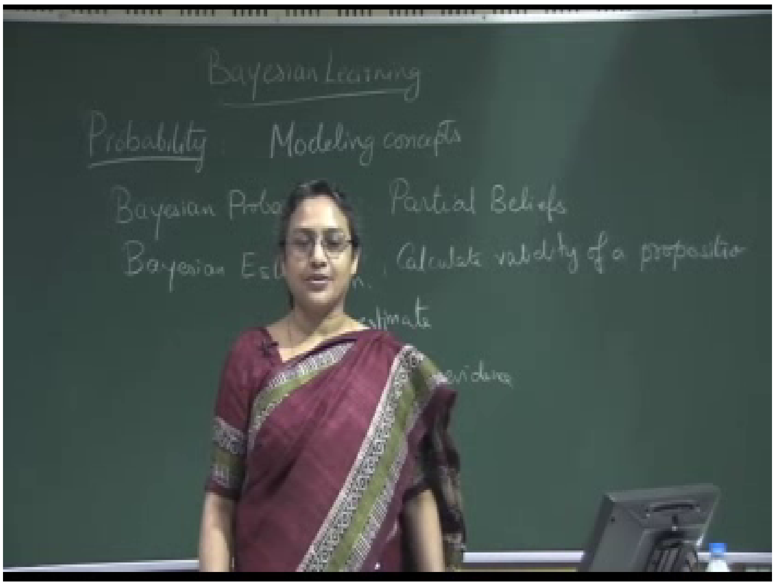
Introduction to Machine Learning
Prof. Sudeshna Sarkar
Department of Computer Science and Engineering
Indian Institute of Technology, Kharagpur

Lecture – 16
Bayesian Learning

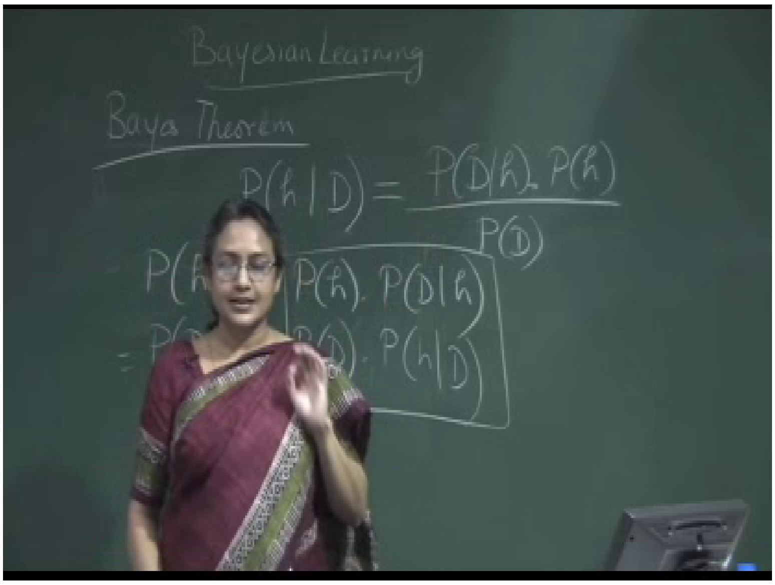
Good morning. Welcome to today's lecture. Today we will talk about Bayesian learning which is Part B of module 4. In the last class, we gave a crash course on probability and today we will see how probability is used for learning especially for classification, probability how it is used for modeling concepts.

(Refer Slide Time: 00:25)



So, Bayesian probability is the notion of probability which talks about partial beliefs. So, Bayesian probability talks about probability interpretation as partial beliefs and Bayesian estimation, it calculates the validity of a preposition. The validity of the preposition is calculated based on two things; number one the prior estimate. It is based on the prior estimate of its probability and in fact, new evidence, and new relevant evidence. Based on this, the posterior Bayesian estimation is done and the key to this is an important theorem called Bayes theorem which we will introduce now.

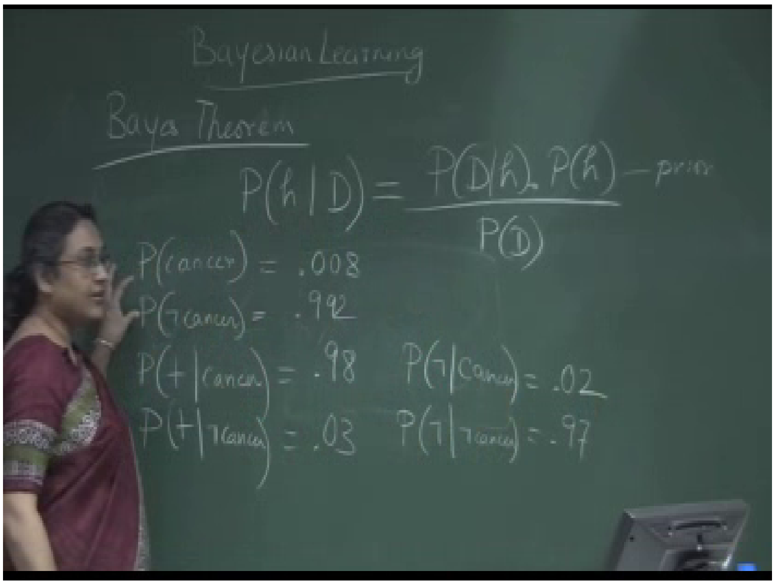
(Refer Slide Time: 02:45)



So, Bayes theorem deals with how to find the probability of a hypothesis given the data you have different possible competing hypothesis and you can find out the probability of the individual hypothesis given the data, so that you can find out which is the most probable or most likely hypothesis according to the Bayes theorem probability.

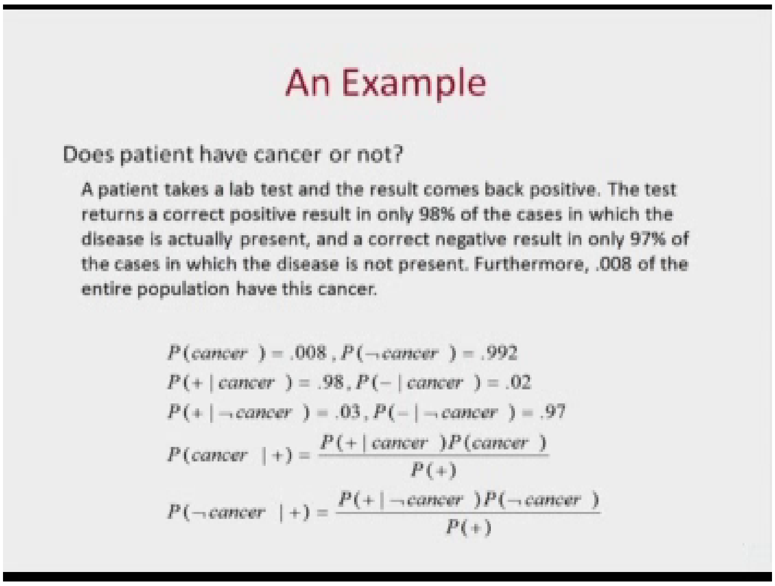
If hypothesis given data is given by probability D given h times prior probability of the hypothesis h divided P D. This is very easy to derive you know that by the law of products, you can see that probability h D equal to probability h times probability D given h and you can also because it is commutative; this is also equal to probability D h which is equal to probability D times probability h given D. So, if you consider these two equal, then by manipulating them you can come up with Bayes role. So, Bayes role is the most important formula form which we can look for at Bayes learning. So, P h is the prior probability of the hypothesis h. So, this is the prior probability.

(Refer Slide Time: 04:30)



Probability D given h is the probability of the data. If the hypothesis is true, what is the likelihood of that data being generated? If h was true what is the probability of D being generated and P D is the likelihood of the data. So, based on this, we have Bayes theorem. Now, let us see an application of Bayes theorem for this we may look at the slide.

(Refer Slide Time: 05:04)



An Example

Does patient have cancer or not?

A patient takes a lab test and the result comes back positive. The test returns a correct positive result in only 98% of the cases in which the disease is actually present, and a correct negative result in only 97% of the cases in which the disease is not present. Furthermore, .008 of the entire population have this cancer.

$P(cancer) = .008, P(\neg cancer) = .992$

$P(+ | cancer) = .98, P(- | cancer) = .02$

$P(+ | \neg cancer) = .03, P(- | \neg cancer) = .97$

$$P(cancer | +) = \frac{P(+ | cancer)P(cancer)}{P(+)}$$

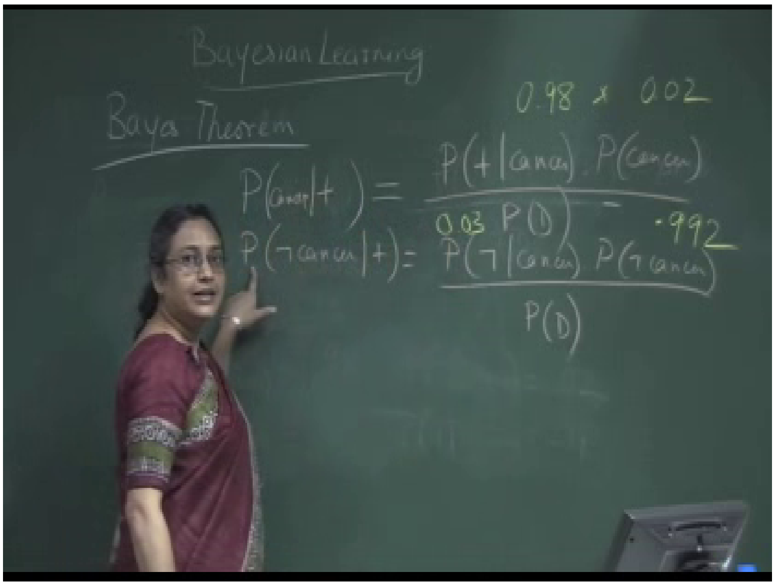
$$P(\neg cancer | +) = \frac{P(+ | \neg cancer)P(\neg cancer)}{P(+)}$$

Suppose you want to know whether a patient has cancer or not. So, this particular example is taken from Tom Meshes. Secondly, machine learning a patient takes a lab test and the result is positive. Now, the test returns a correct positive result in only 98 percent of the cases in which the disease is actually present and a correct negative result in only 97 percent of the cases in which the disease is not present.

Furthermore you know that 0.008 of the entire population have this cancer. So, we can write down this as probability of cancer that is the prior probability of cancer among the population is equal to 0.008 and therefore, probability of not cancer equal to 1 minus 0.008 that is 0.992. Now, what is the probability of the test being positive given that cancer is present? If this is given us 0.98 by this statement of the problem also probability of therefore, the probability of the test being negative given cancer is equal to 1 minus 0.98 that is 0.02. What is the probability of test being positive given not cancer? This is given by 1 minus 0.97 that is 0.03. Similarly probability of not given cancer is given to be 0.97.

So, these are the values that are supplied to you in the problem. Now, based on this we can use Bayes theorem. We want to find out the probability of cancer. The hypothesis is that the patient has cancer given that the probability that the patient has a cancer.

(Refer Slide Time: 07:31)

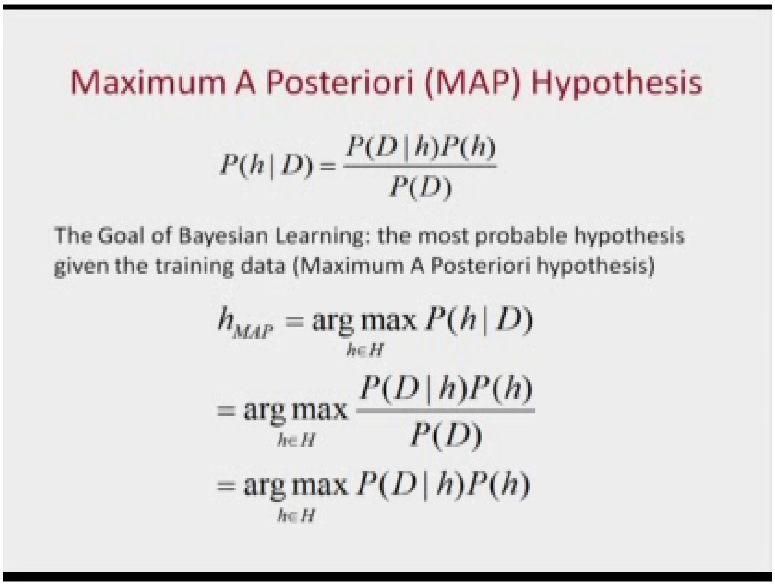

Bayesian Learning
Baya Theorem
0.98 x 0.02
P(cancer|+) = P(+|cancer). P(cancer) / P(D)
0.03
P(¬cancer|+) = P(¬|cancer). P(¬cancer) / P(D)
.992

Given that the test is positive, this is given by probability of positive given cancer times prior probability of cancer divided by probability of the data. Similarly you can write probability of not cancer given that the test is positive equal to probability of test being negative given cancer times probability of not cancer times probability of the data and now, you can put the values here.

What is probability of plus given cancer? It is 0.98. So, this is 0.98, this is 0.02 whereas, probability negative given cancer is 0.97, not sorry 0.03. So, this is 0.03 times 0.992. So, probability cancer given the test is positive is 0.98 into 0.002 divided by P D probability not cancer given the test is positive is 0.03 times 0.992 given P D divided by P D. Now, divide by P D this denominator is common to both the expressions. Secondly, probability of having cancer is proportional to 0.98, 0.02 and not cancer is proportional to 0.003 into 0.992. Based on this, you can figure out the probabilities.

So, this plus, this will sum to 1 and you can find out that this is more likely. So, it is more likely that the patient does not have cancer given this. So, this is an application of Bayes theorem.

(Refer Slide Time: 09:53)



Maximum A Posteriori (MAP) Hypothesis

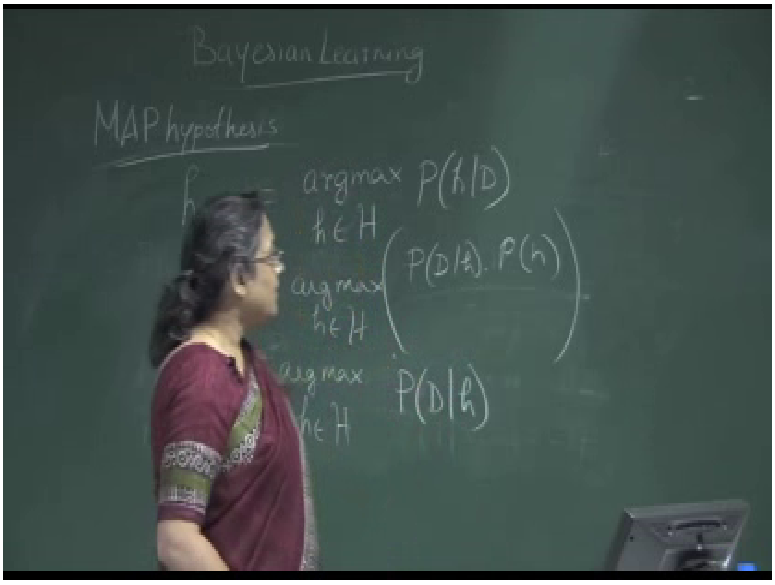
$$P(h \mid D) = \frac{P(D \mid h)P(h)}{P(D)}$$

The Goal of Bayesian Learning: the most probable hypothesis given the training data (Maximum A Posteriori hypothesis)

$$\begin{aligned} h_{MAP} &= \arg\max_{h \in H} P(h \mid D) \\ &= \arg\max_{h \in H} \frac{P(D \mid h)P(h)}{P(D)} \\ &= \arg\max_{h \in H} P(D \mid h)P(h) \end{aligned}$$

Now, the goal of Bayes learning; now, can Bayes theorem we apply it to find a hypothesis in machine learning. So, based on the Bayes theorem, we can find out the most likely hypothesis which is called the maximum posterior hypothesis.

(Refer Slide Time: 10:19)


Bayesian Learning
MAP hypothesis
$h = \underset{h \in H}{\operatorname{argmax}} P(h|D)$
$\underset{h \in H}{\operatorname{argmax}} \left(P(D|h).P(h) \right)$
$\underset{h \in H}{\operatorname{argmax}} P(D|h)$

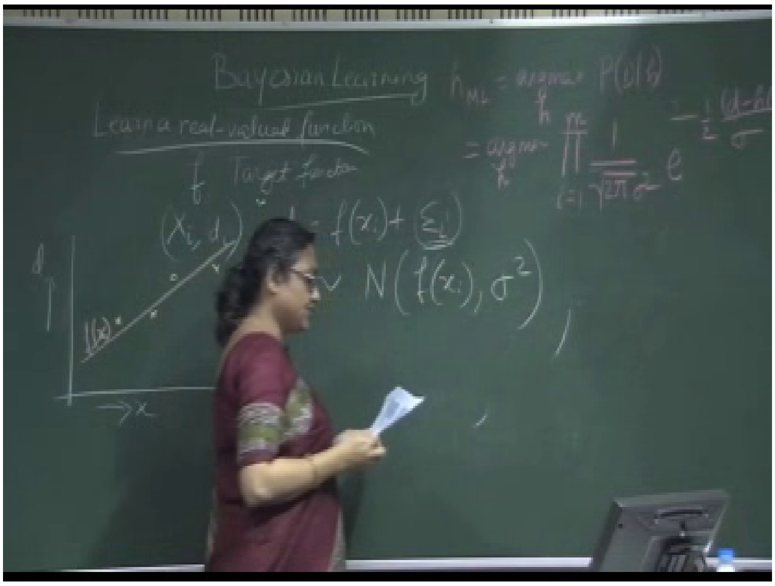
So, the map hypothesis is given by that value of h for which probability h given data is maximize. Now, by Bayes theorem, this is same as that hypothesis. We just expand this by Bayes theorem. We get probability D given h times probability h divided by P D. So, now, h capital H is the hypothesis space and small 8 out of all hypothesis in the hypothesis space, you want to find that hypothesis for which this expression is maximized.

Now, P D is independent of the particular hypothesis. So, we can say this is the same hypothesis for which this part is maximized. So, the posterior hypothesis, posterior probability is given by probability D given h proportional to probability D given h times D h and the maximum. Posteriori hypothesis is one for which probability D given h times P h is maximum. This is the prior probability of the hypothesis and we choose hypothesis based on their posterior probability.

Now, in the event if for all hypotheses, the probability is equal, then you choose that hypothesis for which probability D given h is maximum. So, 8 m l is the maximum likelihood hypothesis. It is applicable in those cases where the prior probability of all hypotheses is equal. That is initially before you have any data. The entire hypothesis are equally probable, in that case you choose the hypothesis for which the probability D given h is maximum; so the application of Bayesian theorem, in order to find out maximum a posteriori hypothesis and the maximum likelihood hypothesis.

Now, we will see an example of how in a finding the least squared line, we can apply the Bayes theorem to find out the most likely hypothesis. So, suppose you have to learn a real valued function.

(Refer Slide Time: 13:29)


Bayesian Learning
Learn a real-valued function
f Target function
(x_i, d_i)
$= f(x_i) + \varepsilon_i$
$\sim N(f(x_i), \sigma^2)$
$h_{ML} = \arg\max_h P(D|h)$
$= \arg\max_h \prod_{i=1}^{m} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\frac{(d_i - h}{\sigma}}$
d
f(x)
x

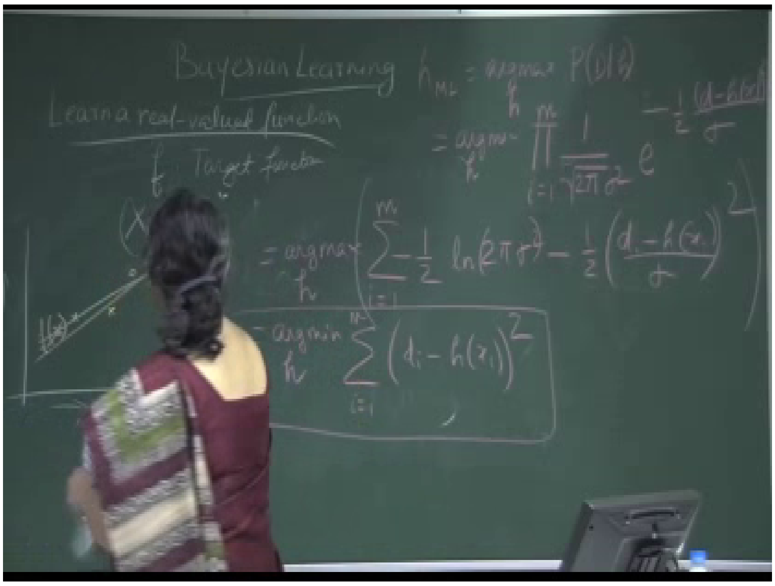
We have already talked about leaner regression which can be used to learn a real valued function and suppose the data is generated in the following fashion. So, there is a target function f. F is the target function and the individual data are generated. So, the data is given as x i D i are the individual data points and D i is generated as f x i plus epsilon i.

Epsilon i is the error and we assume that this error follows a normal distribution with mean zero and a standard deviation sigma. So, we can think that D i is coming from a normal distribution whose mean is f x i and whose error is given by sigma square, where sigma square is the variance corresponding to this error term. So, this is how the data is being generated and let us assume that epsilon i is independently generated for the individual instances where epsilon i are independent for in different instances, and it is a Gaussian with zero mean and variance sigma square and therefore, we can say that data is generated as normal distribution f x i sigma square.

So, what we have is that this is our x and this is our d, suppose this is the true function. So, this is f x and the data that we get are let us say generated like this. So, these are the data points that we have. Now, we have to find a function which estimates f. Now, how do we find this function? Let us say we use the maximum likelihood hypothesis.

So, what is h m l? H m l is the maximum likelihood hypothesis which is given by that hypothesis for which probability D given h is maximum. Now, what is this? This is this arg max h and probability D given h is given by product. Over all the training examples 1 by route over 2 pi sigma square e to the power minus half D minus h x i whole square by sigma because they follow the Gaussian distribution. Now, this can be written as let me rub out this portion of the board, so that we can write this formula here.

(Refer Slide Time: 17:21)

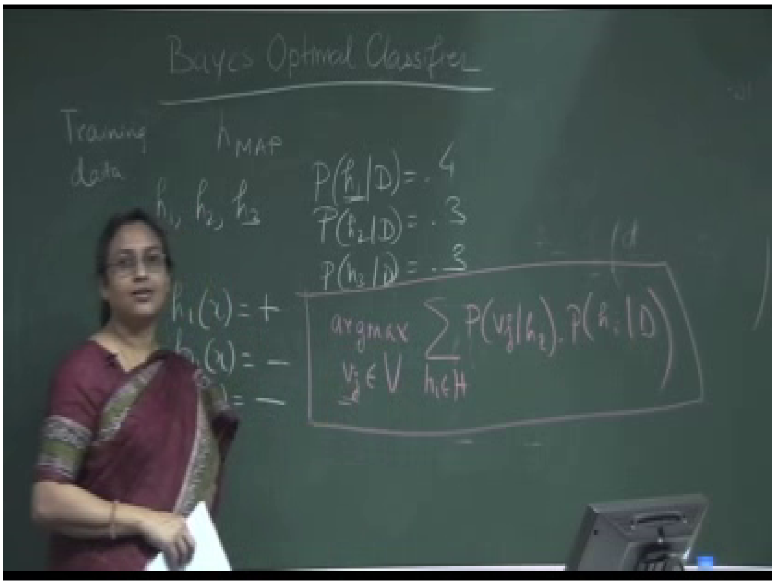

Bayesian Learning
Learn a real-valued function
f Target function
$h_{ML} = \arg\max_h P(D|h)$
$= \arg\max_h \prod_{i=1}^{m} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{d-h(x)}{\sigma}\right)^2}$
$= \arg\max_h \sum_{i=1}^{m} -\frac{1}{2}\ln(2\pi\sigma^2) - \frac{1}{2}\left(\frac{d_i - h(x_i)}{\sigma}\right)^2$
$= \arg\min_h \sum_{i=1}^{m} (d_i - h(x_i))^2$
f(x)

So, this turns out to be Arg max h. So, that function which maximizes this product is the same as which maximizes the sum of the logs. So, we convert it to the log do mine which is summation i equal to 1 to m, where m is the number of training examples half l n. So, we have taken logarithm of this part. So, it is half l n. So, minus half minus half l n 2 pi sigma square minus half D i minus h x i by sigma whole square by taking logarithm, we can get this, so by simplifying what we get? It is that function for which sigma i equal to 1 to m D i minus h x i whole square is minimized.

Why? It is because this part is constant. When I am taking the hypothesis for which this expression is maximized, this part does not play a role because this is constant. So, this part plays a role which arg max hypothesis minus of half by D i minus h x i by sigma whole square. So, half we can ignore because whatever maximizes minus half of that also maximizes only this part. So, if you want to maximize negative of this, it is the same of minimizing the positive part of this. So, it is that hypothesis, the maximum likelihood hypothesis for this linear regression problem is that hypothesis for which D i minus h x i whole square is maximized and this is exactly the least square criteria.

So, based on this, we will get a function and that function could be something like this, but this is that function for which the sum of square errors is maximized. So, this is the Bayesian explanation to why we would choose a sum of square error to minimize in order to find out the linear regression. Now, next we will study about what is bayes optimal classifier.

(Refer Slide Time: 20:31)



Now, the question is suppose we are given some training data which each of the training instances we are given the class that it belongs to, then we are given a test instance and we are asked what is the optimum classification of x. The live answer would be that you find out the most probable hypothesis using the map criteria and then, you apply that hypothesis to the test example, but this is not necessarily the case. So, if you are given the training data form, the training data we learn h map. So, h map is the most probable hypothesis, but h map is not the most probable classification.

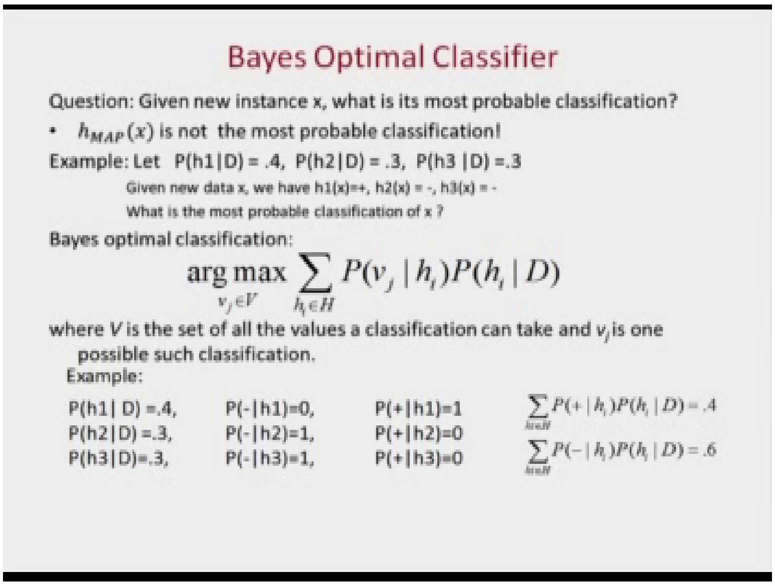
For example, suppose h1, h2, h3 are three candidate hypothesis belonging to the hypothesis space and suppose probability h1 given D is 0.4, probability h2 given D is 0.03 and probability h3 given D is 0.3. So, which is the map classifier? H1 is the map classifier because it has the maximum posterior probability.

Suppose we are a new data x and suppose h1 x is positive, h2 x is negative, h3 x is negative, what is the most probable classification? The most probable hypothesis is h1. H1 is saying that x is positive, but h2 and h3 both are saying that h is negative. So, the most in this case, the most probable classification would be actually negative because the sum of the probabilities of these two hypotheses is 0.6 which is larger than the probability of this hypothesis is 0.4.

So, what we have is, we have what we call the Bayes optimal classification. In Bayes optimal classification for a particular example we take the class to be. So, the capital V is the set of all possible classes, u hypothesis your algorithm will output that class included in all the classes for which the summation over all the hypothesis included in the hypothesis space probability v j given h i times probability h i given D is accepted. So, this is called the Bayes optimal classifier.

Bayes optimal classifier will output that class for a classification problem for which if you take summation over the entire hypothesis space of probability v j given h i times probability h i given D that will be maximum. So, in order to find out the Bayes optimal classifier, you have to for that test instance, you have to apply the possible hypothesis on that test instance in order to find out the bayes optimal classification. This is the optimal classifier, but this turns out to be interactive. So, we can quickly look at the slide here.

(Refer Slide Time: 24:57)



Bayes Optimal Classifier

Question: Given new instance x, what is its most probable classification?

- $h_{MAP}(x)$ is not the most probable classification!

Example: Let P(h1|D) = .4, P(h2|D) = .3, P(h3 |D) =.3

Given new data x, we have h1(x)=+, h2(x) = -, h3(x) = -

What is the most probable classification of x ?

Bayes optimal classification:

$$\arg\max_{v_j \in V} \sum_{h_i \in H} P(v_j \mid h_i) P(h_i \mid D)$$

where V is the set of all the values a classification can take and v_j is one possible such classification.

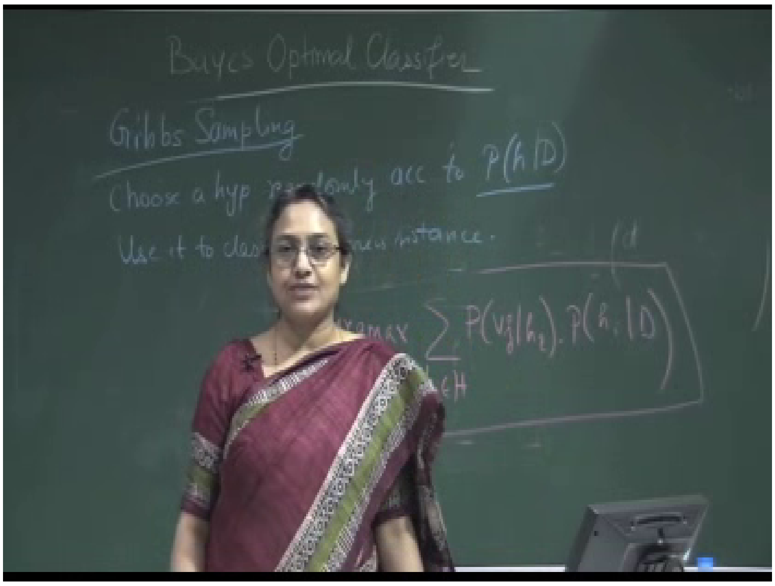
Example:

P(h1| D) =.4,	P(-|h1)=0,	P(+|h1)=1	$\sum_{h_i \in H} P(+ \mid h_i) P(h_i \mid D) = .4$
P(h2|D) =.3,	P(-|h2)=1,	P(+|h2)=0	$\sum_{h_i \in H} P(- \mid h_i) P(h_i \mid D) = .6$
P(h3|D)=.3,	P(-|h3)=1,	P(+|h3)=0	

So, the Bayes optimal classification is given by that v j for which sigma h included in capital H. Secondly, this is maximum and this is an example in that we can work out which you have seen probability h1 given D is 0.4, h2 given D is 0.3, h3 given D 0.2. Therefore, probability negative given h1 is zero negative, given h 2 is one probability negative, given h 3 is 1 and probability plus given h 1 is zero and if you apply the bayes optimal classifier, we see that probability of plus is 0.4, probability of minus is 0.6.

So, why is this called optimal? It is optimal in the sense that no other classifier using the same hypothesis space and same prior knowledge can outperform this on the average. So, this is called the bayes optimal classifier, but as you can see since typically the size of the hypothesis space is huge, it is not possible to apply the bayes optimal classifier. So, we have to use some approximation of the bayes optimal classifier and for that we can use Gibbs sampling.

(Refer Slide Time: 26:22)


Bayes Optimal Classifier
Gibbs Sampling
Choose a hyp randomly acc to P(h|D)
Use it to class new instance.
argmax $\sum_{h_i \in H} P(v_j|h_i).P(h_i|D)$

So, what we do in Gibbs sampling is that instead of applying all possible hypothesis on x v sample from the hypothesis space, we choose a hypothesis randomly according to probability h given D. So, for each hypothesis we have a probability associated with it. So, we have a probability distribution over the hypothesis space based on our training data that is our evidence we get a posterior probability distribution over the hypothesis space. In the bayes optimal classifier, each of the hypothesis according to their probability will apply on the each of the hypothesis will apply on the test instance.

Secondly, their contributions according to their posterior probabilities, but in Gibbs sampling, we will choose a randomly high hypothesis according to P h by t and use it to classify the new instance.

So, we just choose one hypothesis from the distribution and use it to classify the newsiest. Fortunately it is a surprising result that it has been found that the error for Gibbs algorithm is quite bounded. So, if the expected value is taken over the target hypothesis drawn at random according to the prior probability distribution, then the expected error of the Gibbs classifier is less than equal to twice the error of the Bayes optimal classifier.

So, the Gibbs classifier which is very much practical in the sense that you can need only to apply one hypothesis you know if once the posterior distribution has been computed, Gibbs sampling can be used to choose one hypothesis which can be used to classify the instance and it gives an error which is no more than twice the error of the Bayes optimal classifier. With this we come to conclusion of today's lecture.

Thank you.