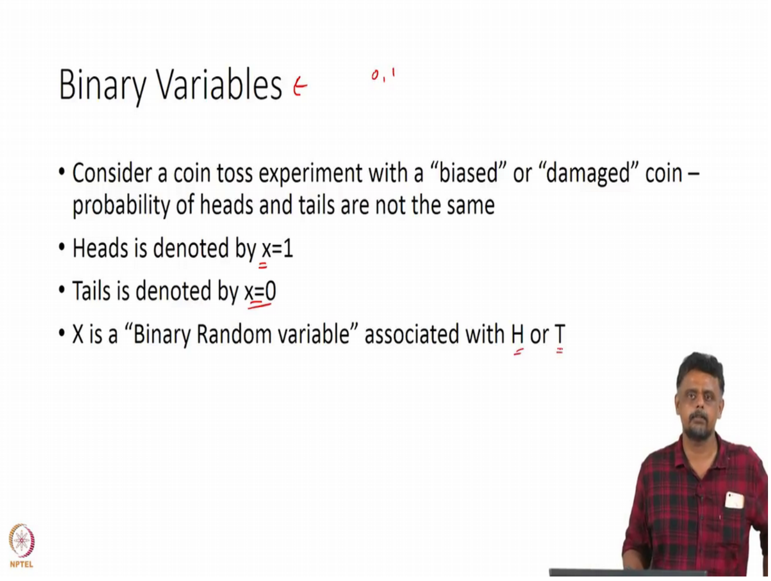
Probability Distributions

Bernoulli & Gaussian distribution

Slides provided by Dr. Christopher Bishop based on his text "Pattern Recognition and Machine Learning"

Hello and welcome back in this video we will look at probability distributions notably Bernoulli and Gaussian distribution. All the slides were provided by Dr Christopher Bishop based on his text book PRMLD 'Pattern Recognition and Machine Learning' textbook.

(Refer Slide Time: 0:29)



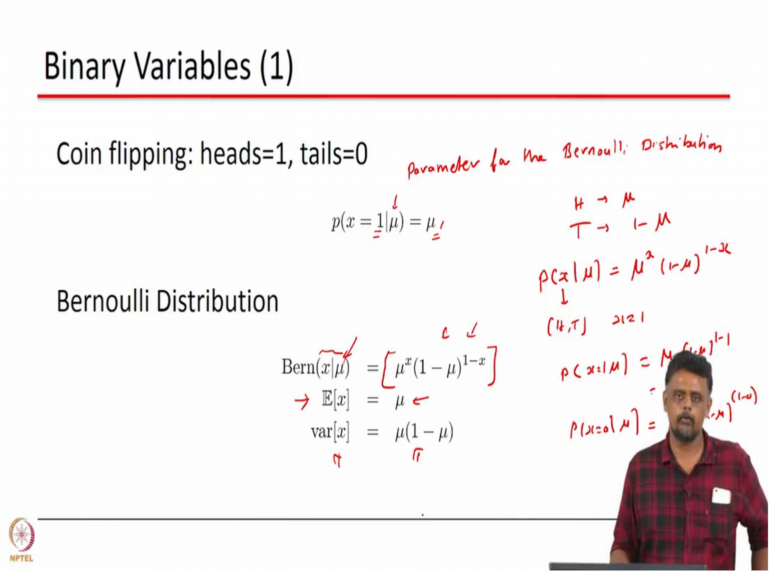
Binary Variables 0, 1

- Consider a coin toss experiment with a "biased" or "damaged" coin – probability of heads and tails are not the same
- Heads is denoted by x=1
- Tails is denoted by x=0
- X is a "Binary Random variable" associated with H or T

So we will first consider the Bernoulli distribution which deals mostly with binary variables, so basically these are 2 states either 0 or 1 okay, so if you have a feature which can take either of 2 values that would be a good example of binary random variable okay, so just have an example we consider coin toss experiment with a damage or biased coin which means that the probability of either getting heads or tails is not the same, so maybe there is higher probability of getting a head than getting a tale when you toss it away okay.

So we will consider the random variable x which is linked to the event heads or tails okay, so depending upon the event x will take a certain value, so if you toss the coin and you get heads x gets the value 1 and if you toss the coin and you get tails x gets a value 0 okay, so that is you variable x is referred to as a binary random variable and it is either associated with you getting either heads or tails based on coin toss.

(Refer Slide Time: 1:40)



Binary Variables (1)

Coin flipping: heads=1, tails=0

Parameter for the Bernoulli Distribution

$p(x = 1|\mu) = \mu$

$H \rightarrow \mu$

$T \rightarrow 1-\mu$

$p(x|\mu) = \mu^{x}(1-\mu)^{1-x}$

(H, T) $x=1$

$p(x=1|\mu) = \mu (1-\mu)^{1-1}$

$p(x=0|\mu) =$ $(1-\mu)^{(1-0)}$

Bernoulli Distribution

$$\begin{aligned}\mathrm{Bern}(x|\mu) &= \left[\mu^{x}(1-\mu)^{1-x}\right] \\ \mathbb{E}[x] &= \mu \\ \mathrm{var}[x] &= \mu(1-\mu)\end{aligned}$$

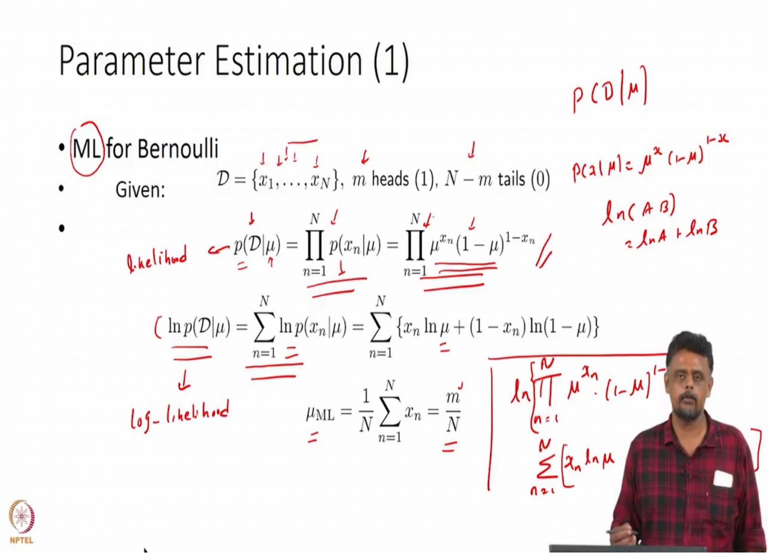
Now what we do is we will assign a probability with that event, so since this is not an unbiased coin this has a bias, we will say that the probability of getting a head given 1 coin toss is given by the value mu okay, so mu (())(1:55) probability that if you toss the coin once the probability of getting a heads is mu and this is a notation, so this mu is the parameter for the Bernoulli distribution right. So if the probability of getting a head is mu okay, what would be the probability of getting a tale should be since heads or tails are mutually exclusive it will be 1 minus mu, right so given a coin toss event if you want to predict the probability of getting either a heads or a tails then this can be given by this cumulative expression.

So which is basically probability of x given mu where x is the event heads or tails is given by this expression mu raise to x 1 minus mu raise to 1 minus x, right so how does this works? Because if x is heads (())(3:11) then x equal to 1, so the probability of x equal to 1 given mu would be just be mu times 1 minus mu time 1 minus 1 it will be just mu it is correct right and the probability of x equal to 0 given mu is mu raise to (())(3:32) time 1 minus 2 mu raise to 1 minus 0 which is 1 minus mu okay so that is correct, so the probability of getting a heads or tails is what this…

If you want to link it to some experiment in that way this is what it denotes, the mu is the parameter for the Bernoulli distribution, so in problems where we define…where we are dealing with this probability distribution most of the time the problem will be about estimating mu that is what we will be doing most of the time. So given mu we know we can now formulate the expression for getting a heads or a tails, so probability of x equal to 1 probability of x equal to 0 given mu.

The average value of mu of x is mu okay we will see that with a small example some time later and the variants of x is mu times 1 minus mu okay so these 2 you can actually calculate based on your formula for expectation as well as the variants right, we will look at more concrete examples later on but this is just an introduction to the probability mass function for the Bernoulli distribution right, so this is given by mu raise to x times 1 minus mu raise to 1 minus x where x is the event x can be either 1 or 0 depending on whether it is heads or tails okay so this is just not for coin flip whenever you have any variable that you can associate with like 1…either or choices like there are only 2 choices for that variable can take them in associate Bernoulli random variable with that quite easy okay.

(Refer Slide Time: 5:08)



Parameter Estimation (1)

- ML for Bernoulli
- Given: $\mathcal{D} = \{x_1, \ldots, x_N\}$, m heads (1), $N - m$ tails (0)
- $p(\mathcal{D}|\mu) = \prod_{n=1}^{N} p(x_n|\mu) = \prod_{n=1}^{N} \mu^{x_n}(1-\mu)^{1-x_n}$

$$\ln p(\mathcal{D}|\mu) = \sum_{n=1}^{N} \ln p(x_n|\mu) = \sum_{n=1}^{N} \{x_n \ln \mu + (1 - x_n)\ln(1-\mu)\}$$

$$\mu_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N} x_n = \frac{m}{N}$$

likelihood

log-likelihood

$p(D|\mu)$

$p(x|\mu) = \mu^{x}(1-\mu)^{1-x}$

$\ln(AB) = \ln A + \ln B$

$\ln \prod_{n=1}^{N} \mu^{x_n} \cdot (1-\mu)^{1-}$

$\sum_{n=1}^{N} [x_n \ln \mu$

So the parameter estimation for the Bernoulli distribution, so it is called the maximum (()) (5:10) technique we will see what is that again later on but just from a common sense point of view we should be able to follow this argument, so we have a set of these events, so let us say tail N coin tosses okay, so out of which we get m heads and N minus m tails okay this is (()) (5:29) I think okay.

So what is the probability of observing this sequence of m heads and N minus m tails okay this is what is note so probability of observing the data set given that probability of heads, this is the condition probability is basically since each of these events is independent we just to do the product of the (())(5:56) each of this probability so that is what this expression does, so since you are doing N coin tosses then for each coin toss what is the probability of observing the head that is probability of x given mu and so since there are capital N coin tosses you just multiply all of them that is what gives us product and instead of probability of x (())(6:15) substitute the expression here if you saw for the probability mass action for Bernoulli distribution.

So recall that probability of x even mu is mu raise to x times 1 minus mu raise to 1 minus x. Here we have sequence of N coin tosses, capital N coin tosses, so for each coin toss we can write this as a probability, since there are capital N coin tosses each of them being independent we just take a product, so this symbol here represents the product okay of m capital N terms okay.

From a computational point of view since multiplication can lead to some of these numerical errors if you take a log of this expression since you take the log of products then it decomposes into a sum of the logs okay because you can take the logarithm, so we will say log of let us say 2 a times b is log A plus log B you can do that right so that is possible, so that is what we have done for each of those terms, so that is the expression rightly here. Taking the log of this product decomposes it into this sum right. Then if we substitute the value in here we can expand it in this fashion, so if you to write it down in a concrete way.

So log of the products N equal m capital N terms mu raise to x. 1 minus mu raise to 1 minus xn, so the log of this entire expression, so this a product of capital N terms, so we can write this as a sum of capital N terms by taking the logarithm inside, so that will get cap summation over n equal to 1 to n, so the log of mu raise to xn is we will get xn log mu right plus…these are products you can this become a summation here plus 1 minus xn log 1 minus mu, so this summation is for this entire term okay.

So simple thing of writing the logarithm of products as sum of logs okay that is what I have done, so then if you want to estimate mu you can set the derivative of this log pDu to 0 with respect to mu and then you can calculate mu ML as total number of heads divided by the total number of tosses, so m is the total number of heads okay, so you might wonder why would we consider this expression and why do you want to take the derivative of this, so you can consider this log pDu like a loss function okay so how is that loss function?

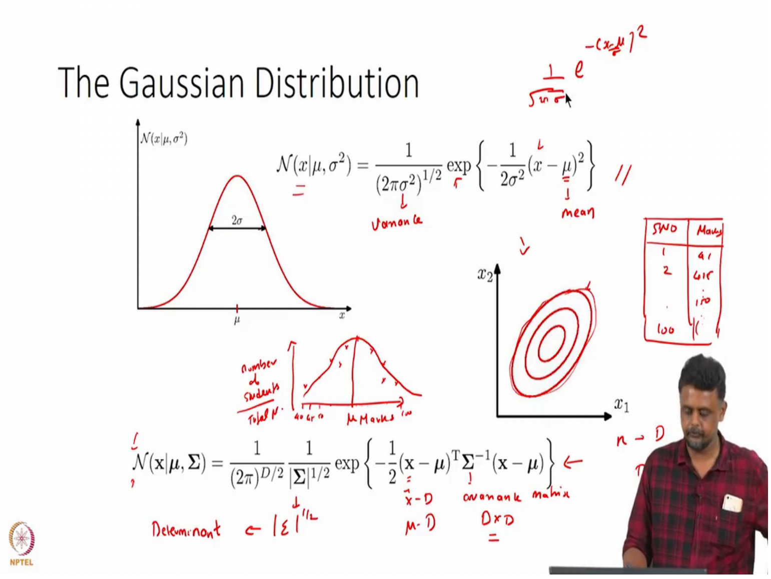
So we formulated this probability of capital D giving mu but probability of capital D is here data, data is basically the sequence of coin tosses, what is the probability of observing this coin tosses given this Bernoulli parameter that is what this expression evaluates to right because each coin tosses independent of each other, we write the probability of observing the data as the probability of observing x1 times the probability of observing x2 times he probability of observing x3 so on and so forth up to the probability of observing x1 so the product all these terms each of the product but then we know the expression for the probability it is Bernoulli random variable x.

So we can plug that expression in there and the rest is just algebra, so now we have got the probability of observing this data set that is this sequence of coin tosses given this mu. Now what we want to do is we want to maximise the probability of observing this data set right so that is the idea (())(10:19) construct in this probability distribution. So if you want to maximise the profitability of observing the data set, the same as maximising the log of the probability of observing this data set the same as maximising the log of the probability of observing these datasets with respect to mu the parameter that you want to estimate.

So by taking the derivative with respect to mu of this function, this log pDu then you can calculate mu ml so this is referred to as the, this log of p this terms is always referred to as the log likelihood and this p of d given mu is referred to as the likelihood. What is the likelihood of observing this data set given this parameter mu, so this is Bernoulli trial, sequence of Bernoulli trials and it is characterised we assumed that it is characterised by this mu which is the probability of observing x equal to 1 okay so we then consider construct the probability of observing the entire data set which is x1 to x1 through xN we assume that each trial is independent of each other.

So it is just the product of the probabilities of observing each individual trials that is what this (())(11:28) in this expression and you just to plug-in the formula for P of x given mu for Bernoulli distribution and then take the log because then it is easier to process that way and then differentiate this law with respect to mu and set it to 0 and that will give you…and from there you can derive mu ML, I have not gone through the algebra but it is not too hard to do okay so this expression this probability of observing a dataset given the parameters of your distribution is referred to as the likelihood and taking the logarithm of that usually referred to as the log likelihood okay.

(Refer Slide Time: 12:05)



The Gaussian Distribution

$$\mathcal{N}(x|\mu,\sigma^2) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}$$

$$\mathcal{N}(\mathbf{x}|\mu,\Sigma) = \frac{1}{(2\pi)^{D/2}}\frac{1}{|\Sigma|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\mu)^{\mathrm{T}}\Sigma^{-1}(\mathbf{x}-\mu)\right\}$$

Now we consider another distribution which we are often using from now on, it is known as Gaussian distribution here again the form of the distribution is given here. We are not familiar with it (())(12:15) by just going through this question, the Sigma Square is the variance again note that this is for a one-dimensional variable, so x is a one-dimensional variable and we are looking at one x and Sigma square is the variance, mu is the mean of the distribution okay and what does this expression evaluate?

So if you have a value x of that event, so this is for continuous variables so if you have a value x this expression will evaluate is the probability of observing x, what is the probability of getting x? okay so what is plotted here on the graph is the graph of the probability distribution wherein the x axis is s and y is the probability of observing x okay and the mean is an mu. This width of the distribution is 2 Sigma where Sigma is the standard deviations, Sigma square is the variance.

Now if you are not sure as to like okay what kind of variables lead to this kind of the distribution okay, so just to give you a practical example let us consider let us see a class with hundreds of students and then they take an exam, so you get the marks of every students, so we have a student serial number I will just have a table here okay. Serial number of students then marks okay for that particular test, so the student serial number goes from 1, 2, let us say 100 and the marks also well it will also go from in this case let us say they go from again 40 to 100 everybody passed let us say we have worked hard threshold of 40 and everybody got over 40.

You can imagine that marks will go from let us say 41, 41.5, so on and so forth but let us say nobody got less than 40 let us say something like that okay and somebody could have got 100 so on and so forth okay, so this is our data, we have about 100 students and the marks in an examination, so what you do is you construct a histogram. How do you construct a histogram? So let us take an axis, this is marks okay and it goes from just for the sake of convenience you have 4200 okay so to care to construct a histogram is you make them bins on your axis, so we will make bins of size 5 so 40 to 45 is one bin, 45 to 50 is another bin so on and so forth okay up to 100 okay, so what you do is you count the number of students whose marks fall between 40 to 45 okay let us say some number, this axis is the number of students okay this is the number of students.

So you count the number okay so you do that for let us say all of them, typically what you will get will be something along these lines I am drawing its move so you will get something like this. So you make bins of size 5, so 40 to 45, what are number of students who got marks between 40 to 45? What are the number of students what who got marks between 45 to 50 so on and so forth and you just plot it as a histogram wherein the x axis give you the…you can plot with the center of the bin, so instead I have just indicated here as 40 to 45.

Instead of doing that you can just say 42.5 and plot that number 52.5 and plot that number so on and so forth 47.5 and plot the number of students between 45 to 50, so you can plot that and you get this histogram okay and if you normalise the y-axis here the number of students by total number, what you get is the probability distribution of the marks that the students scored in that particular examination okay so it is a very good way of summarising what your class performance is like okay so you can easily say if you can let us say somebody gets 40 marks you can plug that and see you know whether is one of the few students, what are the probability that he got…how far away he is from the main thing of that sort.

So to summarise the statistics of a class this is a good way of doing it because then you can easily assigned grades using this okay, so that is one of the things that people do usually the calculate this curve and use this curve to assign grades right, so you know the mean, this is the mu we will see how the mean the standard deviation are estimated for a Gaussian distribution, something that is very familiar to you but the upfront assumption that you are making every time you do this calculation is that you are assuming that your data falls under Gaussian distribution right.

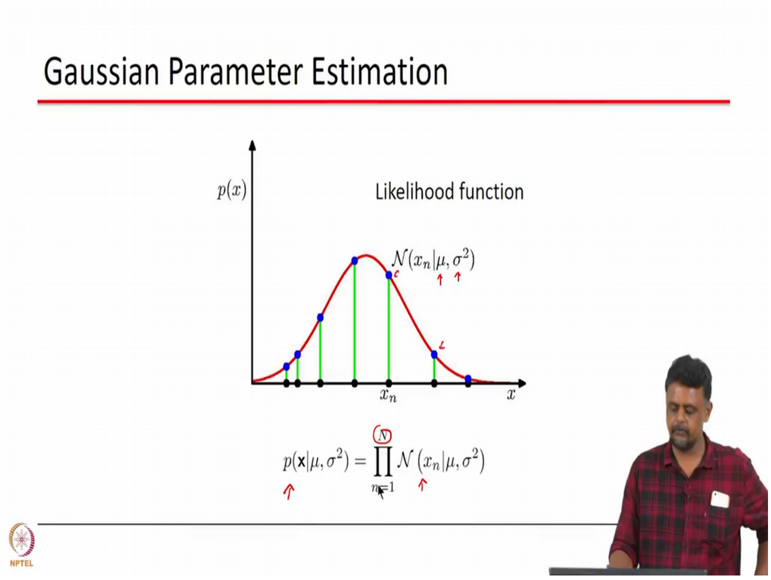
So I am giving you examples of marks, can do that for let us say you measure all the heights of all the students in your class okay you have a class of hundred or several hundred and you measure all their heights and if you plot again histogram of heights based on some bins then also the distribution will look very close to a normal distribution or for that matter many of the quantities that you measure might look like a Gaussian distribution because any calculation that you usually do makes this assumption, we will see what those calculations are but this is in 1d okay this is for 1d example, one-dimensional example where in you have marks okay.

Sometimes the variable x you are looking at is multi-dimensions let us say of dimensional you are used to the dimension small n but for the purposes of this lecture I am going to use capital D okay, so if your dimensionality of x is capital D so let us say x has many features that is one way of looking at it right. If x has many features then x should be some vector of link d, this mu will be a vector of link d because each dimension will have its mean so mu will also be like d.

This Sigma it is called the covariant metrics and it has size d cross d okay and this is the modulus is determinant of the covariant metrics square root okay so this is the determinant okay, so this is for a multi-dimensional x this is how at the Gaussian distribution formula is link okay so what you have plotted here in this figure basically is the let us say if d is 2 okay, 2 variable x1 and x2 okay and these red lines are contours of constant probability or constant probability mass function.

So for instance if you plug-in this x1 and x2 find out all values find out the range of all values of x1 and x2 for which n is similar or same then you can plot this…the locus of all those x1 and x2 is basically this okay so what it shows is that the x1 and x2 are co-related slide kind of right because x1 seem to…x2 seems to linearly increase with x1 okay. So this plots are that way useful to figure out whether there is co-relation between your variables when you are dealing with multi-dimensions variable okay. So this is a general form of the Gaussian distribution, so it comes in the exponent so you will be typically writing it as something like this for one-dimensional and you will have 1 over square root 2 pi Sigma okay.

(Refer Slide Time: 20:38)



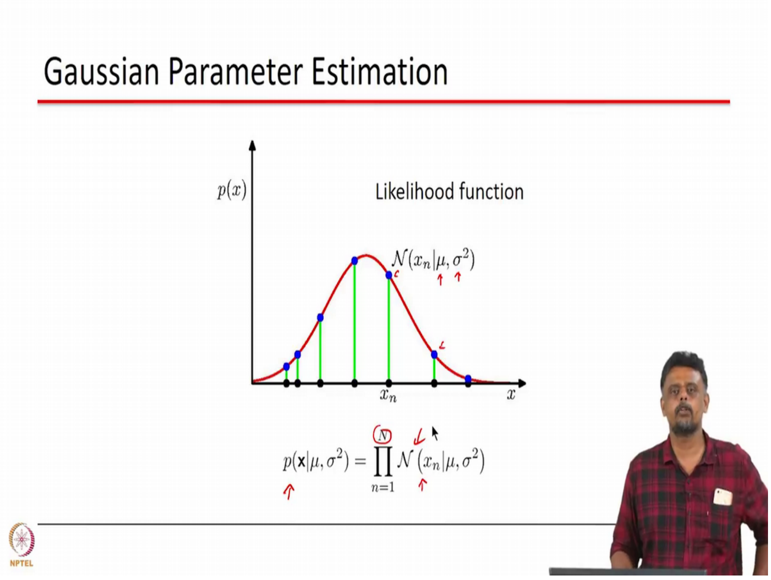
Gaussian Parameter Estimation

$$p(\mathbf{x}|\mu,\sigma^2) = \prod_{n=1}^{N} \mathcal{N}(x_n|\mu,\sigma^2)$$

So then for a Gaussian parameter estimation, so remember we saw for Bernoulli distribution we estimated, we looked at how we can estimate the Bernoulli parameter mu which is a probability x equal to 1right, so how do you do the similar things because we have the normal distribution, it is also called normal distribution or the Gaussian distribution is denoted as 2 parameters mu and sigma square, so we want to estimate mu and sigma square, how do you do this? Okay so recall that this is for continuous random variable, so x can take any real value but usually data is discreet right we only observe for specific x right that is the thing.

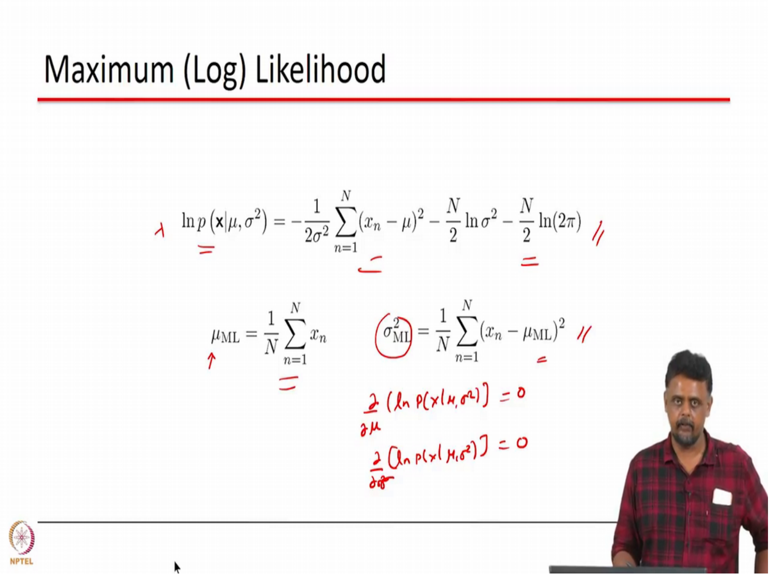
So we then again once again what you do is based on their observation, so this blue points or the observations we have, so based on this observation we calculate probability of observing the data or the likelihood. Likelihood of observing capital X refers to the data that we have observed, so let us say we have capital N points once again like we saw in the Bernoulli distribution the probability of observing the data set is the probability of observing each one of the individual data points and we know that the probability of observing each data point is given by the normal distribution, so the total probability of observing this data set is the product of each of the individual probability which is what is given by this expression okay.

(Refer Slide Time: 22:06)



Gaussian Parameter Estimation

$$p(\mathbf{x}|\mu,\sigma^2) = \prod_{n=1}^{N} \mathcal{N}(x_n|\mu,\sigma^2)$$



Maximum (Log) Likelihood

$$\ln p(\mathbf{x}|\mu,\sigma^2) = -\frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n-\mu)^2 - \frac{N}{2}\ln\sigma^2 - \frac{N}{2}\ln(2\pi)$$

$$\mu_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N} x_n \qquad \sigma^2_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N}(x_n-\mu_{\mathrm{ML}})^2$$

$$\frac{\partial}{\partial\mu}\left(\ln P(x|\mu,\sigma^2)\right] = 0$$

$$\frac{\partial}{\partial\sigma^2}\left[\ln P(x|\mu,\sigma^2)\right] = 0$$

So then what we do is similarly we take the log of that is likelihood and the log of the product transforms to summation of the logs which is what is given here. I urge you to actually work this out yourself because it is virtually algebra but it will also be comfortable with the expressions here. Now once we have this expression what we want to do like we saw earlier we use to maximise the log likelihood of observing the data right that is what we…mu and sigma square should be such that the probability of observing this data is very high, then the way to do that would be to take the derivative of this expression.

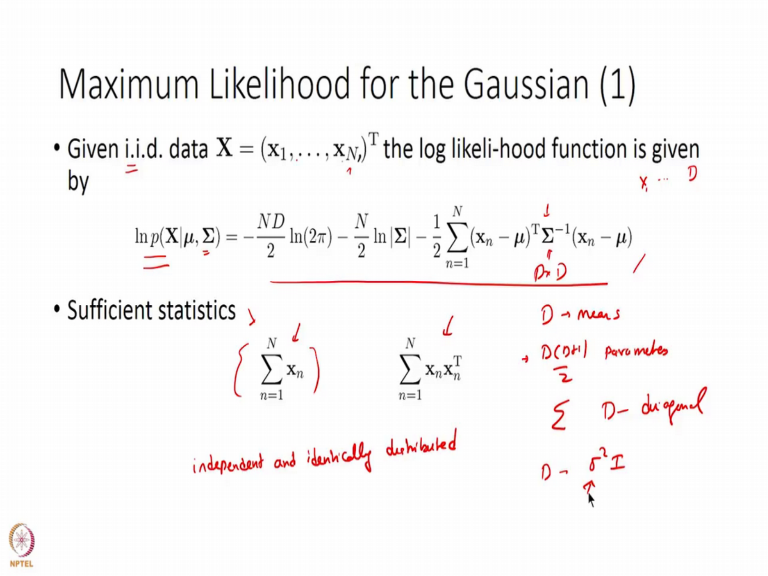
So d Delta over delta mu of log of P x square set that to 0 and then you will do Delta over delta sigma square of the same…also Delta Sigma set that to 0 and then once you solve for it you get 2 results which none…both of it should not surprise you. It says that the mu ML this is called the maximum likelihood technique mu ML is nothing but the average of your data points okay and your sigma square which the variants in this parameter in the Gaussian terminology is called variants which is nothing but the you calculate the variants of your data given mu ML that is what you get okay which is basically the mean of x minus mu ML square that is the mean of x minus mu ML square is the Sigma square value here.

So this is typically the statistics that you always calculate right whenever you have any data set or you are like we will characterise it by using the mean and standard deviation. When you do that what you are assuming is that that your data comes from a normal distribution, so your data set come from a normal distribution and these are the assumptions that you are implicitly making when you calculate mean and standard deviation, the assumption being that your data is normally distributed that they are drawn from this normal distribution with a particular sigma and particular mu okay.

So in the context of machine learning what we will again, we will model data using these distributions, so for instance where we have problems which involves 0 or 1 choice we will use the Bernoulli distribution and where the problems are continuous variable involved will use the Gaussian distribution and the problem… you will see that the problem and I will explain this in a later video, we will see that what you trying to do is to…we will end up modelling this mu using our data and for even linear (())(24:47) can be bought in this form we will do that in a video soon.

So we end up modelling or we are trying to estimate this mu sigma or in the case of the Bernoulli distribution again the parameter Bernoulli parameter mu those are what we try to estimate every time okay. As the output that we are looking at and implicitly… What I wanted to show with this video is that typically for a given data set of you calculate mu and Sigma okay. This is done implicitly in many of the models that you are using, we will see that in a later video.

(Refer Slide Time: 25:30)



Maximum Likelihood for the Gaussian (1)

- Given i.i.d. data $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_N)^{\mathrm{T}}$ the log likeli-hood function is given by

$x_1 \cdots D$

$$\ln p(\mathbf{X}|\mu, \Sigma) = -\frac{ND}{2}\ln(2\pi) - \frac{N}{2}\ln|\Sigma| - \frac{1}{2}\sum_{n=1}^{N}(\mathbf{x}_n - \mu)^{\mathrm{T}}\Sigma^{-1}(\mathbf{x}_n - \mu)$$

$D \times D$

- Sufficient statistics

$$\left(\sum_{n=1}^{N}\mathbf{x}_n\right) \qquad \sum_{n=1}^{N}\mathbf{x}_n\mathbf{x}_n^{\mathrm{T}}$$

independent and identically distributed

$D \rightarrow$ means

$\rightarrow \frac{D(D+1)}{2}$ parameters

Σ D- diagonal

$D \rightarrow \sigma^2 I$

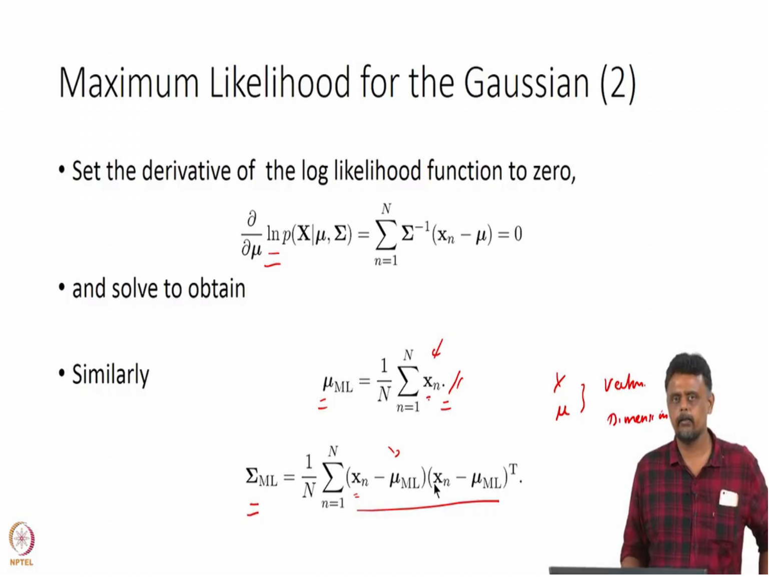
So if you want to consider a multi-dimensional data where each of the x1 have let us say D dimensions okay the procedure is the same so we still construct the log likelihood of the data given a dataset consisting of n points and independent and identically distributed that is what i.i.d stands for independent and identically distributed. So the probability of observing x1 times the probability of observing x2 times the probability of observing xN so for all the N data points the products of each of these probabilities is the probability of (())(26:18) in the entire dataset okay and we do that like we did it for the 1 time (())(26:23) case and we can still do the log of the probability and you will get an expression like this.

Once again can take the derivative with respect to mu remember now mu is a multi-dimensional variable and you also take the derivative with each of the common again sigma is a matrix here so it has d into d plus it has d square elements okay so once we do that we can do the similar process wherein taking the derivative with respect to new and this capital sigma set it to 0 to obtain the value of x and set okay.

In this context remember this, in order to describe a Gaussian probability distribution, in order to estimate or calculate a Gaussian probability distribution we just need 2 quantities one is this remember we just need the mean and this variance okay, so these are referred to as the sufficient statistics we call that right so we are going to need all the data points in fact that is one of the reason why you construct this probability distribution, if you have a large data set you can actually summarise the data set with just 2 parameters. In the one-dimensional case it is just mean and standard deviation, in the multi-dimensional case you will have more than that because remember this is d cross D matrix.

So you will have D means right because each dimensions has a mean and feature has a mean and this will turn out to be symmetric matrix so you will have D into D plus 1 by 2 parameters, independent parameters right because if the symmetric matrix these are the unique D to D plus 1 by 2 unique elements of sigma okay so it is still a lot remember if there are hundred features it is a lot of parameters to estimate, so typically when you are solving these problems using for probability techniques what people usually do is they assume that D is diagonal okay, so that you have only D parameters to estimate. In fact you can even…they assume that you know if D is given by some sigma square times identity matrix, so then you just have again only one parameter okay, so depending on how you model you can reduce the number of parameters you would like to estimate using multidimensional Gaussian okay.

(Refer Slide Time: 29:00)



Maximum Likelihood for the Gaussian (2)

- Set the derivative of the log likelihood function to zero,

$$\frac{\partial}{\partial \mu} \ln p(\mathbf{X}|\mu, \Sigma) = \sum_{n=1}^{N} \Sigma^{-1}(\mathbf{x}_n - \mu) = 0$$

- and solve to obtain

- Similarly

$$\mu_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N} \mathbf{x}_n.$$

$$\Sigma_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N} (\mathbf{x}_n - \mu_{\mathrm{ML}})(\mathbf{x}_n - \mu_{\mathrm{ML}})^{\mathrm{T}}.$$

So (())(28:59) we will set the derivatives of the log likelihood function to zero again with respect to mu and you can actually solve to obtain that for mu again for the multidimensional case is the mean of your data points again remember that these are vectors so for every dimensional you have to independently calculate the mu and once again the covariant matrix is given by this expression the average or for the expectation of your this term here x minus mu ML times x minus mu ML (())(29:35) again member x and mu are vectors of dimension D okay so we have looked at 2 distributions Bernoulli and Gaussian, Bernoulli distribution is used to describe variable that can take 0 or 1 values.

A typical example is coin toss, so it has only one parameter which is basically the probability of observing x equal to 1 okay x equal to one might respond to any event like for instance even we have talked about this is the coin toss where in x equal to 1 corresponds to heads x equal to 0 corresponds to tales, so what the distribution is characterised by probability of x equal to 1 which is given by mu and we also saw how we estimate mu, mu is just the number of…if you have given sequence of N coin tosses we just calculate the total number of coin tosses where it landed as heads which corresponds to the total number of N x equal 1 and the ratio of the total number where x equal to 1 divided by the total number of actual coin tosses this gives you the estimate of mu there are in the Bernoulli distribution.

For the Gaussian distribution there are 2 parameters one is the mean under standard deviation of covariance in higher dimensions and mean is basically the mean of your data points observed data points remember that you have to take the mean across every feature independently and the covariants is nothing but the covariants of x again it is 0 centered so x minus mu ML times x minus ML (())(31:19) the mean of this value is your covariants matrix.

Again remember x and mu are again D dimensional where D is the dimensionality of x. Okay so subsequent to this we will look at various techniques, leftover techniques in machine learning, we will look at SVM we will look at name base classifier and then we will move onto maximum like (())(31:48) destination, how would a price that says linear regression then maximum a posteriori methods and then finally Bayesian regression. We might postponed Bayesian regression to the next week but maximum likelihood estimation in map techniques we will look at this week. Thank you.