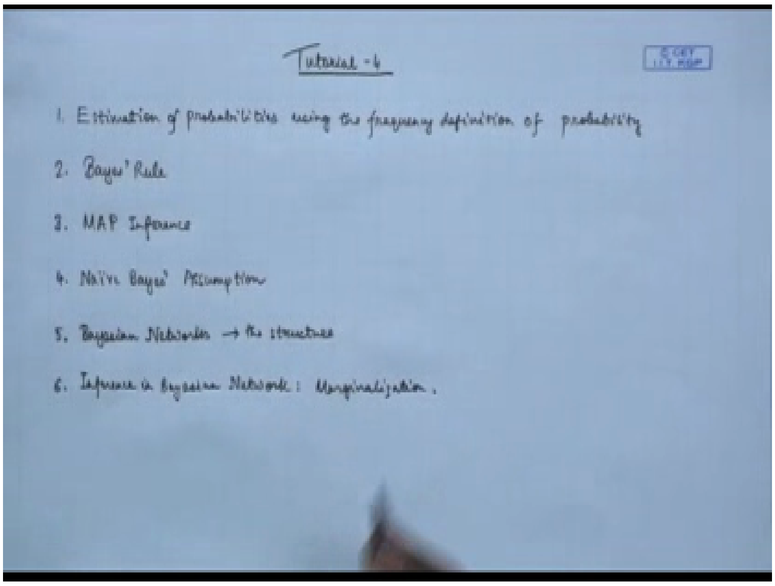
**Introduction to Machine Learning**
**Prof. Mr. Anirban Santara**
**Department of Computer Science and Engineering**
**Indian Institute of Technology, Kharagpur**

# Tutorial IV

Hello friends, this is Anirban, welcome to the tutorial session of the fourth week of this course. In this session, we are going to summarise everything that has been covered in the course, and learn how to solve problems in the exam. So, let us see what all topic we are going to cover today.

(Refer Slide Time: 00:32)

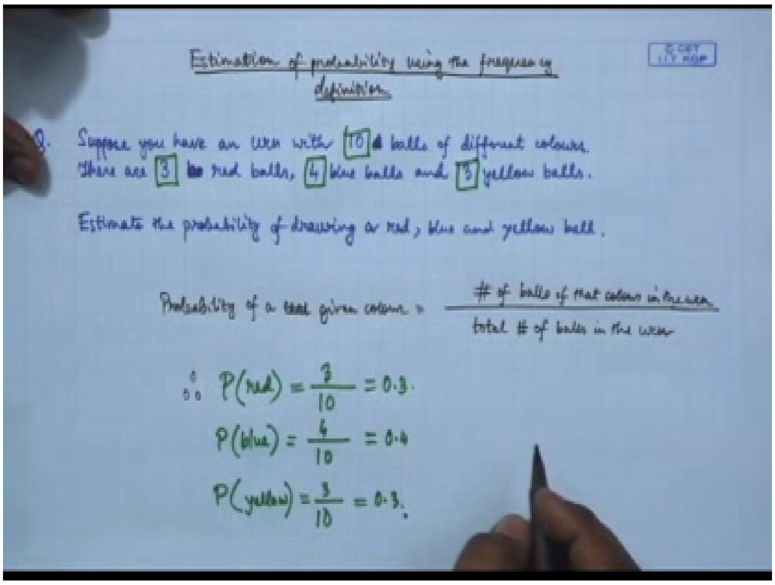


Tutorial - 4

1. Estimation of probabilities using the frequency definition of probability
2. Bayes' Rule
3. MAP Inference
4. Naive Bayes' Assumption
5. Bayesian Networks → the structures
6. Inference in Bayesian Networks : Marginalization.

Today, we are going to first cover estimation of probabilities using the frequency definition of probability. The second topic will be Bayes' rule. The third topic is maximum aposteriori probability inference. Fourth will be Naive Baye's assumption. The fifth topic will be Bayesian network, the structure of Bayesian network what is it all about. Sixth will be inference in Bayesian network and the concept of marginalization.

(Refer Slide Time: 02:04)



Estimation of probability using the frequency definition

Q. Suppose you have an urn with $\boxed{10}$ balls of different colours. There are $\boxed{3}$ red balls, $\boxed{4}$ blue balls and $\boxed{3}$ yellow balls.

Estimate the probability of drawing a red, blue and yellow ball.

$$\text{Probability of a given colour} = \frac{\text{\# of balls of that colour in the urn}}{\text{total \# of balls in the urn}}$$

$$\therefore P(\text{red}) = \frac{3}{10} = 0.3.$$

$$P(\text{blue}) = \frac{4}{10} = 0.4$$

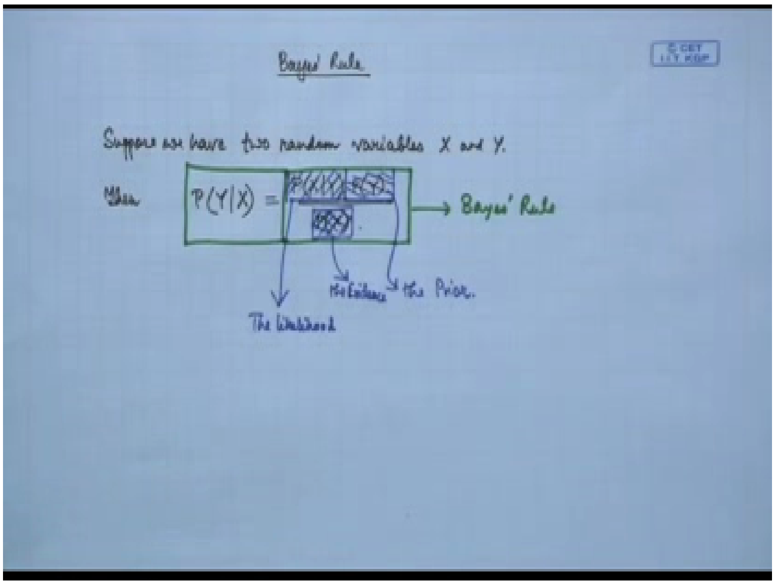
$$P(\text{yellow}) = \frac{3}{10} = 0.3.$$

So, let us take up the first topic of today, which is estimation of probabilities using the frequency definition. The frequency definition of probability says that the probability often event is the fraction of times that particular event happen, for example, we have (Refer Time: 02:34) coin, say we toss it 100 times and 45 times the heads come up, head side comes up, and the remaining 55 times the tail shows up. So, the probability of head will be 45 by 100 or 0.45, and the probability of tails will be 55 by 100 which is 0.55. And so let us take up problem which you can face in the exam and try to solve it.

Suppose you have an urn with 10 balls of different colour. There are 3 red balls, 4 blue balls and 3 yellow balls. Estimate the probability of drawing a red, blue and yellow ball. So you have been given an urn which has balls of three different colours. And now you have been asked to just close your eyes, and pick up an at random. So, what is the probability that you would pick red balls, what is the probability of a blue ball, and what is the probability of a yellow ball? So, from the frequency definition, probability of a given colour will be number of balls of that colour in the urn divided by total number of balls in the urn.

So, probability therefore, probability of red of drawing a red ball is equal to 3 times red because there are 3 red balls, so 3 divided by all together, it is 3 plus 47 plus 30 total number of balls 10 or 0.3. Probability of blue equal to there is 4 blue balls, so 4 divided by 10 is equal to 0.4. Probability of yellow will be equal to 3, since there are 3 yellow balls divided by the total number of balls which is 10, which is equal to 0.3. You add these three things up 0.3 plus 0.3 is 0.6 plus 0.4 is 1. So, this is how you calculate the probability using the frequency definition. Let us go to the next topic and which is the Bayes rule.

(Refer Slide Time: 06:23)



Bayes' Rule

Suppose we have two random variables X and Y.

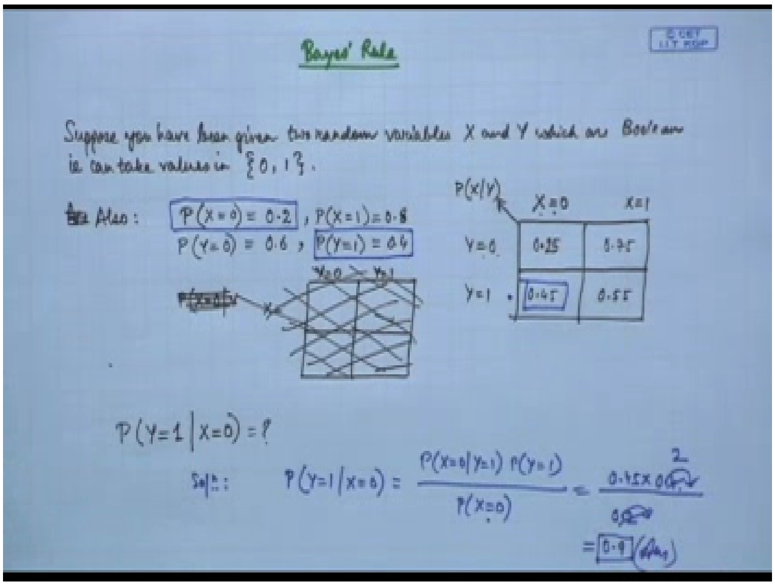
Then $P(Y|X) =$

The Bayes' rule of probability says that suppose we have two random variables X and Y, then probability of Y given X. Suppose, that you have been a given a particular value of X and then ask what the most probable value of Y should be, or what is the probability of different value of Y given a particular value of X. So, probability of Y given X will be equal to probability of X given Y into probability of Y divided by probability of X now each of this is the Bayes' rule. So, this is called the rather the entire thing it is called the Bayes' rule of probability

Now, each of this term in the Bayes rule has a name. P of X given Y this quantity is called the likelihood of X given Y. The term P of Y, it is called the prior. So, even you are asking what is the probability of a certain value of Y, you would like to know what was the prior probability of Y that is without any given information what how likely how probable that particular of value of Y is, so that is the prior probability. And the denominator it is called P of X this is called evidence. So, when you are giving A value of X you are also going to say that how common that particular value of X is so that X i can evidence so this particular term is called the evidence. So, this is the Bayes rule.

Now let us go ahead and see what kind of problems can come in the exam from this part. So, you will be given the value of different you will given this these quantities and you will asked to would these things in formula, and ask and do you know Bayes estimation maximum aposteriori estimation. So, let us talk about it next. So, given that so this is a question that can come in the exam; given that P of so let us go to a new page and see what kind can come in the exam.

(Refer Slide Time: 09:30)



Bayes' Rule

Suppose you have been given two random variables X and Y which are Boolean ie can take values in $\{0, 1\}$.

Also: $\boxed{P(X=0) = 0.2}$, $P(X=1) = 0.8$
$P(Y=0) = 0.6$, $\boxed{P(Y=1) = 0.4}$

| P(X/Y) | X=0 | X=1 |
|---|---|---|
| Y=0 | 0.25 | 0.75 |
| Y=1 | $\boxed{0.45}$ | 0.55 |

$P(Y=1 \mid X=0) = ?$

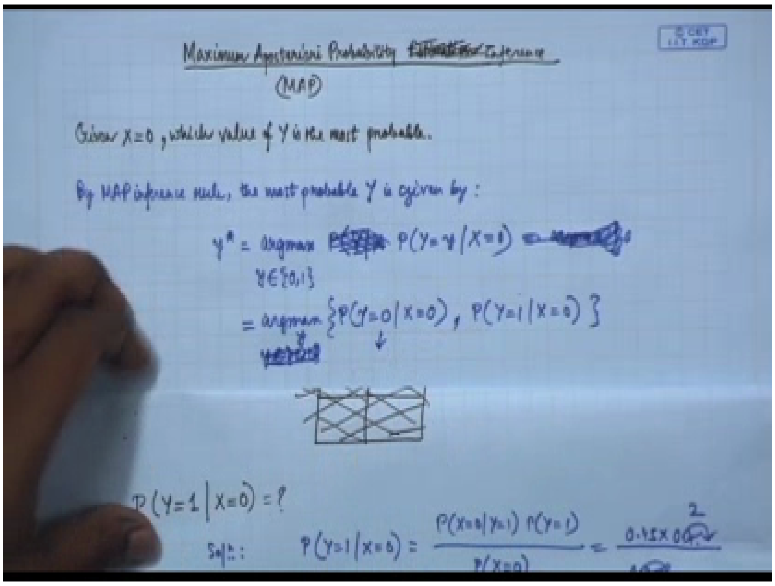
Soln: $P(Y=1/X=0) = \frac{P(X=0/Y=1)\,P(Y=1)}{P(X=0)} = \frac{0.45 \times 0.4}{0.2}$
$= \boxed{0.9}$ (Ans)

Bayes rule continued. Suppose you have been given two random variables X and Y, which are Boolean, so Boolean random variables are those which can take values either 0 or 1 that is can take values in 0, 1. Also probability of X equal to 0 is 0.2, probability X equal to 1 is 0.8; probability Y equal to 0 is 0.6, probability Y equal to 1 is 0.4. Probability of X equal to 0 given; now you are given a table of the conditional probability distributions.

So, here you have Y equal to 0, here you have Y equal to 1, X equal to the other way around yeah, say X equal to 0, and X equal to 1, you have been given this table Y equal to 0, Y equal to 1. And this quantity is probability of X given Y, so whatever you entry is coming here is going to be probability of X given Y. So, probability of X equal to 0 given Y equal to 0 is say 0.25, probability of X equal to 1 given Y equal to 0 is 0.75, so these two term should add to 1. Probability of X equal to 0 given Y equal to 1 is it 0.45 and this is 0.55. So, this is given to you. And you have been asked what should be the probability of Y equal to 1 given X equal to 0. So, what is this quantity?

So, you can directly use Bayes rule over here. How do you solve so probability solution is Y equal to 1 given X equal to 0 is equal to probability of X equal to 0 given Y equal to 1 into probability of Y equal to 1 divided by probability of X equal to 0, which is equal to probability of X equal to 0 is this 0.2. Probability of X equal to 0 given Y equal to 1 is X equal to 0 given Y equal to 1 is this quantity 0.45 times probability of Y equal to 1 probability of Y equal to 1 is 0.4. So, just shift the points this is 2 and 0.45 into 2 is 0.9, perfect. So, the probability, so this is your answer, so this is how you saw using the Bayes rule, the probability of Y equal 1 given X equal to 0 is equal to 0.9. So, you will be asked this kind of question in the exam from the Bayes rule, pretty simple. The next topic that we will take up is MAP inference.

(Refer Slide Time: 13:56)



Maximum Aposteriori Probability ~~Estimation~~ Inference

(MAP)

Given X=0, which value of Y is the most probable.

By MAP inference rule, the most probable Y is given by:

$$y^* = \underset{y \in \{0,1\}}{\mathrm{argmax}}\ P(Y=y \mid X=0)$$

$$= \underset{y}{\mathrm{argmax}} \{P(Y=0 \mid X=0),\ P(Y=1 \mid X=0)\}$$

$P(Y=1 \mid X=0) = ?$

Soln: $P(Y=1 \mid X=0) = \frac{P(X=0 \mid Y=1)\, P(Y=1)}{P(X=0)} = 0.45 \times 0.4$ 2

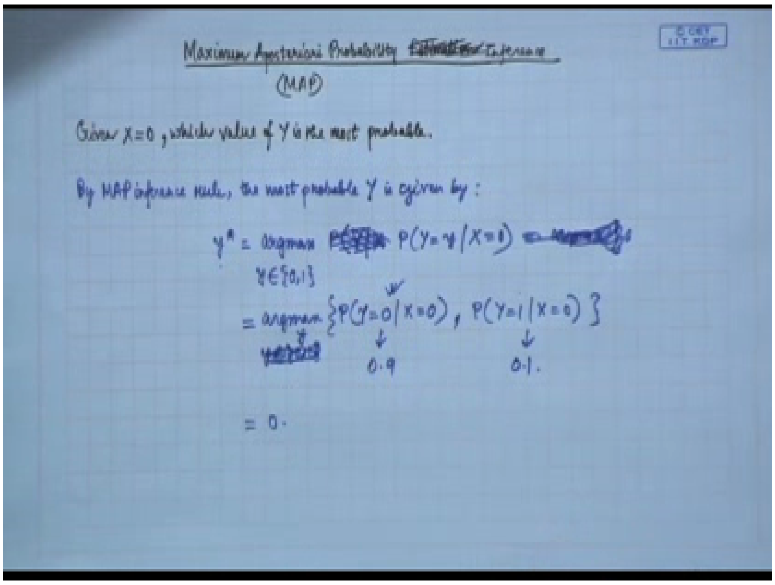
So, I will spell it out it is maximum aposteriori probability estimation. So, the maximum aposteriori probability estimation is also known as MAP inference or MAP estimation, in rather I should say inference. So, it is it goes like this, so you have been given that the same let us have the same problem statement as before.

Suppose, you have been given two random variables X and Y which are Boolean and takes values in 0 and 1. And this is the different prior probability, evidences and the conditional, the likelihood of X given Y and you have been ask to find out. So, given this condition given a particular value of X what is the most probable value of Y. So, I am going to fold this in a half, and keep it like this, let me fold it again.

So, now, you have the problem. So, this is the scenario, I hope you can see it, yes, now you should be able to see it. And this is your statement and you have been asked that given this scenario and a value of X given X equal to 0, which value of Y is the most probable. So, the MAP probability MAP estimation would go like this, it returns, so by MAP inference rule the most probable Y is given by Y star equal to A argmax over Y taking values in 0 and 1. Probability of Y given X or I should write probability of Y equal to Y given X equal to 0.

So, this you are going to choice the particular value of Y which as the maximum aposteriori probability given the value of X. So, this is the basic philosophy. And in this case, you will find as we calculate before that you were going to find out so we calculated that probability of Y equal to 1 given X equal to 0 is 0.9. So, it is going to be argmax rather yes, so it is going to be equal to argmax, Y belonging to into 0 and 1 of probability of Y equal to 0 given X equal to 0, and argmax over Y. So, I will write small y and probability Y equal to 1 given X equal to 0. So, this quantity was calculated to be 0.9, it was calculated to be 0.9 before.

(Refer Slide Time: 18:16)



Maximum Aposteriori Probability ~~Estimation~~ Inference

(MAP)

Given $X=0$, which value of $Y$ is the most probable.

By MAP inference rule, the most probable $Y$ is given by :

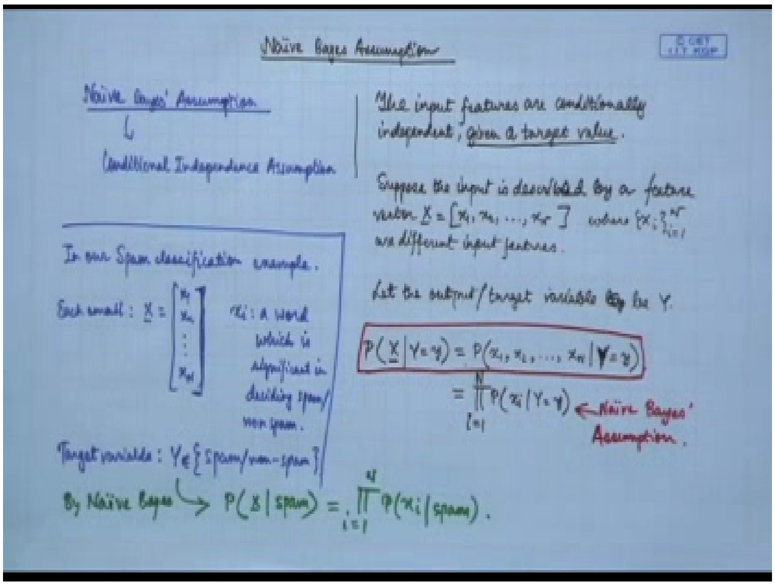
$$y^* = \underset{y \in \{0,1\}}{\text{argmax}} \; P(Y=y / X=0)$$

$$= \text{argmax} \{ \underset{0.9}{\underbrace{P(Y=0 / X=0)}}, \underset{0.1.}{\underbrace{P(Y=1 / X=0)}} \}$$

$$= 0.$$

And similarly, we can calculate this is 0.1 and so which value of Y gives more value more the maximum probability maximum aposteriori probability Y equal to 0, so this is going to be 0. So, the most probable value of Y given this particular value of X is 0. So, this is how MAP inference goes. And this is theoretically the best possible inference. So, given all the estimation given the likelihood, which comes from domain knowledge and the priors, which also come from the domain knowledge and their problem definition, the best you can predict is by this rule so you are going to find put that particular value of the output which is the most probable given the input.

(Refer Slide Time: 19:16)



Naïve Bayes Assumption

Naïve Bayes' Assumption

$\downarrow$

Conditional Independence Assumption

"The input features are conditionally independent, given a target value.

Suppose the input is described by a feature vector $\underline{X} = [x_1, x_2, \ldots, x_N]$ where $\{x_i\}_{i=1}^{N}$ are different input features.

Let the output/target variable be Y.

$$\boxed{P(\underline{X} \mid Y = y) = P(x_1, x_2, \ldots, x_N \mid Y = y)}$$
$$= \prod_{i=1}^{N} P(x_i \mid Y = y) \leftarrow \text{Naïve Bayes' Assumption.}$$

In our Spam classification example.

Each email : $\underline{X} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_N \end{bmatrix}$ $x_i$ : a word which is significant in deciding spam/non spam.

Target variable : $Y \in \{\text{spam}/\text{non-spam}\}$

By Naïve Bayes $\hookrightarrow$ $P(\underline{X} \mid \text{spam}) = \prod_{i=1}^{N} P(x_i \mid \text{spam})$.

In this context, another important topic which, comes in is the Naive Bayes assumption. So, what does the Naive Bayes assumption say the Naive Bayes assumption says that Naive Bayes assumption; this is a conditional independence assumption. So, the Naive Bayes assumption says that so this is the statement of Naive Bayes the input features are conditionally independent given a target value, so this is important. So, if you say that I am sure that the target is going to take a particular value say 0 given that the target has been prefixed to a given value the con the all the input parameters the input variables or the input features, it becomes conditionally independent. So, this is the Naive Bayes assumption.

So, how does it look like in the mathematics side, so this says that suppose the input is described by a future vector X, which is a vector x 1, x 2 till x n, and where x i is i equal to 1 through n are different input features. So, let the output or target variable be Y.

Then probability of X this vector given a particular value of Y, it is going to be so all the variable, so what is this, so this is basically a co-occurrence of all of these random variables, for A joint probability distributions of all of these random variables. So, this can be approximated as a product of the conditioner probability distributions of each of these features given the value of Y. So, i equal to 1 through n p of x i given Y equal to y. So, you assumed that all of these input features become conditionally independent when a particular value of the target variable is given, so this is the Naive Bayes assumption. It is very helpful in certain scenarios, but often is not right to assume Naive Bayes assumption.

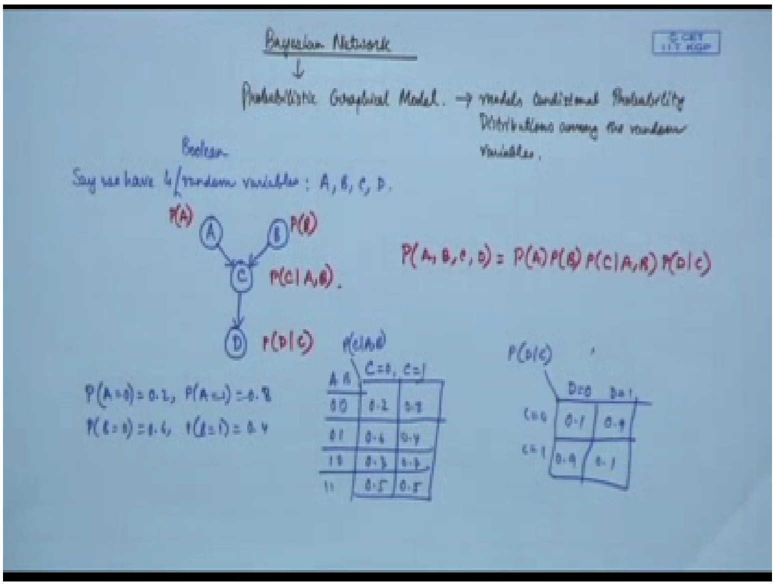
Yes, so what kind of problems can come from this section, you will be given a certain scenario. So, maybe you have already seen the hands on exercise python exercise, and there we used Naive Bayes assumption to classify emails as spam or non spam. There we had assumed at given Y equal to say now let us take up that problem again. So, what kind of assumption was done there in that case so in our so if you have miss this video then go ahead and look at assignment or like exercise, hands on exercise number 3 or week 4 hands on exercise as will be easier for you to find out in the course website. So, we are going to take up that example again and we consider it for a moment.

In our spam classification example, each email was described as a vector of X's where sorry where each x i was a word which is significant in deciding spam or non spam so given that and given this kind of input descriptions and the target value was Y which can takes values either in spam or non spam. So, the Naive Bayes assumption says that would like following the Naive Bayes assumption we wrote that by Naive Bayes we wrote that probability of X given spam. So given that the mail is spam, the value of the output is known, so this is going to be the product over all the different words of probability of that particular word occurring in spam email.

So, this is quite straightforward, and it is helpful in many machinery scenarios. And particularly, when you do not want to have too many parameters, because you do not have too much data, and if you are not going to if you do not assume the Naive Bayes assumption, if you do not consider the Naive Bayes assumption then they might be too many parameters in modelling or describing the joint distributions or the features. So in those scenarios, the Naive Bayes assumption becomes quite handy.

In the exam you will be given certain values of the input and certain value of the output, and you will ask to calculate joint distributions. So, you basically we ask to calculate this using Naive Bayes assumption or maybe this particular this is just the likelihood part. So, this particular likelihood part will as to estimate using the Naive Bayes assumption and you maybe have to plug it into the Bayes rule you calculate the aposteriori distributions and from there. Once you have the aposteriori distributions ready at your hand you can make A MAP inference about the most probable value of Y for a given value of the input variable X like we did in the spam classifier.

(Refer Slide Time: 27:36)



Bayesian Network

↓

Probabilistic Graphical Model. → models conditional Probability Distributions among the random variables.

Say we have 4 Boolean random variables : A, B, C, D.

$P(A)$ $P(B)$ $P(C|A,B)$. $P(D|C)$

$P(A,B,C,D) = P(A)P(B)P(C|A,B)P(D|C)$

$P(A=0) = 0.2$, $P(A=1) = 0.8$
$P(B=0) = 0.6$, $P(B=1) = 0.4$

$P(C|A,B)$

| A B | C=0, | C=1 |
|---|---|---|
| 00 | 0.2 | 0.8 |
| 01 | 0.6 | 0.4 |
| 10 | 0.3 | 0.7 |
| 11 | 0.5 | 0.5 |

$P(D|C)$

| | D=0 | D=1 |
|---|---|---|
| C=0 | 0.1 | 0.9 |
| C=1 | 0.9 | 0.1 |

So, let us move on to the next topic which is Bayesian networks. A Bayesian network is probabilistic graphical model, and it is a probabilistic graphical model, and it models conditional probability distributions among the random variable. So, how the Bayesian look like say you have four random variables, again four Boolean random variables makes life simpler A, B, C and D. And a Bayes net would look something like this.

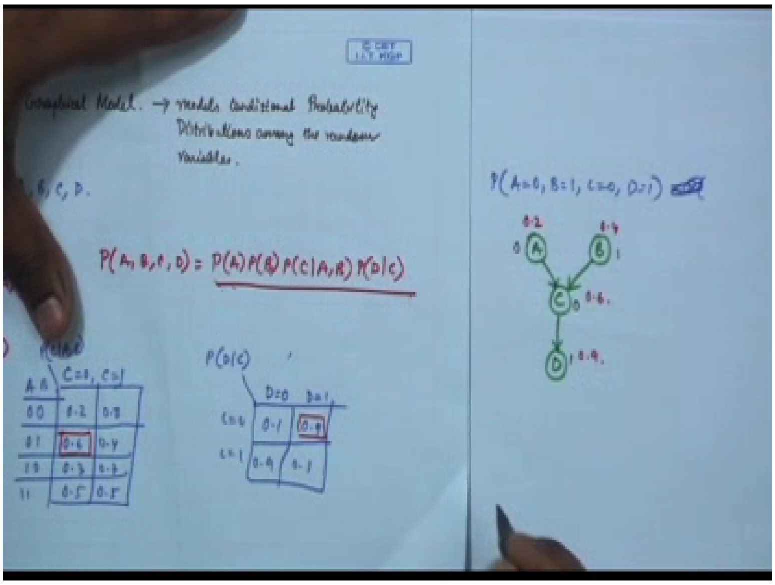
This is a Bayes net. And each of these head nodes the once that come before that come in the beginning, they have these prior probability distributions, so you have the prior probability distributions of A and B. And all the remaining nodes will have conditional distributions of those values of their values given the values of the parents - the immediate parents. So for C, you will have probability of C given A and B, and for D you are going to have probability of D given C.

Now, the Bayesian networks this theory says that if the variables can be represented in the form of a Bayesian networks of in this form then the joint probability distributions of the random variables can be written as a product of the priors and the conditionals. So, this gives a very elegant way of modelling the joint probability distributions. And also sometimes reach to the use of much smaller number of parameters and hence chances an over fitting are also reduced. So, Bayes nets are extremely helpful in making full models of variables which are highly interlinked among each other, and it also gives a very systematic and methodical way of modelling. And there could be Bayes nets could be built automatically in a data driven fashion or it could be hand crafted as well hence while it used in the industry in the machinery learning community.

So, from Bayesian networks, so what kind of questions can come from Bayesian networks? The first question that can come is given certain so this Bayes net will be given the prior distributions will be given, the conditionals will be given. And you will be asked to calculate the joint distributions of these random variables. So, say the probability of A, so let us write down some values. Say probability of A equal to 0 is 0.2, probability of A equal to 1 is 0.8.

Probability of B equal to 0 is 0.6; probability of B equal to 1 is 0.4. And this C B D is also given, so these will be there will be this will be A pictables, so I will tell you. So, say this is like A and B and these are the different values of C equal to 0, and C equal to 1. And let us and this A and B values. So, I am writing here is probability of C given A and B. So, this quantity is going to be for 0 A equal to 0 B equal to 0. Say probability of C equal to 0 is 0.2, this is 0.8. Say for 0 1 its 0.6, 0.4. Say for 1 0, it is 0.3, 0.7. Say for 1 1, it is 0.5, 0.5. And also probability of D given C this will be given in the exam in the question. So, D equal to 0, D equal to 1; and C equal to 0 C equal to 1. So probability of D equal to 0 given C equal to 0 say this is 0.1 0.9, and this is 0.9, 0.1. Let us have this kind of distributions. So, you have been asked to calculate or given this model of probability.

(Refer Slide Time: 32:55)



Graphical Model. → models Conditional Probability Distributions among the random variables.

B, C, D.

$P(A,B,C,D) = \underline{P(A)P(B)P(C|A,B)P(D|C)}$

P(C|A,B)

| A B | C=0 | C=1 |
|---|---|---|
| 0 0 | 0.2 | 0.8 |
| 0 1 | 0.6 | 0.4 |
| 1 0 | 0.3 | 0.7 |
| 1 1 | 0.5 | 0.5 |

P(D|C)

| | D=0 | D=1 |
|---|---|---|
| C=0 | 0.1 | 0.9 |
| C=1 | 0.9 | 0.1 |

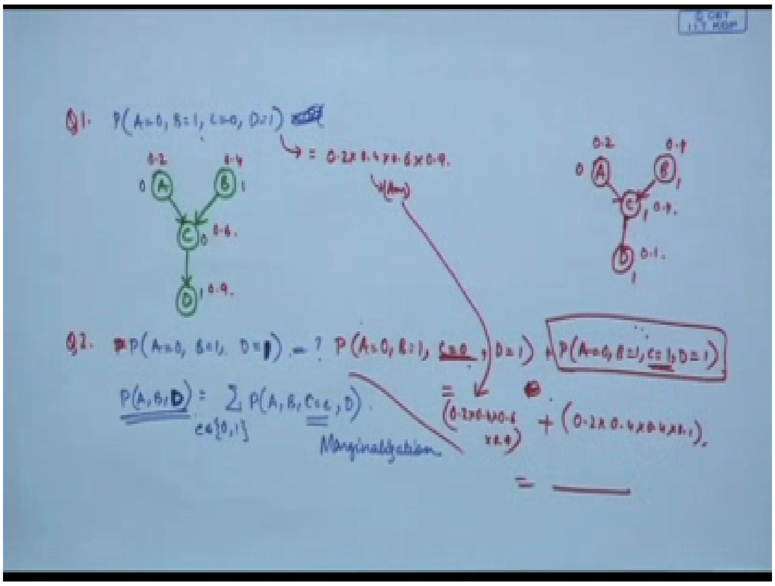
P(A=0, B=1, C=0, D=1)

Suppose you have been ask to calculate probability of A equal to 0, B equal to 1, C equal to 0, D equal to 1, you will be give given this kind of question. So, what will this be, this is going to be probability, so as I said before so I will tell you how to solve this problem in a very elegant and in a nice way. So, you draw the decision tree the sorry the Bayes net pack again draw the Bayes net again. You draw the nodes now and it is always good to use different colours while solving this question. Now the questions says that the values of the variables if like this so A is going to take A value of 0, B is going to take 1, C is going to be 0, D is going to be 1, and so you write these values and now you look up the tables and find out.

So, probability of A equal to 0 is 0.2, so you write 0.2 here, so this is P of A. P of B, B is going to be 1, so it is 0.4. Probability of C given A equal to 0, B equal to 1, probability of C equal to 0, given A equal to 0 B equal to 1, so this entry is what I what are we are looking for. So, C equal to 0 given A equal to 0 given B equal to 1, so this is going to be 0.6. Now D equal to 1, given C equal to 0; D equal to 1, given C equal to 0, this quantity, so this is going to be 0.9.

Now you multiply all of them together. So, all we are going to do is to do this product, this product and the values have been noted down here so you simply multiply these values. So, this is going to be equal to 0.2 times 0.4 times 0.6 times 0.9. Yes, multiply this out this is going to be the value, this is going to be the answer. Let us multiply and find the value. So, this kind of question is going to come.

(Refer Slide Time: 35:14)



Q1. $P(A=0, B=1, C=0, D=1)$

$= 0.2 \times 0.4 \times 0.6 \times 0.9$

Q2. $P(A=0, B=1, D=1) = ?$ $P(A=0, B=1, \underline{C=0}, D=1) + \boxed{P(A=0, B=1, \underline{C=1}, D=1)}$

$\underline{P(A,B,D)} = \sum_{c\in\{0,1\}} P(A,B,\underline{C=c},D)$

Marginalization

$= (0.2\times0.4\times0.6\times0.9) + (0.2\times0.4\times0.4\times0.1)$

$=$ ____

Now, suppose you do not know the value of C, so this is one question. The second question that can come from this part is you have been asked just this part let me write in blue again P of A equal to 0, B equal to 1 and D equal to 0 say, D equal to 1 I am sorry let us keep D equal to 1. So this has been asked, what is this. So, this is basically asking for probability of A, B and D, which is equal to summation over C taking values in 0 and 1 of probability of A, B, C equal to C and D. So, you are marginalizing so this particular thing is called marginalization of a probability distributions. So you are marginalizing C out you summing it up over all the different values of C. And all you end up with the interest of the variables this is D.

So, suppose you have been given this question. So, what you are going to do so this quantity is going to be equal to P of A equal to 0, B equal to 1, D equal to 1, for C equal to 0. And then an another term A equal to 0, B equal to 1, C equal to 1, D equal to 1. We are marginalizing C out, we are summing it up over summing the terms probability values for this different value for the different values of C.

So, this quantity as we calculated before is 0 this quantity right so this plus this quantity. So whatever this number is do not have a calculator right now. So, just calculate this so this comes in sits here plus you calculate this quantity in the same way as before, so you calculate so what is this going to be so this will result in Bayes net which looks like this. Now the values are 0, 1, C equal to 1, and D is also equal to 1, so the priors will be 0.2 and 0.4 will be where this conditional will change.

So, probability of C equal to 1 given A equal to 0 and B equal to 1, C equal to 1, given A equal to 0, and B equal to 1 is 0.4, so the value here will be 0.4. And the probability of D equal to 1 given C equal to 1 is 0.1. So, just go and put 0.1 here. So, this is going to be equal to so let me right it down 0.2 into 0.4 into 0.6 into 0.9 this quantity plus this is going to be 0.2 into 0.4 into 0.4 into 0.1, so add these to up the answer is there, so you will be asked this kind of questions.

So, that the topics that we cover today are estimation of probability using the frequency definition very, very basic and the most important bayes rule, MAP inference, Naive Bayes assumption, Bayesian networks as structure of Bayesian networks what it does it models the condition of probability distributions and also the conditional independence condition independence assumptions rather. Then we also discussed how to do inference in A Bayes net so given certain values of its of the random variables, how to calculate the joint probability distributions, and also how to do the marginalization over different variables which have not been asked in the question, and given an inference for the joint distributions of the subset of the variables of the Bayes net.

So, this is the content that was covered this week. All the questions in the assignment 4 are going to come from these topics, so best of luck for solving the assignment. And the deadline will not be extended, so make it very make sure make sure that you start for you know early enough, and you finish before the deadline which will not be extended by any means, and practise this problem for the exam, because you are going to get quite similar questions.

So, bye-bye, see you next time.