

**Artificial Intelligence**  
**Prof. Sudeshna Sarkar**  
**Department of Computer Science and Engineering**  
**Indian Institute of Technology, Kharagpur**  
**Lecture - 38**  
**Probabilistic Learning**

Today we have our last lecture on machine learning. Today we will talk about probabilistic learning and may be discuss about computational learning theory. Earlier you have learnt about uncertainty and probabilities. So what we discuss now is how Bayesian learning can be used for the sort of concept learning tasks.

(Refer Slide Time: 01:17)



You have already looked at some simple machine learning algorithms that are used for concept learning. For example, you have studied decision trees and looked at algorithms for top down training of decision trees. In the last two classes we have also looked at neural network so mainly multilayer feed forward neural networks and we have seen how to train the neural networks using algorithms like back propagation so that the neural network can recognize the concept. Today we will see that the Bayesian learning framework also provides some ways of concept learning and we will also discuss the applications of these things as well as their limitations.

Why Bayesian learning is as important framework to study?

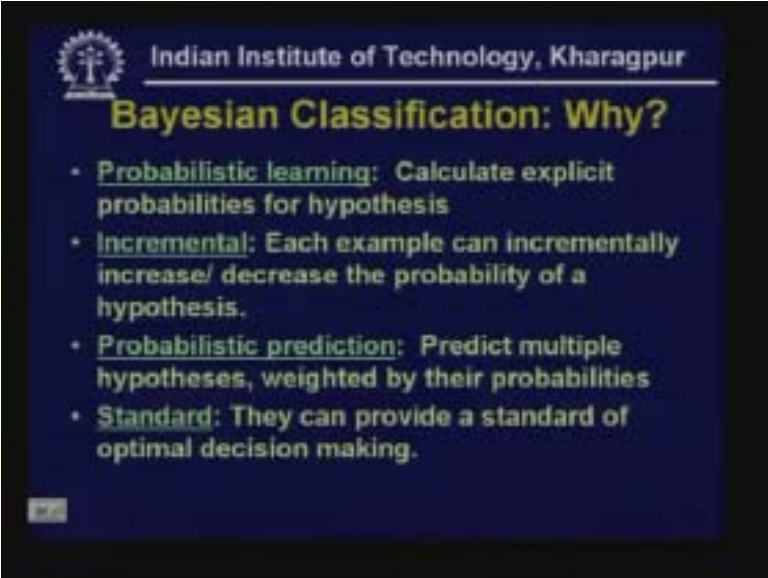
Firstly, Bayesian learning provides many practical learning algorithms. For example, today we will discuss a very simple learning called Naïve Bayes learning which is especially easy to learn and to train. And then we will also discuss how the surprising simple method works quite well for certain classification tasks and it is widely used for applications such as text classification and other areas. Then if you want to have more

sophisticated methods you have already learnt about Bayesian belief networks it is possible to learn also Bayesian belief network. However, we will discuss that in the Bayesian learning framework it is possible to easily combine prior knowledge.

For example, in the concept learning problem what you are trying to find out you are trying to learn a concept, you are trying to represent a concept and you are trying to find out an hypothesis which is a good representation of the concept. If you have some prior knowledge on which hypothesis are more likely for example we will see that the Bayesian framework allows you to integrate this prior knowledge into your learning methods so that the posterior hypothesis that you output depends on your prior probabilities or prior knowledge of the hypothesis.

Also, Bayesian learning provides many foundations for machine learning. And the Bayesian description Bayesian approach can be used to evaluate learning algorithms including other learning algorithms. And it can be used to guide the design of new algorithms and it allows us to learn from models or meta learning. Bayesian learning is a very important concept.

(Refer Slide Time: 05:29)



Indian Institute of Technology, Kharagpur

### Bayesian Classification: Why?

- **Probabilistic learning:** Calculate explicit probabilities for hypothesis
- **Incremental:** Each example can incrementally increase/ decrease the probability of a hypothesis.
- **Probabilistic prediction:** Predict multiple hypotheses, weighted by their probabilities
- **Standard:** They can provide a standard of optimal decision making.

How is Bayesian classification useful?

Bayesian classification is useful for probabilistic learning. In probabilistic learning we calculate the explicit probabilities for hypothesis. For example, if you have some training example you can output more than one hypothesis and you can attach probabilities with the different hypothesis. For example, let us say you have a decision tree so in neural network or decision tree you give only one hypothesis. When you are giving your training examples you come up with one hypothesis. But in probabilistic learning framework it is possible to you to come with multiple hypotheses and also associate the confidence of probability that you have in each of these hypothesis. The second property

of Bayesian classification is that it can be used in an incremental fashion. Each example can incrementally increase or decrease the probability of a hypothesis.

You can find out for the posterior probability of the hypothesis and when you have processed a number of examples you have attached certain posterior probabilities with different hypothesis. For example, your probabilities can get updated incrementally. So, this is one advantage of Bayesian classification techniques.

Thirdly, it allows you to make probabilistic prediction. That is, you can predict multiple hypotheses and you can associate with each of them their probabilities. Also the Bayesian framework provides a standard of optimal decision making. Even when you are not using this technique but using some other technique you can try to evaluate how well the technique corresponds to this model. We can try to evaluate the type of the hypothesis that a learning algorithm outputs whether it satisfies the different criteria that we will study.

(Refer Slide Time: 07:51)

Indian Institute of Technology, Kharagpur  
Basic Formulas for Probabilities

**Product Rule**  
 $P(A, B) = P(A | B)P(B) = P(B | A)P(A)$

**Sum Rule:**  
 $P(A + B) = P(A) + P(B) - P(AB)$

**Theorem of Total Probability :** if events  $A_1, \dots, A_n$  are mutually exclusive

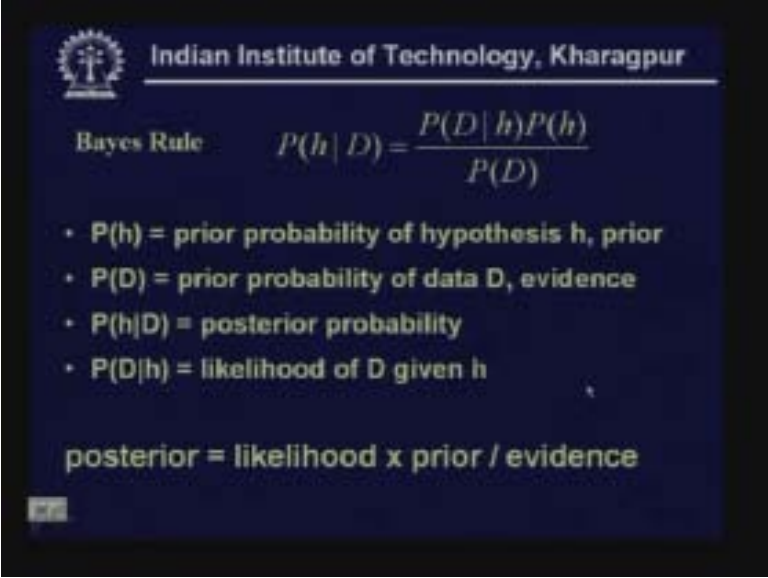
$$P(B) = \sum_{i=1}^n P(B | A_i)P(A_i)$$

Some basic formula for probabilities:

For example, the product rule is used to find out the product of the probabilities of A and B. Now P(A and B) can be written as the product of condition of P(A by B) times the P(B). As this is commutative the P(A and B) is the same as the P(B and A). Therefore it can also be written as P(B by A) times the P(A). We can combine this expression P(A by B) times P(B) is equal to P(B by A) times P(A) to derive a very important learning hypothesis called the Bayesian hypothesis. Here are some other rules of combining probabilities. Sum rule allows you to find the probability of the sum of two random variables P(A plus B) is equal to P(A) plus P(B) minus P(A) and B or A intersection B. So this is called sum rule. We have the theorem of total probability. If we have events  $a_1, a_2, \dots, a_n$  and these events are mutually exclusive then we can say that P(B) is equal to sum over all i's P(B by  $A_i$ ) times P( $A_i$ ). So we can condition on the different  $A_i$ 's and we can

combine these the weighted sum to get the P(B). These three formulas are often used in different manipulation of probabilities. Another very important rule is the Bayes Rule.

(Refer Slide Time: 10:20)



The slide features the IIT Kharagpur logo and name at the top. Below it, the Bayes Rule formula is presented as  $P(h|D) = \frac{P(D|h)P(h)}{P(D)}$ . A list of four items explains the terms: P(h) is the prior probability of hypothesis h, P(D) is the prior probability of data D, P(h|D) is the posterior probability, and P(D|h) is the likelihood of D given h. At the bottom, a summary states 'posterior = likelihood x prior / evidence'.

This expression can be used to derive Bayes Rule. From this expression we can write the P(A by B) is equal to P(B by A) times P(A) by P(B). And this we can rewrite in the form of P(h by D) is equal to P(D by h) times P(H) by P(D). This follows from the previous expression that we saw.

Now let us see how we can interpret this. If you take that H represents the hypothesis and D represents the data or evidence. So, in the concept learning problem you are given some training examples or the evidence and you have to find out a hypothesis which describes the concept. So this can be represented probabilistically as finding out for different hypothesis. You can consider different hypothesis and find out the probability of that hypothesis given the data. And you can evaluate this probability for the entire possible hypothesis and output that hypothesis which is most likely.

What we are trying to do is evaluate the P(A) by hypothesis by data or evidence that we have. Now we can rewrite P(h by D) in terms of probability D by h. Why we do this rewriting? It is often easier to evaluate the P(D by h). So, if you evaluate P(D by h) from that we can find P(h by D). But we need some other factors for example we need P(H) and P(D). Let us see what these mean. So P(H) can be interpreted as the prior probability of the hypothesis H. So, before you have got any evidence you have a set of possible hypothesis that you are considering and you have your initial belief about the probabilities of each of these hypotheses.

Each of these hypotheses may be equally likely or may be you consider that some of these hypotheses are more likely than the other. So P(H) denotes your belief about the prior probability of the hypothesis H and we can refer to this as the prior. Now what does

$P(D)$  stands for? It is the prior probability of the data  $D$  or the evidence which is not very intuitively clear what it means. But  $P(D)$  is the prior probability of that data.

What is  $P(h \text{ by } D)$  signify? It is the probability of hypothesis given the data. So after you have seen the evidence what is your posterior probability? What is your current belief, posterior belief after seeing the data about the hypothesis? This is called  $P(h \text{ by } D)$ . After you have seen some training examples  $D$  you want to find out based on this training example what is the probability of each of the hypothesis. And as you increase your training examples this will be updated. Then what is  $P(D \text{ by } h)$  signify? It means the likelihood of the  $(D \text{ by } h)$ . So you have some hypothesis  $H$ .

Now what is the likelihood that the hypothesis could rise to the data?

The  $P(H \text{ by } d)$  is the posterior probability is equal to  $P(h)$  the prior probability, this is the prior probability and this is the posterior probability. And what is  $P(D \text{ by } h)$ ? It is the likelihood of the data given the hypothesis and this  $P(d)$  is the evidence that you have. This is evidence and this is likelihood. You can write it as posterior is equal to likelihood into prior by evidence. So this is the base rule that we have  $P(H \text{ by } d)$  is equal to  $P(D \text{ by } h)$  into  $P(H)$  by  $P(D)$ . Generally we want most probable hypothesis given the training data, the hypothesis which has the maximum posterior probability. We refer to it as  $h_{\text{map}}$ .  $h_{\text{map}}$  is called the map hypothesis.

(Refer Slide Time: 16:20)

Indian Institute of Technology, Kharagpur

Bayes Rule  $P(h | D) = \frac{P(D | h)P(h)}{P(D)}$

Generally we want the most probable hypothesis given the training data

$$h_{\text{map}} = \arg \max_{h \in H} P(h | D)$$
$$= \arg \max_{h \in H} \frac{P(D | h)P(h)}{P(D)}$$
$$= \arg \max_{h \in H} P(D | h)P(h)$$

The map hypothesis is that hypothesis for which  $P(h \text{ by } D)$  is maximum among all the hypothesis that you are considering. We know what a hypothesis space is. It is the set of hypothesis which you are considering. Your hypothesis space can finite and finite but suppose your hypotheses is finite and suppose there are hundred hypothesis in the hypothesis space.

You want to find out the probabilities, the posterior probabilities for each of the hypothesis so that you can find out that hypothesis for which the posterior probability is the highest. So  $h_{MAP}$  is that value of  $h$  for which the  $P(h \text{ by } D)$  is maximized. This can be written as  $\text{argmax}_{h \in H} P(h \text{ by } D)$  included in  $h$ ,  $h$  included in the hypothesis space  $H$   $P(H \text{ by } d)$ .

Now if we apply Bayes Rule to this expression we can substitute the right hand side of this equation so we can say that  $h_{MAP}$  is that hypothesis for which  $P(D \text{ by } h)$  times  $P(H)$  times by  $P(D)$  is maximized. Now when you inspect this expression you are trying to find out that value of  $h$  for which the combination of these three terms is maximized out of which the particular hypothesis we are considering affects these two terms but not this term. So, that hypothesis for which this entire thing is maximized the same hypothesis for which only the numerator is maximized. So a map hypothesis is a hypothesis for which  $P(D \text{ by } h)$  times probability  $h$  is maximized.

(Refer Slide Time: 19:24)

Indian Institute of Technology, Kharagpur

## Choosing Hypotheses

Maximum a posteriori hypothesis  $h_{MAP}$

$$h_{MAP} = \text{argmax}_{h \in H} P(D|h) P(h)$$

$P(h)$  → prior

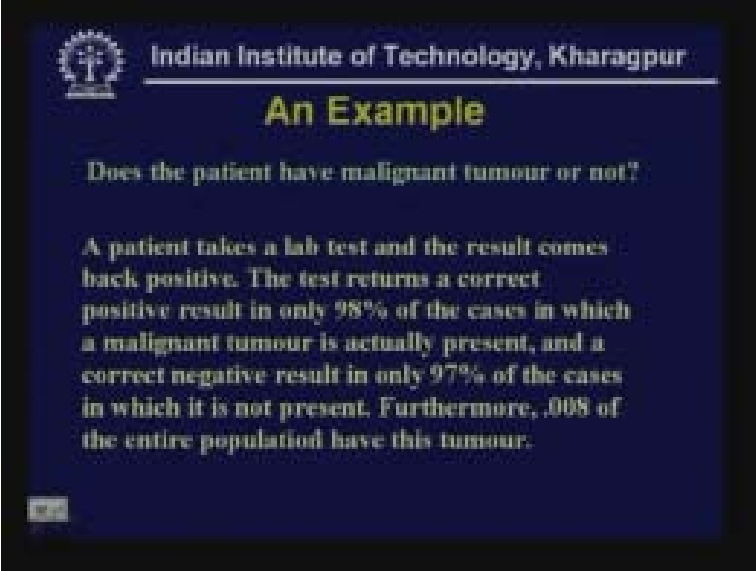
If the priors of hypothesis are equally likely  
 $P(h_i) = P(h_j)$  then one can choose the maximum likelihood (ML) hypothesis

$$h_{ML} = \text{argmax}_{h \in H} P(D|h)$$

Suppose you are given a learning problem and you are able to find out these different probabilities you would like to choose the maximum a-posteriori hypothesis  $h_{MAP}$  and  $h_{MAP}$  is given by  $\text{argmax}_{h \in H} P(D \text{ by } h)$  times  $P(H)$ . Now, as we discussed  $P(h)$  denotes the prior probability of the different hypothesis. Suppose that initially you consider all hypothesis to be equally likely then you can say that  $P(H)$  is the same for all  $h$  included in the hypothesis space. In that case  $h_{MAP}$  would be that hypothesis for which  $P(D \text{ by } h)$  is maximized if you consider  $P(H)$  is the same. So such a hypothesis for which only  $P(D \text{ by } h)$  maximized is called the maximum likelihood hypothesis and this is called a maximum a-posteriori hypothesis. Now map and ML hypothesis gives you the same hypothesis if the prior probability of the entire hypothesis are equal otherwise they might give you different values.



(Refer Slide Time: 19:24)



The slide features the IIT Kharagpur logo and name at the top. The title 'An Example' is in yellow. The main text is in white on a dark blue background. It asks a question about a patient's tumor status and provides test accuracy statistics and a population prevalence rate.

Indian Institute of Technology, Kharagpur

### An Example

Does the patient have malignant tumour 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 a malignant tumour is actually present, and a correct negative result in only 97% of the cases in which it is not present. Furthermore, .008 of the entire population have this tumour.

Here is an illustration of what we mean by prior, posterior and how they affect the decision making. Here is a simple standard problem. In this problem your objective is to find out whether a patient has a malignant tumor or not. And you have the following:

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 a malignant tumor is actually present, and a correct negative result in only 97% of the cases in which it is not present. Furthermore, it is known that only 0.008 of the entire population have this particular tumor. So this is the data that you are given. What you have to find out is whether the patient is likely to have malignant tumor.

Here the two possible hypotheses are; the patient has malignant tumor and the patient does not have malignant tumor. You have to find out whether the probability of the patient having malignant tumor given your evidence is higher or the probability of not having malignant tumor given the evidence is higher. What is the evidence? The evidence is that the patient takes a lab test for which the result is positive so the result of the lab test is positive and the lab test is taken to identify whether the patient has tumor or not. However, the lab test is not 100% accurate.

Now you are told that, if the patient does have malignant tumor the lab tests returns yes only in 98% of the cases. In 2% of the cases the lab test returns wrong result. It tells you that you do not have tumor but actually you have tumor. Also it tells us that if the malignant tumor is not present then in 97% of the cases the lab test returns negative. And 3% of the cases the lab test returns positive. Therefore based on this you have to find out if the result is positive how likely the patient is to have malignant tumor. In addition to this you are given some other data that your prior probability on malignant tumor and not malignant tumor are not equal. Hence, only 0.008 of the entire population has this tumor. That is, 0.992 of the population does not have tumor. So the prior probability of tumor is only 0.008 and the prior probability of no tumor is 0.992. **Here is a way to proceed in this**

**problem.** From the description given in the previous slide we can write that  $P(\text{tumor})$  is equal to 0.008 so we can write the  $P(\text{tumor})$  is equal to 0.008 and  $P(\text{not tumor})$  is equal to 0.992. These are your prior probabilities. Now you are also given that, if there is tumor then  $P(\text{lab test})$  being plus is 0.98 and if there is tumor  $P(\text{negative result})$  in the lab test is 0.02.

(Refer Slide Time: 24:15)

Indian Institute of Technology, Kharagpur

### An Example

$$P(\text{tumor}) = .008, \quad P(\text{not-tumor}) = .992$$

$$P(+ | \text{tumor}) = .98, \quad P(- | \text{tumor}) = .02$$

$$P(+ | \text{not-tumor}) = .03, \quad P(- | \text{not-tumor}) = .97$$

$$P(\text{tumor} | +) = \frac{P(+ | \text{tumor})P(\text{tumor})}{P(+)}$$

$$P(\text{not-tumor} | +) = \frac{P(+ | \text{not-tumor})P(\text{not-tumor})}{P(+)}$$

You are further given that if there is no tumor then  $P(\text{plus})$  in lab test is 0.03 and  $P(\text{negative})$  in lab test is 0.97. Now, the evidence that you are given is that the lab test returns positive and you have to find out the posterior probability of the two possible hypothesis  $P(\text{tumor by plus})$  and  $P(\text{no tumor by plus})$ . So  $P(\text{tumor by plus})$  can be written by using Bayes Rule as  $P(\text{plus by tumor}) P(\text{tumor})$  by  $P(\text{plus})$  and  $P(\text{not tumor by plus})$  is equal to  $P(\text{plus by not tumor})$  times  $P(\text{not tumor by } P(\text{plus}))$ . Now you can evaluate each of these values from the data you are by 0. What is  $P(\text{plus by tumor})$ ?  $P(\text{plus by tumor})$  is equal to 0.98 and  $P(\text{tumor})$  is equal to 0.008.

What is  $P(\text{plus by not tumor})$ ?

$P(\text{plus by not tumor})$  is equal to 0.03 and  $P(\text{not tumor})$  is equal to 0.992.

And what is  $P(\text{plus})$ ?

It is not known to us but it does not matter. When you have  $P(\text{tumor by plus})$  and when  $P(\text{not tumor by plus})$  we want to find out which one of them is higher. You can find out which one of them is higher by only evaluating the numerator. Also you know that probability of these two sums to 1. You know that any random variable  $P(A)$  and  $P(\text{not } A)$  is equal to 1. So you can take this  $P(\text{plus})$  as some constant alpha and you can say alpha times 0.98 into 0.008 plus alpha times 0.03 into 0.992 is equal to 1, you can find out alpha and you can find the exact values of each of these probabilities.



So you need not know the P(plus). In this problem you will see that P(not tumor by plus) is much higher and the reason to why this is the case is because the prior P(not tumor) is much higher.

Prior P(not tumor) is equal to 0.992 and P(tumor) is only 0.002. If these two prior probabilities were equal then P(plus by tumor) is equal to 0.098 and P(plus by not tumor) is equal to 0.03 so definitely this would have been more likely. The maximum likelihood hypothesis is that the patient has tumor. But if we combine it with the prior probabilities on the hypothesis whether a patient has tumor or not tumor we see that most likely the patient does not have tumor. So this example illustrates to you the role that prior probability plays in finding out the posterior probability.

(Refer Slide Time: 28:25)

Indian Institute of Technology, Kharagpur

## MAP Learner

For each hypothesis  $h$  in  $H$ ,  
calculate the posterior probability

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

Output the hypothesis  $h_{\text{map}}$  with  
the highest posterior probability

$$h_{\text{map}} = \max_{h \in H} P(h|D)$$

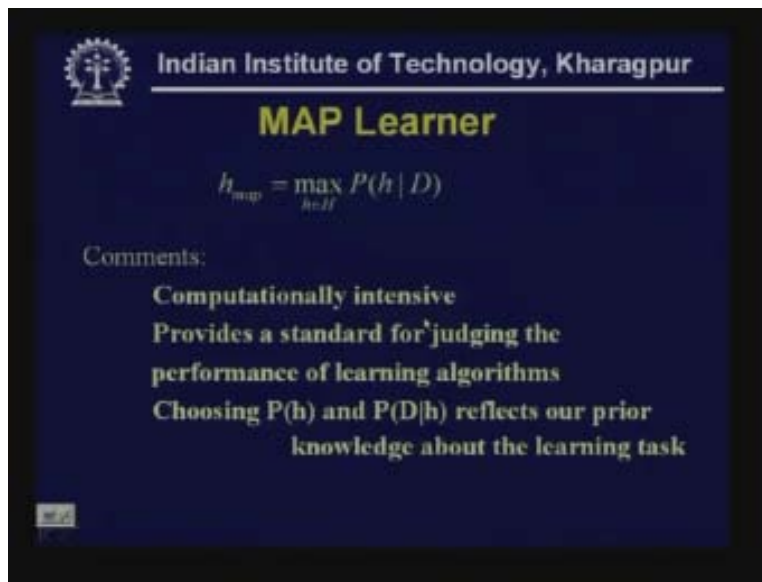
How does a map learner work?

You have the hypothesis space  $H$  and you are trying to find out for each hypothesis  $h$  in the hypothesis space you compute the posterior probability  $P(h$  by evidence) and you compute it using Bayes Rule and then you output that hypothesis for which this is highest. So the posterior probability  $h_{\text{map}}$  is maximum and that hypothesis for which  $P(h$  by  $d$ ) is maximum. This is called a map learner maximum a-posteriori probability learner. The Bayesian framework provides a survey of evaluating the probability of hypothesis.

In practice the Bayesian method is not always possible to apply. The main reason is that in the particular case we considered we had only two possible hypothesis in the hypothesis space, the patient has tumor or the patient does not have tumor. Consider a case whether a number of possible hypothesis is very large. So, if you follow this procedure for each hypothesis you have to find out the posterior probability for each of the hypothesis given to it for each of the hypothesis. You have to do it for each of the hypothesis.

If the number of hypothesis is very large it is not practical to enumerate each of the probabilities. Therefore this sort of rule is not applicable. But this nevertheless gives you a very good framework to show what is happening here. And if you study further into machine learning you will see that many of the learning algorithms are evaluated based on whether the hypothesis that they output is the maximum a-posterior hypothesis, whether it is a map hypothesis or ML hypothesis or whatever criteria that satisfies. Therefore this gives us standard for judging the efficacy of different learning algorithms and it can be applied when your hypothesis space is not very large.

(Refer Slide Time: 28:25)



Indian Institute of Technology, Kharagpur

## MAP Learner

$$h_{map} = \max_{h \in H} P(h | D)$$

Comments:

- Computationally intensive
- Provides a standard for judging the performance of learning algorithms
- Choosing  $P(h)$  and  $P(D|h)$  reflects our prior knowledge about the learning task

So, the map learner is computationally intensive because you got to find out the posterior probability for each of the hypothesis. It provides a standard for judging the performance of learning algorithm. You can analytically find out whether a particular learning algorithm gives the map hypothesis or not. And, by choosing  $P(h)$  the prior reflects the prior knowledge about the learning task. And  $P(D|h)$  also can be computed in many cases depending on our knowledge about how the hypothesis affects the data. **So far we have looked at a map hypothesis.**

A map hypothesis gives you the hypothesis which is most likely. In some cases it is not the hypothesis that you really care about but you want to really find out the class. So you have a concept learning problem, you are given some data, you are by a new example and you want to find out the class of the data. If the hypothesis corresponds to the class then a map hypothesis is what you are looking for. However, consider these cases, if the hypothesis that is returned by the map learner is applied on the instance  $x$  does it give you the most likely classification of the instance  $x$ ?

(Refer Slide Time: 32:38)

Indian Institute of Technology, Kharagpur

## Bayes Optimal Classifier

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

$H_{\text{map}}(x)$  ? Not necessarily!

Ex: Let  $P(h_1|D) = .4$ ,  $P(h_2|D) = .3$ ,  $P(h_3|D) = .3$   
Given  $x$ , we have  $h_1(x)=+$ ,  $h_2(x) = -$ ,  $h_3(x) = -$   
What is the most probable classification of  $x$  ?

It may not be the case. It depends on what form the hypothesis is in. For example, suppose you have some evidence  $D$  and you have three competing hypothesis. So your hypothesis space consists of three hypothesis  $h_1$ ,  $h_2$  and  $h_3$ . And suppose you compute the posterior  $P(h_1)$  to be 0.4 and that of  $h_2$  to be 0.3 and that of  $h_3$  to be 0.3. Now, which is the hypothesis with the highest posterior probability? Definitely it is  $h_1$  because  $P(h_1|D)$  is equal to 0.4 which is more than  $P(h_2|D)$  or  $P(h_3|D)$ .

Now we pick  $h_1$  and we apply  $h_1(x)$ . We find out that when we apply  $h_1(x)$   $h_1$  applied to  $x$  is plus whereas  $h_2$   $x$  is minus  $h_3$   $x$  is also minus. Now what is the most likely classification of  $x$ ? So you can just look at the examples and see that combined probability  $h_2$   $h_3$  is 0.36 and they predict that  $x$  is minus whereas  $h_1$  which has the single highest probability predicts that  $x$  is plus but this probability is only 0.4. So in this example  $x$  is most likely best classified as minus. Therefore the map hypothesis applied to the instance does not necessarily give you the most probable classification of  $x$ . In such cases what we really want is to find out a Bayes optimal classifier. **Now let us see what a Bayes Optimal Classifier is.**

Suppose you have different classes  $V$  so  $V$  is the set of possible plus. In your classification problem  $V$  is the set of possible classes. Therefore possible classes could be positive or negative in the case of the tumor example. Suppose you are trying to classify a news article as sports, politics, health and entertainment so your possible classes are sports, politics, health and entertainment. These are the four possible classes. So what you are trying to find is the class. So the class  $v_j$  which is an element of  $V$ ,  $V$  is the set of all classes and you want to find out that class in  $V$  for which this quantity for which this is the highest. So you have different hypothesis  $h_1$ ,  $h_2$  etc and all these hypothesis belong to the hypothesis space  $H$ . So, for each hypothesis  $h_i$  in the hypothesis space you find out the probability that the probability of a particular value of  $v_j$  sigma over all hypothesis

$P(v_j \text{ by } h_i)$  into  $P(h_i \text{ by } D)$  so this is the posterior probability of the hypothesis  $h_i$  and this is the probability that the classification is  $v_j$  given the hypothesis  $i$ .

(Refer Slide Time: 36:58)

Indian Institute of Technology, Kharagpur

### Bayes Optimal Classifier

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

Example:

$P(h_1|D) = 0.4, P(-h_1) = 0, P(+h_1) = 1$   
 $P(h_2|D) = 0.3, P(-h_2) = 1, P(+h_2) = 0$   
 $P(h_3|D) = 0.3, P(-h_3) = 1, P(+h_3) = 0$

$$\sum_{h_1} P(+h)P(h|D) = 0.4$$

$$\sum_{h_2,3} P(-h)P(h|D) = 0.6$$

Suppose you have three hypothesis  $h_1, h_2, h_3$  the posterior  $P(h_1)$  is equal to 0.4,  $h_2$  is equal to 0.3,  $h_3$  is equal to 0.3. Then what is the probability that your class is positive? It is  $P(\text{plus by } h_1)$  into 0.4 plus  $P(\text{plus by } h_2)$  into 0.3 plus  $P(\text{plus by } h_3)$  into 0.3. So this is the  $P(\text{plus})$ .

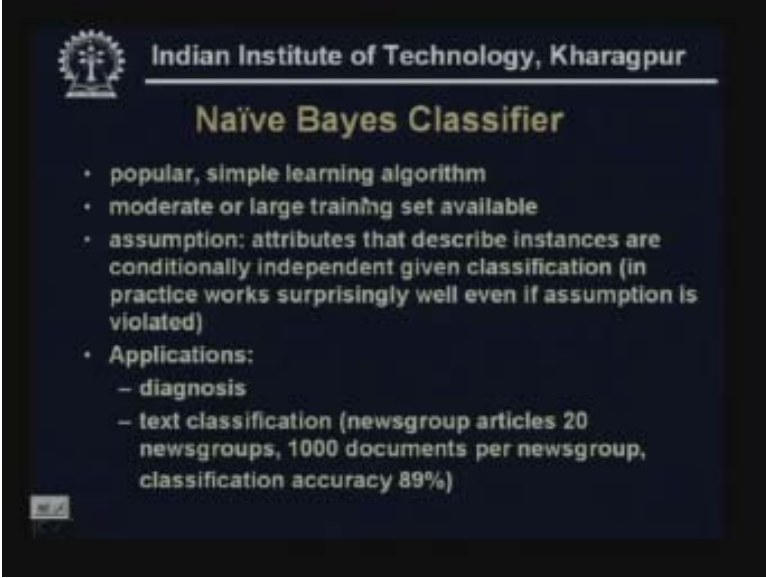
What is the  $P(\text{minus})$ ?

Similarly, it is  $P(\text{minus by } h_1)$  into posterior  $P(h_1)$  plus  $P(\text{minus by } h_2)$  into posterior of  $(h_2 \text{ minus by } h_3)$  into posterior  $(h_3)$ . So let us try to apply this to the particular problem we saw in the previous slide. We have already seen that  $P(h_1 \text{ by } D)$  is equal to 0.4,  $P(h_2 \text{ by } D)$  is equal to 0.3,  $P(h_3 \text{ by } D)$  0.3 and  $h_1(x)$  is positive that means  $P(\text{plus by } h_1)$  is equal to 1,  $P(\text{minus by } h_1)$  is equal to 0. So this is given to us and we can therefore write as  $P(h_2 \text{ by } D)$  is equal to 0.2,  $P(\text{minus by } h_2)$  is equal to 1,  $P(\text{plus by } h_2)$  is equal to 0,  $P(\text{minus by } h_1)$  is equal to 0,  $P(\text{plus by } h_1)$  is equal to 1,  $h_3$  classifies instance as negative therefore  $P(\text{minus by } h_3)$  is equal to 1 and  $P(\text{plus by } h_3)$  is equal to 0. So we can find out that  $P(\text{plus by } h_i)$  into  $P(h_i \text{ by } d)$  over all  $h_i$ 's gives us 0.4. It is because there are three hypothesis  $h_1, h_2, h_3$  and only  $h_1$  classifies the positive. So this expression of  $P(\text{plus by } h_1)$  into  $P(h_1 \text{ by } d \text{ plus by } h_1)$  is equal to 1 and  $P(h_1 \text{ by } D)$  is equal to 0.4 and the rest of the things are 0. So the  $P(\text{plus})$  is finally 0.4 whereas  $P(\text{minus})$  is equal to 0.6. So this is the most likely classification therefore the most likely classification of instance is negative.

Therefore, once you have found out the posterior probability of hypothesis you can apply the Bayes Optimal Classifier to find out the most likely classification of examples.

Naive Bayes Classifier: This is a very simple classification algorithm but quite effective in especially problems where the number of attributes is very large.

(Refer Slide Time: 40:16)



Indian Institute of Technology, Kharagpur

## Naïve Bayes Classifier

- popular, simple learning algorithm
- moderate or large training set available
- assumption: attributes that describe instances are conditionally independent given classification (in practice works surprisingly well even if assumption is violated)
- Applications:
  - diagnosis
  - text classification (newsgroup articles 20 newsgroups, 1000 documents per newsgroup, classification accuracy 89%)

So the Naive Bayes Classifier is a popular simple learning algorithm. It works well when a moderate or a large training set is available. To explain the Naive Bayes Classifier we make a very restriction assumption and the assumption is that the attributes that describe instances are conditionally independent given classification. Suppose we have some attributes  $a_1, a_2, \dots, a_n$  for a classification problem we say that the  $P(a_1 \text{ by } c)$  is independent of  $a_2$  or of the other attributes. So the attributes are independent given the class. So this is an assumption that we make. This is a very restricting assumption. We are trying to say that this attribute features are completely independent features and they do not interact. In practice, in realistic learning problems it is very difficult to find examples of attributes whether all these attributes are independent.

However, even with this wrong assumption which does not hold certain realistic problems the output of the Naive Bayes Algorithm in practical problem sometimes is surprisingly good. And there are some applications on which Naive Bayes has been applied and gives quite good results. For example, in certain diagnosis problems. And in text classification your objective is that you are given some text documents and you have some classes. For example, it could politics, sports, health, entertainment etc so you want to find out which class this article belongs to. And for this classification task Naive Bayes Algorithm has been quite successful in many such classification tasks. For example, there is a particular experiment people have performed on the 20,000 newsgroup data set where you have 20 newsgroups and you have 1000 documents per newsgroup. And on this Naive Bayes was applied and it gave a classification accuracy of 89% which is quite good. Now let us look at what the Naïve Bayes Classifier does.

We assume that we have a discrete target function which takes you from the attribute space  $X$  to the class  $C$ . And suppose each instance  $x$  is described by the attributes  $a_1, a_2, a_n$  so there are  $n$  attributes  $s_{sub 1}, s_{sub 2}, s_{sub n}$  and you have a classification from.....

(Refer Slide Time: 43:20)

Indian Institute of Technology, Kharagpur

### Naïve Bayes Classifier

Assume discrete target function  $f: X \rightarrow C$ , where each instance  $x$  described by attributes  $\langle a_1, a_2, \dots, a_n \rangle$ .

Most probable value of  $f(x)$  is:

$$C_{MAP} = \operatorname{argmax}_{c_j \in C} P(c_j | \langle a_1, a_2, \dots, a_n \rangle)$$

$$= \operatorname{argmax}_{c_j \in C} P(\langle a_1, a_2, \dots, a_n \rangle | c_j) \cdot \frac{P(c_j)}{P(\langle a_1, a_2, \dots, a_n \rangle)}$$

$$= \operatorname{argmax}_{c_j \in C} P(\langle a_1, a_2, \dots, a_n \rangle | c_j) P(c_j)$$

The  $C$  is the set of classes and  $x$  is the input instance. Now what is the most probable value of  $f(x)$ ? It is given by the  $C_{MAP}$ ,  $C_{MAP}$  is the class which has the highest probability. The class  $C$  with the maximum a-posterior probability is that class  $c_j$  for which the following maximized.  $P(c_j)$  given the attributes. So you want to find out  $C_{MAP}$  which is that value of  $c_j$  such that  $c_j$  is included in  $C$  for which the  $P(c_j$  by  $a_1, a_2, a_n)$  is maximized. We can rewrite  $P(c_j$  by  $a_1, a_2, a_n)$  as  $P(a_1, a_2, a_n$  by  $c_j)$  into  $P(c_j$  by  $P(a_1, a_2, a_n)$  by applying Bayes Rule. So  $P(c_j$  by  $a_1, a_2, a_n)$  is  $P(a_1, a_2, a_n$  by  $c_j)$  times  $P(c_j$  by  $P(a_1, a_2, a_n)$ . And since  $P(a_1, a_2, a_n)$  does not depend on  $c_j$  we can say that  $C_{MAP}$  is that class for which  $P(a_1, a_2, a_n$  by  $c_j)$  times  $P(c_j)$  is maximum. So  $C_{MAP}$  is that class  $c_j$  for which  $P(a_1, a_2, a_n$  by  $c_j)$  times  $P(c_j)$  is maximum.

The  $P(c_j)$  we can estimate from either we can have prior belief or we can estimate from the data that we have given. If you are given some training examples and the training examples we find out the proportion of examples belonging to the different classes we can estimate  $P(c_j)$ . Now the difficulty is to find out value of this expression:  $P(a_1, a_2, a_n$  by  $c_j)$ . This is a joint  $P(n)$  attributes. Now if we want to estimate the joint probability from the training data we will require a very large number of training data. It is because we have  $n$  attributes. Even in the simplest case when each attribute is Boolean that is each attribute can take two different values the number of combination of  $a_1, a_2, a_n$  is  $2$  power  $n$ . So the number of such joints is very large.



(Refer Slide Time: 43:20)

Indian Institute of Technology, Kharagpur

### Naïve Bayes Classifier

$$c_{MAP} = \arg \max_{c_j \in C} P(a_1, a_2, \dots, a_n | c_j) P(c_j)$$

Naïve Bayes assumption:

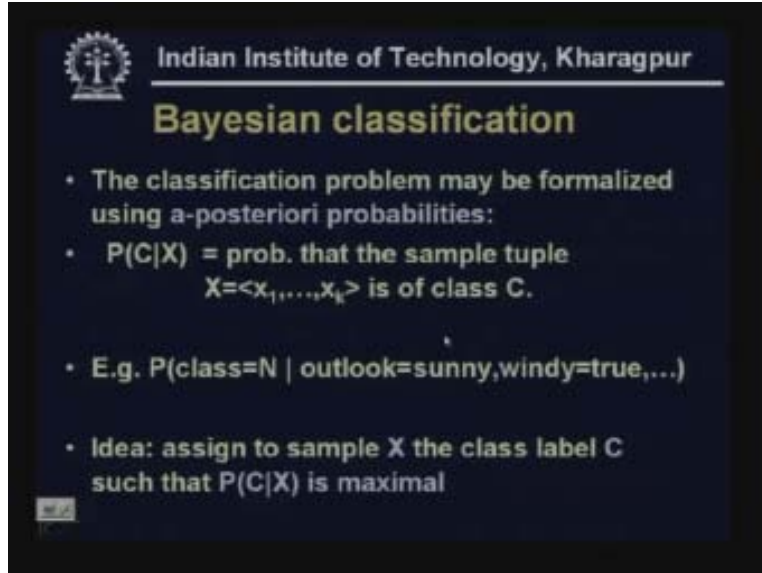
$$P(a_1, a_2, \dots, a_n | c_j) P(c_j) = \prod_i P(a_i | c_j) P(c_j)$$
$$c_{NB} = \arg \max_{c_j \in C} \prod_i P(a_i | c_j) P(c_j)$$

We look at all combinations of the attribute values; the number of such combinations is very large. Now in order to estimate for each of them the probability requires very large **corpus** and also requires a large number of the values that you have to store and this is clearly not doable when  $n$  is large. Therefore in the Naïve Bayes assumption we make a very simplifying assumption.

We assume that  $P(a_1, a_2, a_n \text{ by } c_j)$  is the product for  $P(a_i \text{ by } c_j)$ . We assume that  $(a_1, a_2, a_n)$  are conditionally independent given  $c_j$  so this can be written as  $P(a_1 \text{ by } c_j)$  times  $P(a_2 \text{ by } c_j)$  times  $P(\dots a_n \text{ by } c_j)$ . So we can write  $P(a_1 a_2 a_n \text{ by } c_j)$  as product of  $P(a_i \text{ by } c_j)$  over all  $i$ 's. Therefore  $C_{NB}$  is that class for which this times  $P(c_j)$  is maximized. So  $C_{NB}$  is the class  $c_j$  which is a member of  $C$  so that product over  $i$   $P(a_i \text{ by } c_j)$  into  $P(c_j)$  is maximized. This is the Naives Bayes Classifier, it works under the assumption that the attributes are conditionally independent of each other given that we know the class. A recap: In Bayesian classification the classification problem may be formalized using a-posterior probabilities.



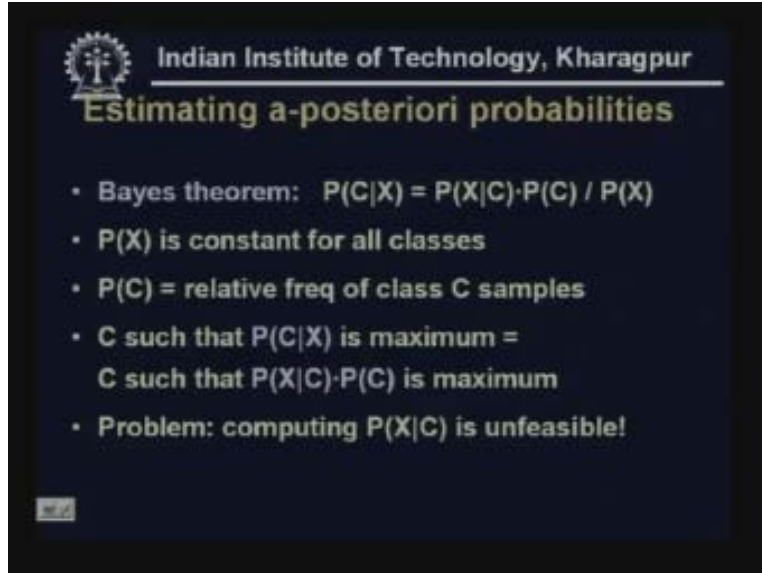
(Refer Slide Time: 48:20)



The slide features the IIT Kharagpur logo in the top left corner. The title 'Bayesian classification' is centered at the top in a yellow font. Below the title, there are four bullet points in white text on a dark blue background. The first bullet point states that the classification problem can be formalized using a-posteriori probabilities. The second bullet point defines  $P(C|X)$  as the probability that a sample tuple  $X = \langle x_1, \dots, x_k \rangle$  belongs to class C. The third bullet point provides an example:  $P(\text{class}=N \mid \text{outlook}=\text{sunny}, \text{windy}=\text{true}, \dots)$ . The fourth bullet point states the goal: to assign a class label C to a sample X such that  $P(C|X)$  is maximized.

We want to find out  $P(C \text{ by } A)$  probability that the sample tuple is of class C. You are given an instance X and you want to find out the probability that instance belongs to class. For example, suppose you want to know given the attributes of a particular day whether it is a good day for playing. Suppose you are given that the outlook of the day is sunny, windy and so on I want to know whether it is a good day to play or bad day to play. So there are two classes yes and no. Therefore the idea is you want to assign to the sample X the class label c such that  $P(C \text{ by } X_0)$  is maximum. This is the problem we were trying to solve and Naives Bayes Classifier is one way of getting the solution provided the independent assumptions hold provided these attributes are independent given the class. And outlook and windy does not depend on each other so whether it is sunny or cloudy does not affect whether it is wind or not which may not be true.

(Refer Slide Time: 50:11)



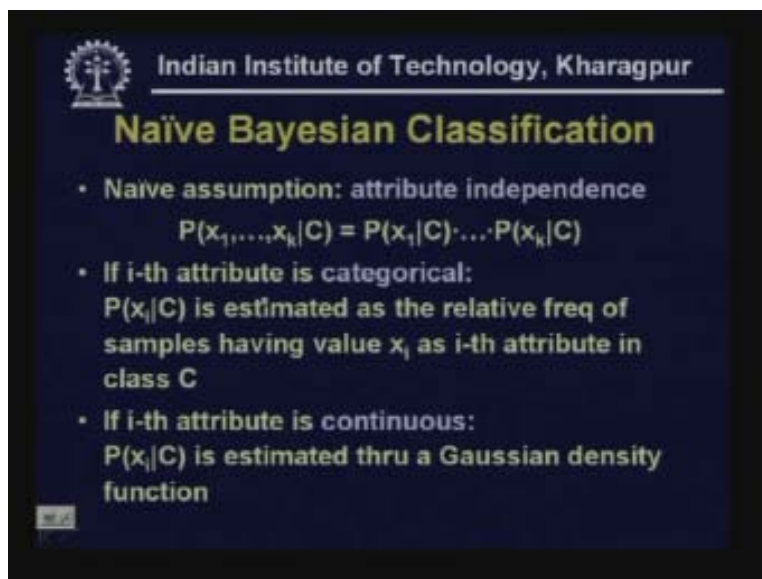
Indian Institute of Technology, Kharagpur

### Estimating a-posteriori probabilities

- Bayes theorem:  $P(C|X) = P(X|C) \cdot P(C) / P(X)$
- $P(X)$  is constant for all classes
- $P(C) =$  relative freq of class C samples
- C such that  $P(C|X)$  is maximum =  
C such that  $P(X|C) \cdot P(C)$  is maximum
- Problem: computing  $P(X|C)$  is unfeasible!

We have seen how to estimate a-posterior probabilities by using Bayes Theorem. Bayes Theorem is stated as  $P(C \text{ by } X)$  is equal to  $P(X \text{ by } C)$  times  $P(C)$  by  $P(X)$ ;  $P(X)$  is constant for all classes and therefore we want to find out that class for which  $P(X \text{ by } C)$  into  $P(C)$  is maximum. But we cannot compute  $P(X \text{ by } C)$  because  $X$  is an instance which depends on potentially a large number of attributes and if you to estimate each of these possibilities the number of such possibility will be very large. So in Naïve Bayes Classification we have made a very restrictive assumption a naïve assumption which says attributes are independent. That is  $P(x_1, x_2, x_k \text{ by } C)$  is product of  $P(x_1 \text{ by } C)$   $(x_2 \text{ by } C)$   $(x_k \text{ by } C)$ .

(Refer Slide Time: 51:12)



Indian Institute of Technology, Kharagpur

### Naïve Bayesian Classification

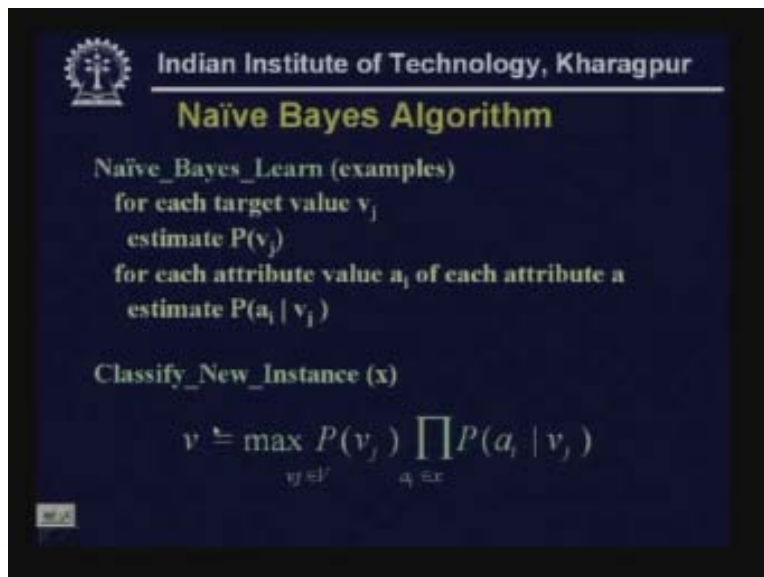
- Naive assumption: attribute independence  
 $P(x_1, \dots, x_k | C) = P(x_1 | C) \cdot \dots \cdot P(x_k | C)$
- If  $i$ -th attribute is categorical:  
 $P(x_i | C)$  is estimated as the relative freq of samples having value  $x_i$  as  $i$ -th attribute in class C
- If  $i$ -th attribute is continuous:  
 $P(x_i | C)$  is estimated thru a Gaussian density function

Suppose  $x_i$  is a categorical attribute then how do I estimate  $P(x_i \text{ by } c)$ ? What we can do is we can look at the training examples and find out for a given class how many times  $x_i$  has a particular value. So  $P(x_i \text{ by } c)$  can be estimated from the training corpus as the relative frequency of those training examples that have value  $x_i$  as the  $i^{\text{th}}$  attribute in class  $c$ . But if your attribute is continuous, suppose you have an attribute like temperature which takes continuous values then  $(e_{x_i} \text{ by } c)$  is often estimated by assuming a probability distribution for these attributes. For example, often we assume a Gaussian density function from which  $P(x_i \text{ by } c)$  can be estimated.

A recap: Naïve Bayes Algorithm works like this. In Naives Bayes learn you are given some examples. For each target value  $v_j$  you estimate  $P(v_j)$ . Suppose you have this news group classification or news classification you have four classes sports, politics heath and entertainment. Now, for each of these classes you estimate  $P(v_j)$  by looking at the corpus and finding out how many news you have from sports, how many from politics, how many from health, and how many from entertainment. So, from this you can estimate  $P(v_j)$ . Now for each attribute value you can estimate  $P(a_i \text{ by } v_j)$ .

So  $v_j$  is sports, you find out how many times this attribute  $a_1$  is true or  $a_1$  has a particular value. In a document classification we often use the words in the document as features. So we look at the set of words that occur in the document. So we want to know how many times the word referee is in a document given that it belongs to a sports domain or how many times the word referee is in the document that belongs to entertainment domain or the health domain.

(Refer Slide Time: 54:57)



Indian Institute of Technology, Kharagpur

### Naïve Bayes Algorithm

Naïve\_Bayes\_Learn (examples)  
for each target value  $v_j$   
estimate  $P(v_j)$   
for each attribute value  $a_i$  of each attribute  $a$   
estimate  $P(a_i | v_j)$

Classify\_New\_Instance (x)

$$v \simeq \max_{v_j \in V} P(v_j) \prod_{a_i \in x} P(a_i | v_j)$$

For each attribute or each word that we are considering we can find out how many times that attribute takes a particular value for those documents that belong to that class. From this we can estimate  $P(a_i \text{ by } v_j)$ . Once we have estimated for all  $j$   $P(v_j)$  that is for all class the probabilities and all conditional  $P(a_i \text{ by } v_j)$  we can use those values to classify a new

instance. So the most likely classification of the new instance given by  $v$  is that value of  $v_j$  for which  $P(v_j)$  into product of  $(p_i \text{ by } v_j)$  for all  $A_i$  is maximum. This is Naive Bayes formula for classifying a new instance.

So how do I compute the  $P(a_i \text{ by } v_j)$ ?

$P(a_i \text{ by } v_j)$  is, suppose  $n$  denotes the number of examples in your training set for which the class is  $v_j$  and  $n_c$  denotes the number of examples in your training set for which the class is  $v_j$  and for which  $A_i$  has this particular value that is the attribute has the value of  $i$ . So  $P(a_i \text{ by } v_j)$  can be estimated  $n_c$  by  $n$ . However, this sort of estimation has a problem. It could be that because you do not have sufficient number of examples in your corpus in your training data there is no example with class is equal to  $v_j$  in attribute is equal to  $A_i$  in which case this probability will be 0. And because of you are taking the product of different probabilities if one of the probabilities is 0 the entire value would be 0. Therefore in order to avoid this problem we do some form of smoothing. So, to ensure that the numerator is not 0 we add  $p$  which is the prior estimate for  $P(a_i \text{ by } v_j)$ .

(Refer Slide Time: 56:57)

Indian Institute of Technology, Kharagpur

### Naïve Bayes Algorithm

$$v = \max_{v \in V} P(v_j) \prod_{a_i \in A} P(a_i | v_j)$$

Typical estimation of  $P(a_i | v_j)$

$$P(a_i | v_j) \leftarrow \frac{n_c + mp}{n + m}$$

where  
 $n$ : examples with  $v=v_j$ ;  $p$  is prior estimate for  $P(a_i|v_j)$   
 $n_c$ : examples with  $a=a_i$ ,  $m$  is the weight to prior

So we ensure the prior estimate is greater than 0. And if  $m$  is the weight given to the prior estimate then we add  $mp$  to the numerator and  $m$  to the denominator. So this is the smoothed value of  $P(a_i \text{ by } v_j)$  to ensure that this never goes to 0. In practice we take  $p$  to be a very small value and  $m$  is the number of your examples. So let us just illustrate it with an example. Suppose we are given the following data:

(Refer Slide Time: 56:57)

Indian Institute of Technology, Kharagpur

### Naive Bayesian Classifier (II)

- Exercise: Given a training set apply the Naive Bayes classifier

Outlook	P	N	Humidity	P	N
sunny	2/9	3/5	high	3/9	4/5
overcast	4/9	0	normal	6/9	1/5
rain	3/9	2/5			

Temperature	P	N	Windy	P	N
hot	2/9	2/5	true	3/9	3/5
mild	4/9	2/5	false	6/9	2/5
cool	3/9	1/5			

Suppose we have some training set and in the training set we have the data where we have the different attributes of a particular day. For example, what is the outlook of the day? Is it sunny or overcast or raining? Is the temperature hot mild or cool is the humidity high or normal is that windy or not? And we want to know whether it is a good day for sport or a bad day for sport. And from the data we find out from the  $P(a_i \text{ by } v_j)$ . If it is sunny then the  $P(\text{good day})$  is equal to 2 by 9 and bad day is 3 by 5 and so on. So we get this data of  $P(a_i \text{ by } v_j)$  and when we are given a new instance the new instance is described by the values of the attributes.

(Refer Slide Time: 58:11)

Indian Institute of Technology, Kharagpur

### Naive Bayes Example

Consider *PlayTennis* and new instance  
< Outlook=Sunny, Temp=cool,  
Humidity=high, Wind=strong >

Compute  $c_{NB} = \text{argmax}_{c_j \in C} P(c_j) \prod_{a_i \in X} P(a_i|c_j)$   
playtennis (9+,5-)  
 $P(\text{yes}) = 9/14$ ,  $P(\text{no}) = 5/14$

In this particular instance outlook is sunny, temperature is cool, humidity is high and wind is strong and we want to find out whether it is a good day for playing tennis or a bad day. I have fourteen examples in my training set and nine of them are good days five are bad days so  $P(\text{yes})$  is equal to 9 by 14  $P(\text{no})$  is equal to 5 by 14 and then we use the value of the previous table to find out the Naïve Bayes example. The independence hypothesis we make in Naive Bayes makes computation simple but is seldom satisfied in practice.

(Refer Slide Time: 59:11)



Indian Institute of Technology, Kharagpur

### The independence hypothesis

- makes computation simple
- but is seldom satisfied in practice, as attributes (variables) are often correlated.
- Attempts to overcome this limitation:
  - Bayesian networks - combine Bayesian reasoning with causal relationships between attributes

And if we realistically want to model the problem we have to use joint probabilities which are not feasible. But the way out is to use Bayesian networks. You can have algorithms for learning Bayesian networks.