PROBABILITY THEORY FOR DATA SCIENCE

Prof. Ishapathik Das

Department of Mathematics and Statistics

Indian Institute of Technology Tirupati

Week - 02

Lecture - 07

Bayes' Theorem

Now, we will discuss Bayes' theorem. But before that, let's go over a particular theorem known as the *Total Probability Theorem*, or the *Law of Total Probability*. So, let's discuss that. What does it say? We've already talked about this kind of theorem before, but we're repeating it here.

Let $A_1, A_2, ..., A_n$ be a set of pairwise mutually exclusive and exhaustive events. What does this mean? Pairwise mutually exclusive means that the intersection of A_i and A_j is empty if $i \neq j$. A mutually exclusive and exhaustive set of events means that the union of $A_1, A_2, ..., A_n$ is equal to the sample space S. So, suppose this is the sample space, and then this is just a partition.

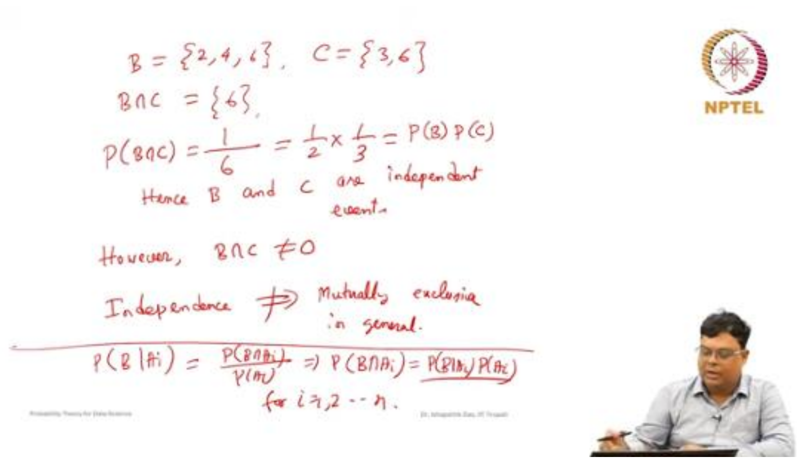
So, we have A_1, A_2, A_3, A_4, and so on, up to A_5, and $\emptyset$. They're pairwise mutually exclusive; they're disjoint sets pairwise, and if you take their union, it will be S. Then, for any event—let's say B—the probability of B can be written as the summation of *i* from 1 to *n*, with $P(B|A_i) \times P(A_i)$. So, this is the conditional probability. And this is also the same as the summation from $i = 1$ *to n* of $P(B \cap A_i)$.

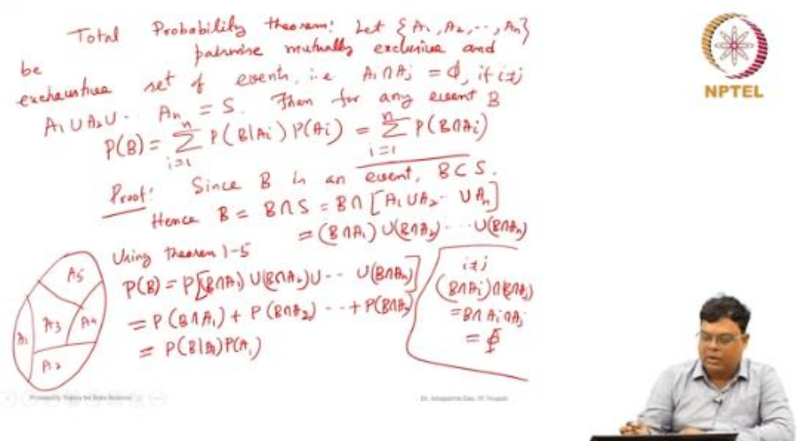
We've actually already proved this. So, let us go through it again. This is just a different expression. Now, because $A_1, A_2, ..., A_n$ are pairwise disjoint and an exhaustive set of events, let's consider an event B. Since B is an event, B is a subset of S.

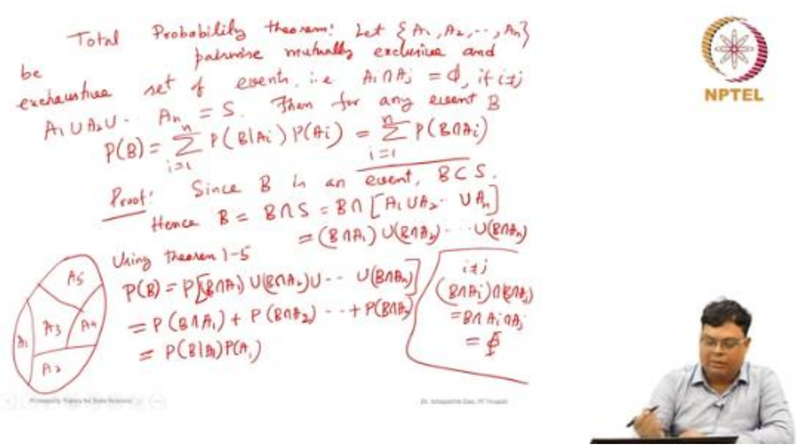
So, B can be written as $B \cap S$. Now, S can be represented as the union of $A_1, A_2, ..., A_n$, because these are pairwise disjoint and exhaustive. Then, using the distributive property, this can be represented as $B \cap A_1 \cup B \cap A_2 \cup ... \cup B \cap A_n$. Now, all these sets will be pairwise disjoint because, if $i \neq j$, then $B \cap A_i \cap B \cap A_j$ is the same as $B \cap A_i \cap A_j$. Since A_i and A_j are pairwise disjoint, this will be $\emptyset$.

So, that's why they are disjoint—they are pairwise disjoint. Now, using Theorem 1.5, this is just a finite version of Axiom 3. The probability of B can be represented as, actually, Theorem 1.5 is used here: $B \cap A_1 \cup B \cap A_2 \cup ... \cup B \cap A_n$. Because they are pairwise disjoint, we use Theorem 1.5, which says that we can add $P(B \cap A_1) + P(B \cap A_2) + ... + P(B \cap A_n)$. So, that's the expression we got.

Now, this means that $P(B \cap A_n)$ can be represented as $P(B|A_1) \times P(A_1)$. So, that can be found. Suppose, here you can see $P(B \cap A_i)$. So, what is the conditional probability? By definition, this is simply $P(B \cap A_i) / P(A_i)$, assuming that $P(A_i) \neq 0$.

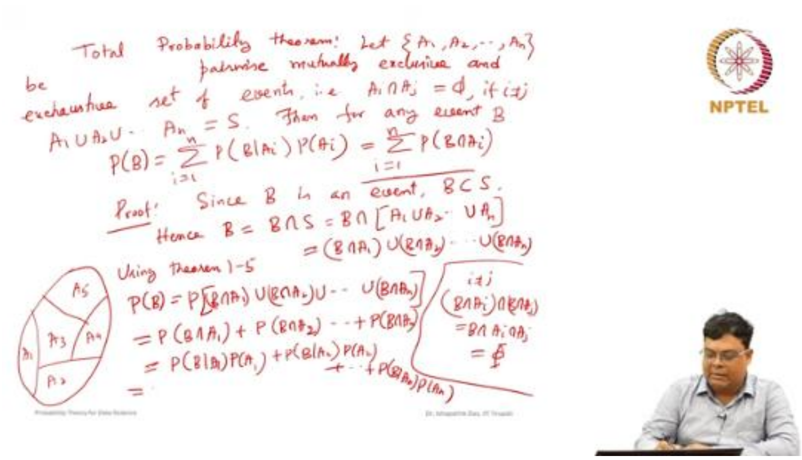




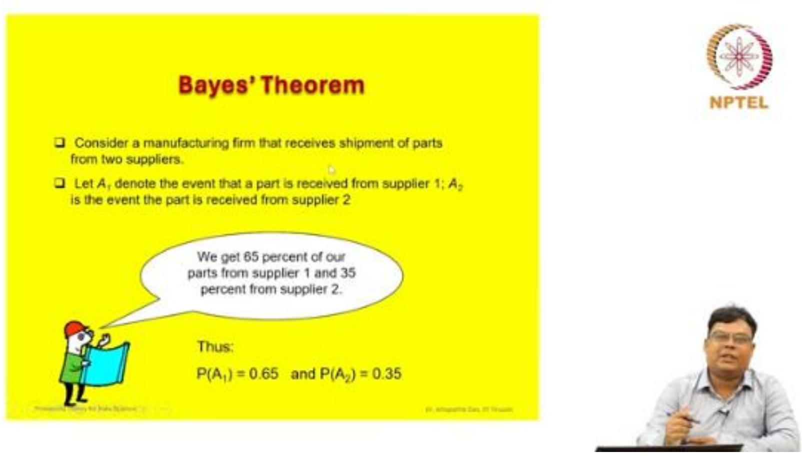


This implies that $P(B \cap A_i)$ can be represented as $P(B|A_i) \times P(A_i)$ for *i* from 1 to *n*. Hence, this can be represented as $P(B|A_2) \times P(A_2) + P(B|A_n) \times P(A_n)$. So, in short form, we can write this as the summation from $i = 1$ *to n* of $P(B|A_i) \times P(A_i)$.

This is useful, and we will discuss it whenever we talk about Bayes' theorem. There are also some examples where the total probability is used for computing this kind of probability.



Suppose we have $P(B|A_i)$ and $P(A_i)$; then any event B can be represented this way.Now, we will discuss Bayes' theorem. You may have heard of it or learned about it. If you have learned it, then it will be very simple to understand. But I am assuming that you do not know this theorem, so I will discuss it very slowly from the beginning, starting with one example.



We will also study some more examples to understand Bayes' theorem better. Bayes' theorem is very useful when working with data sets. Sometimes, some of the conditional probabilities are known and straightforward to identify, but the inverse probabilities are not known. For example, suppose we have attendance data for students based on whether it is a rainy day or a non-rainy day. The probability that a student is present on a rainy day, given that the day is rainy, may be known.

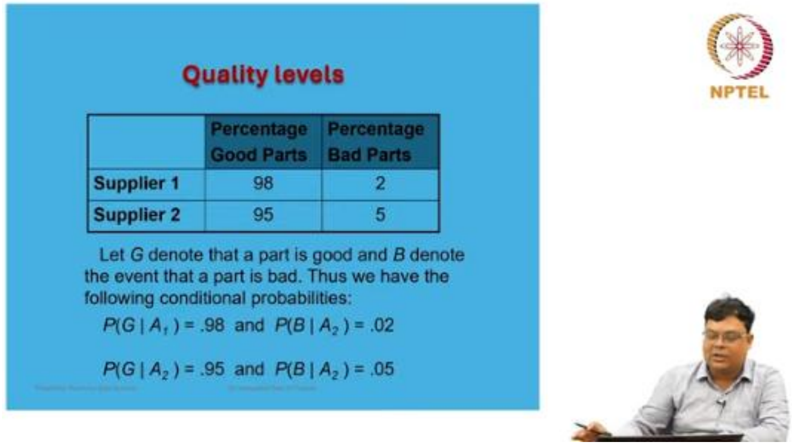
We can also know the probability that a student will be absent, given that the day is rainy, not rainy, or sunny. Additionally, the probability of it being a rainy day or not can be known from the data set. Now, suppose we ask the question in the opposite direction: if a student is absent from class or present in class, what is the probability that the day is rainy? The probability of the student being present or absent, given that the day is rainy, is straightforward to find from the data set. But if we ask the opposite—if we know that the student is present or absent, what is the probability that it was a rainy day?

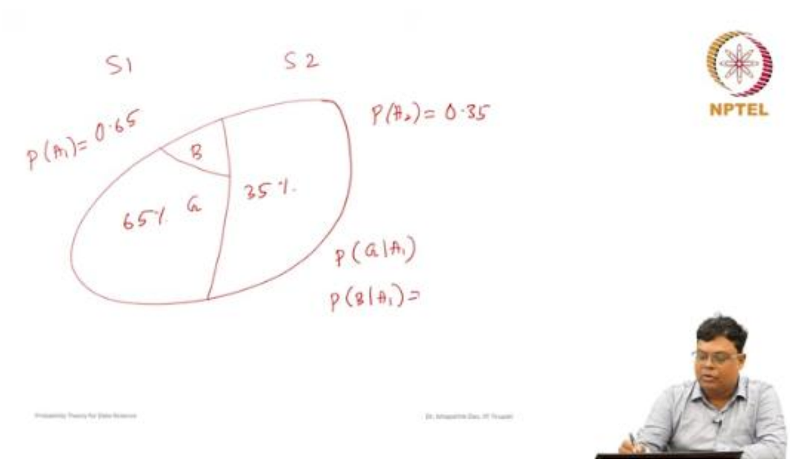
That is very useful. We will discuss more examples related to this. Here, consider a manufacturing firm that receives shipments of parts from two suppliers: Supplier 1 and Supplier 2. The manufacturer takes parts from both suppliers, and we define these events as A_1 and A_2. So, A_1 is the event that a part will be from Supplier 1, and A_2 is the event that a part will be from Supplier 2. Now, we know that we get 65% of our parts from Supplier 1.

That means, if we take a simple classical approach, the probability of receiving a part from Supplier 1 is 65%. This indicates that, for any part you take, there is a 65% chance it will come from Supplier 1. What is the probability that it is from Supplier 1? The probability of A_1 is 0.65.

The probability of A_2 will be 1 - 0.65 = 0.35. So, there is a 35% chance that it is from Supplier 2. Now, if we condition on the data that it is from Supplier 1, what is the probability that this part is a bad part? What is the probability that it is a good part? We know this information, so the probability that the part is good, given that it is from Supplier 1, and the probability that it is a bad part, can be analyzed from the data.

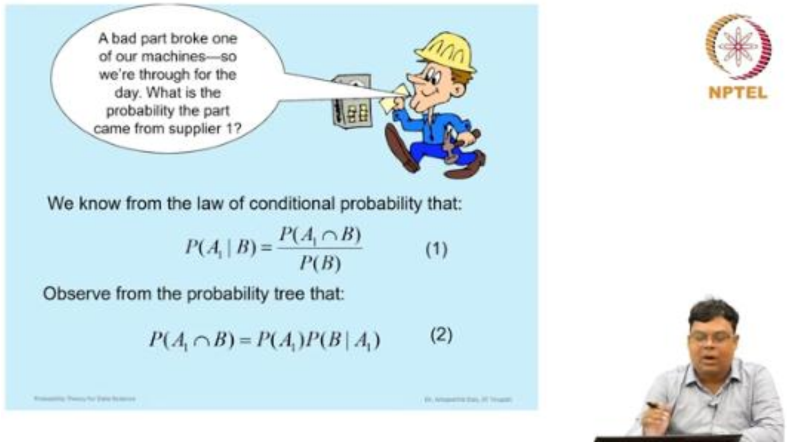
So, this is also given here. So, here it is: the quality levels. The percentage of good parts from Supplier 1, given that they are from Supplier 1, is 98%. This means $P(\text{good} \mid A_1) = 0.98$. For a bad part from Supplier 1, it would be 1 - 0.98 = 0.02, i.e., $P(\text{bad} \mid A_1) = 0.02$.





Similarly, out of the 35% from Supplier 2, the probability of a bad part, given that it is from Supplier 2, is 95%, or 0.95. So, the probability that it is a good part from Supplier 2 is 0.05, which is 1 - 0.95. All this information is very straightforward. Now, suppose it happens that a bad part broke one of our machines. Suddenly, we find that a part is bad, so we are through for the day.

What is the probability that the part came from Supplier 1? All these probabilities are given to us. Based on that, it is asked: given that a part is bad, what is the probability that it is from Supplier 1? So, we need to find the probability that it is from Supplier 1 given that the part is bad. This means we are looking for $P(A_1 \mid B)$.



Similarly, you can also ask what the probability is that the part is from Supplier 2 given that it is bad. Now, let's find out the probability of A_1 given B. This is nothing but $P(A_1 \cap B) \div P(B)$. We just saw that $P(A_1 \cap B)$ can be represented as $P(B \mid A_1) \times P(A_1)$. So, by definition, we have $P(A_1 \cap B) = P(B \mid A_1) \times P(A_1)$.

That's why I wrote it this way, and we divide by P(B). Now, how can we find P(B)? That is the question. We can compute this using the theory of total probability. Here, $A_1 \cup A_2$ is just the whole sample space S.

This means A_1 is for Supplier 1 and A_2 is for Supplier 2. All the parts are coming from either Supplier 1 or Supplier 2, so all parts belong to $A_1 \cup A_2$, which equals S. Also, A_1 and A_2 are disjoint events. In this case, $A_1 \cap A_2 = \emptyset$ because any part must come from either Supplier 1 or Supplier 2, but not both. So, they are disjoint. By using the total probability theorem, we can express P(B) since B is a subset of S. We can represent P(B) as $P(B \cap S)$, which is just $P(B \cap (A_1 \cup A_2))$. By using the distributive property, this becomes $P(B \cap A_1) \cup P(B \cap A_2)$. Hence, because they are disjoint, we have already shown that $P(B \cap A_1 \cap B \cap A_2)$ is nothing but $P(B \cap A_1 \cap A_2)$, which is $P(B \cap \emptyset)$. This is the null set because A_1 and A_2 are disjoint. Therefore, they are pairwise disjoint as well.

Hence, P(B) will be equal to $P(B \cap A_1) \cup P(B \cap A_2)$. Since they are disjoint, using axiom 3, we can express this as $P(B \cap A_1) + P(B \cap A_2)$. This can be represented as $P(B \mid A_1) \times P(A_1) + P(B \mid A_2) \times P(A_2)$. This is the total probability. We will replace here, and this is called Bayes' theorem for two events.

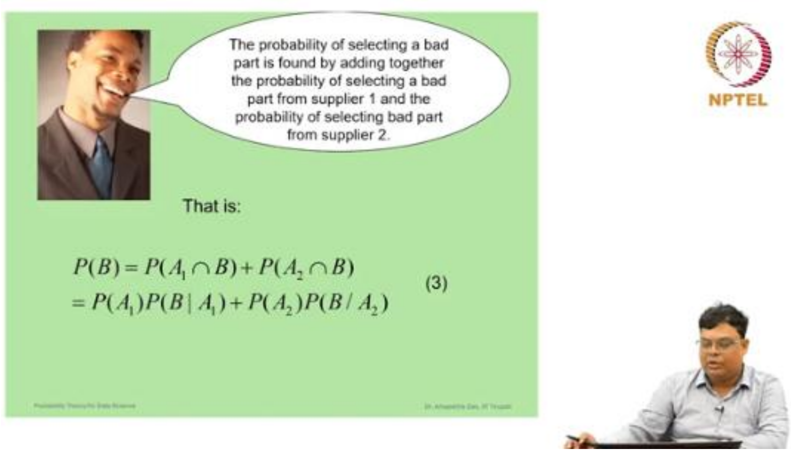
So, $P(B) = P(B \mid A_1) \times P(A_1) + P(B \mid A_2) \times P(A_2)$. This is the probability of A_1 given B.

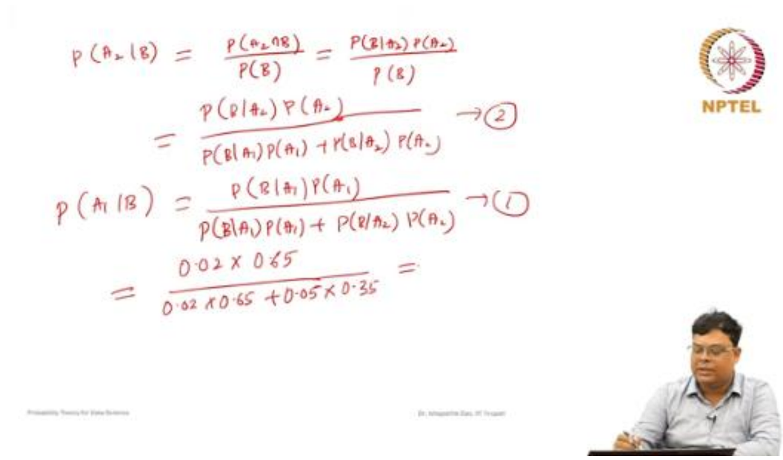
Anyone can ask what the probability is that the part is bad, but it is coming from Supplier 2. It can be found similarly. $A_2 \cap B$, by the definition of conditional probability, is $P(A_2 \mid B)$, which is nothing but $P(A_2 \cap B) \div P(B)$.

This can be represented as $P(B \mid A_2) \times P(A_2) \div P(B)$. $P(B)$, which we computed earlier, is $P(B \mid A_1) \times P(A_1) + P(B \mid A_2) \times P(A_2)$. Note that the sum of $P(A_1 \mid B)$ and $P(A_2 \mid B)$ will equal 1 because this represents the total probability of all possible outcomes. Now, we will numerically compute all these values since they are known to us. Let's write down this equation again for clarity:

$P(A_1 \mid B) = P(B \mid A_1) \times P(A_1) \div (P(B \mid A_1) \times P(A_1) + P(B \mid A_2) \times P(A_2))$. As mentioned, the sum of these two values will equal 1.

Now, we will replace the values here to find the desired probability. What is $P(B \mid A_1)$? It is 0.02. $P(A_1)$ is given as 65%, or 0.65. Thus, we have 0.02×0.65. Next, what is $P(B \mid A_2)$? This is 0.05. $P(A_2)$ is given as 35%, or 0.35. So, then if you compute this numerically, whatever the value it will come. You have to use a calculator. So, you have to use a calculator. So, here everything is explained. Now you can see that this is Bayes' theorem for two events that we have already discussed

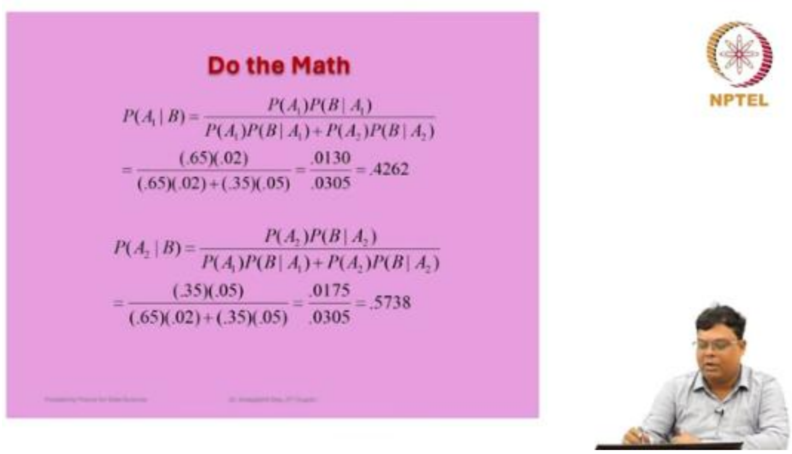




The computation of 0.65 * 0.02 gives us the value of 0.4262. You can review this again. All the explanations are also provided in the slides. For A2 given B, you can replace the values, and it comes out to be 0.5738. So, now you can see that this is the Bayes' theorem for two events that we have already explained. And this is the computation 0.65 * 0.02.

This value is coming 0.4262. So, this value is 0.4262. So, you can go through it again. So, this is whatever we explained here. So, that is also given in the slides, and A2 given B also has its values replaced.

So, you can try here. So, you replace this value, the numerical value, and this is coming 0.5738. Now, if you add these two values, this should be equal to 1. So, this is known as Bayes' theorem for two events. Now, we will discuss Bayes' theorem for n events.



This is just an extension for n events. Let A1, A2, ..., An be a pairwise mutually exclusive and exhaustive set of events. So, these are the notations used here: A1, A2, ..., An. For any event B, we assume that the sample space S and sigma field C follow similar notation, meaning $B \subseteq S$. For any event B, the probability of a specific event Ai given B can be represented as $P(A_i \mid B) = P(B \mid A_i) * P(A_i) / \Sigma (P(B \mid A_j) * P(A_j))$, where $j = 1$ to n.

Now, what is the proof? Similarly, this can be proved for two events; it's just an extension for n events. It's very straightforward, and we will go through it. This is just an extension of Bayes' theorem up to n events. Let $A_1, A_2, \ldots, A_n$ be pairwise mutually exclusive and exhaustive events. So, as you can see, we have the notation here: $A_1, A_2, \ldots, A_n$.

For any event B, this will be in a sigma field. Here, we assume the sample space S and sigma field C, following the same notation, so that any event $B \subseteq S$. Now, for any event B, suppose if you consider any "i"—say, supplier "i" among n suppliers—then the probability of a bad part from A_i given B can be represented as $P(A_i \mid B) = P(B \mid A_i) * P(A_i) / \Sigma (P(B \mid A_j) * P(A_j))$, where $j = 1$ to n. What’s the proof? Similarly, this can be shown for two events; it’s just an extension up to n events. This is very straightforward. We’ll just go through it.

The probability for $i = 1$ to n is fixed here. You take $P(A_1 \mid B)$, $P(A_2 \mid B)$, and find the probability by just changing the value of i. For fixing i, where $i \in \{1, 2, ..., n\}$, $P(A_i \mid B) = P(A_i \cap B) / P(B)$, assuming that $P(B) \neq 0$. This is the definition of conditional probability, and it can be represented as $P(B \mid A_i) * P(A_i) / P(B)$.

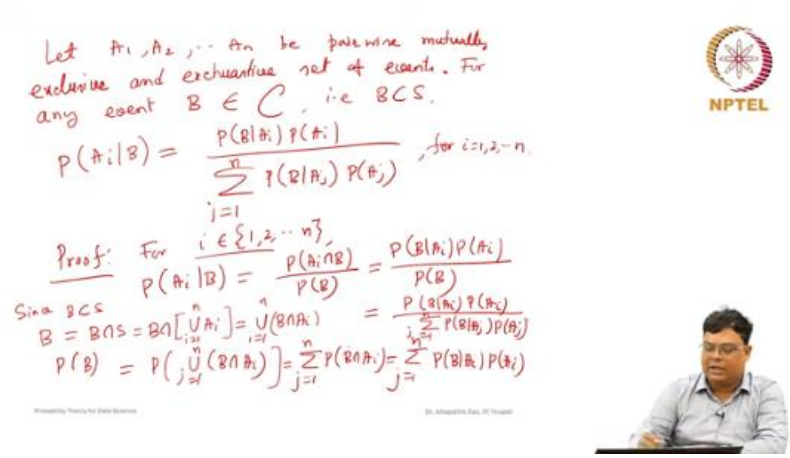
Now, from the theory of total probability, this probability of B can be computed. Since $B \subseteq S$, B can be represented as $B \cap S$. Because $A_1, A_2, ..., A_n$ are pairwise mutually exclusive and exhaustive sets of events, the union of A_i is S. Thus, this can be represented as Σ ($i = 1$ to n). By the distributive property, we can say they are pairwise disjoint. Therefore, $P(B) = \Sigma$ ($i = 1$ to n) $P(B \cap A_i)$. This can be represented as Σ ($i = 1$ to n) $P(B \mid A_i) * P(A_i)$. So, this is the probability of B.

We replace this here, and finally, we get Bayes' theorem: $P(A_i \mid B) = P(B \mid A_i) * P(A_i) / \Sigma$ ($j = 1$ to n) $P(B \mid A_j) * P(A_j)$. I made a mistake because here, we are fixing i, so we cannot use this variable again. So, we will take $j = 1$ to n here instead. When i is fixed, you can change one of the values to get the probability of A_i given B. So, this is nothing but Σ ($j = 1$ to n) $P(B \mid A_j) * P(A_j)$.

Now, from the theory of total probability, this probability of B can be computed since $B \subseteq S$. So, B can be represented as $B \cap S$. Because $A_1, A_2, ..., A_n$ are pairwise mutually exclusive and exhaustive events, the union of A_i is S. This can then be represented as Σ ($i = 1$ to n). By the distributive property, we can write that they are pairwise disjoint.

So, the probability of B can be represented as $P(\bigcup$ ($i = 1$ to n) $B \cap A_i)$. Now, because they are pairwise disjoint by axiom 3 or theorem 1.5, this is nothing but Σ ($i = 1$ to n) $P(B \cap A_i)$. So, that can be represented as Σ ($i = 1$ to n) $P(B \mid A_i) * P(A_i)$. So, this is the probability of B we replace here. Then finally, we get this as Bayes' theorem: P(Ai | B) = P(B | Ai) * P(Ai) / Σ (j = 1 to n) P(B | Aj) * P(Aj). I made a mistake because here we're fixing i, so we can't use this variable again.

So, this is where we'll take j = 1 to n, here also j = 1 to n, because i is fixed here. You can change one of the values, then you'll get the probability of Ai given B. So, this is nothing but Σ (j = 1 to n) P(B | Aj) * P(Aj). So, this is Bayes' theorem; I just proved it. We've already gone over these things, like the theory of total probability when we discussed it earlier.



Then here, we can replace the probability of B. It's just conditional probability, by definition of conditional probability, and we see how this is derived. So, this is Bayes' theorem. So, let us do this. We have already discussed one example.

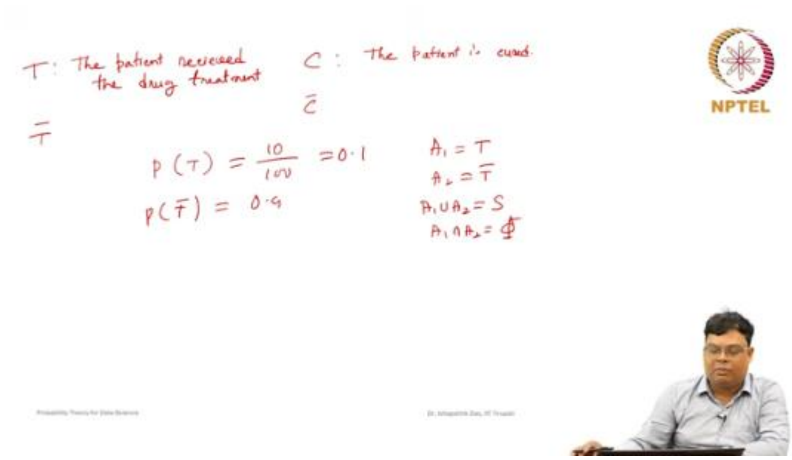
This is the Bayes' theorem for n events, which is given here. You can go through it again. Now, let us discuss a numerical example. There are 100 patients in a hospital with a certain disease. Of these, 10 are selected to undergo a drug treatment that increases the cure rate from 50% to 75%.

Here it is actually given that whenever someone receives treatment, we denote that event as T. If a patient did not get the treatment, we denote that event as T complement (T^c). So, we can write that T is the event that the patient got the treatment. So, it is already given that if a patient is cured, we denote this event as C, and if not, it is denoted as C^c.

Out of 100 patients in the hospital, 10 are selected to undergo drug treatment. Now, what is the probability that a patient received the drug treatment if the patient is known to be cured? So, C denotes the event that the patient is cured, and C^c is its complement. T denotes the event that the patient received the drug treatment, and T^c denotes that the patient did not receive the drug treatment.

Since 10% of the patients received the treatment, the probability of T is P(T) = 0.1, and the probability of T^c is $P(T^c) = 1 - P(T) = 0.9$.

Here, if we represent it using Bayes' theorem, A1 will be T, and A2 will be T^c. The union of A1 and A2 will be all patients—this is the entire sample space S. Additionally, any patient either received the drug treatment or did not, so $A1 \cap A2 = \emptyset$ (null set). So, that is what we found from this problem. Now, 10 patients are selected to undergo drug treatment, increasing the cure rate from 50% to 75%.



This means that if the patient receives the drug treatment, the cure rate is 75%. So, the conditional probability given here is that P(C | T) = 0.75. Similarly, the probability that a patient is cured, given they did not receive the drug treatment, is 50%, or $P(C \mid T^c) = 0.5$.

We need to actually formulate this problem using events because Bayes' theorem is represented by events. So, we have to transform the information in the problem into the form of our formula so that we can use Bayes' theorem.

Here we have to understand that there are 100 patients, and 10 are selected to undergo drug treatment. This means 10% of the patients got the treatment, so P(T) = 0.1. Therefore, $P(T^c) = 0.9$. Now it is given that the drug treatment increases the percentage of the cure rate from 50% to 75%. This means P(C | T) = 0.75. Earlier, the probability that the patient is cured given that the patient did not receive the drug treatment was 50%. So, given T^c, $P(C \mid T^c) = 0.5$, and given T, P(C | T) = 0.75.

Now, what is the question? The question is: what is the probability that the patient received drug treatment? That means T is given here, and if the patient is known to be cured—C is given here. So, what is the probability that the patient received drug treatment? We want to find P(T | C), which is the probability that the patient received drug treatment given that it is known that the patient is cured. So, this probability is what is being asked. All this information is given, but this cannot be found straightforwardly. We have to use Bayes' theorem.

By Bayes' theorem, it is expressed as:

$$P(T \mid C) = (P(C \mid T) \times P(T)) / (P(C \mid T) \times P(T) + P(C \mid T^c) \times P(T^c)).$$

Assume that T is A1, T^c is A2, and C is B in Bayes' theorem. Then it is P(B | A1), which can be expressed as P(A1 | B). Therefore,

$$P(A1 \mid B) = (P(B \mid A1) \times P(A1)) / (P(B \mid A1) \times P(A1) + P(B \mid A2) \times P(A2)).$$

So, you can see the Bayes' theorem for two events that we discussed. So, here this is the probability of A1 given B, nothing but:

$$P(A1 \mid B) = (P(A1) \times P(B \mid A1)) / (P(A1) \times P(B \mid A1) + P(A2) \times P(B \mid A2)).$$

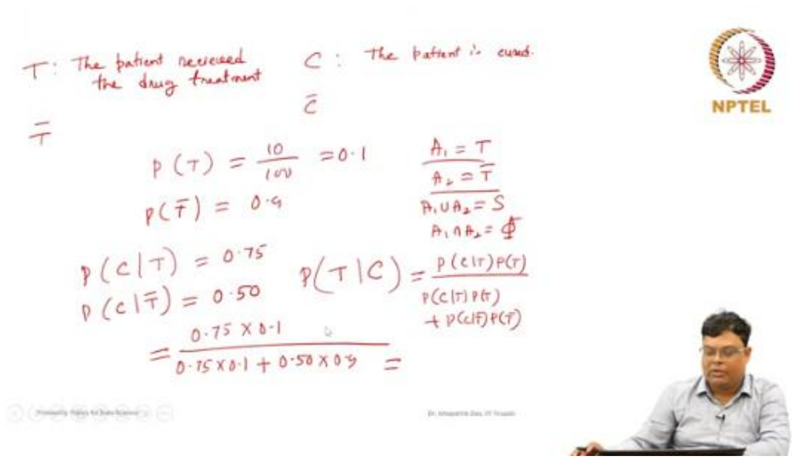
So, if you relate this variable in this way, then you can use the Bayes' theorem here. So, then we will just find out:

$$P(C \mid T) = 0.75 \times P(T) = 0.75 \times 0.1.$$

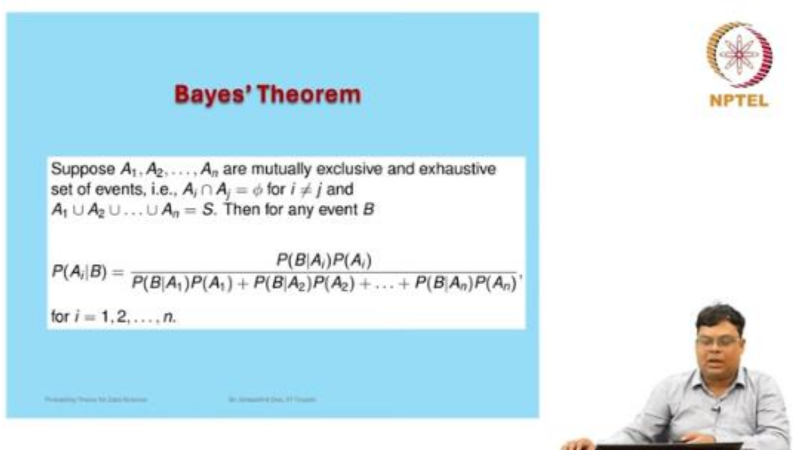
By:

$$P(C \mid T) = 0.75 \times 0.1 + P(C \mid T^c) = 0.50 \times P(T^c) = 0.50 \times 0.9.$$

So, just you have to calculate; you need a calculator. So, actually this value is not provided here. So, you can compute it using a calculator, and you can find out the value. So, this will be the final answer. You can find out the closed, simplified form of this value. So, this is one example of how you can use Bayes' theorem.



You can go through more numerical problems, and then you will understand more clearly how to use the Bayes theorem. So, here we only studied two numerical examples. So, there are many other numerical examples you can find to use the Bayes theorem and how to solve it, which will make it more clear to you. So, we completed the first part of this discussion, starting from the definition of probability, the sample space, and the definition of events, sigma field, or sigma algebra. Then we discussed the classical approach and the axiomatic approach, and before that, the classical approach and frequency approach, along with their drawbacks.



Then we covered the axiomatic approach to defining probability and discussed some important theorems. We proved those theorems, studied some numerical examples, and then discussed conditional probability and its applications. We talked about independence and the difference between independence and mutually exclusive events, the total probability theorem, and finally, we discussed the Bayes theorem. We went through two numerical examples for Bayes theorem, so I hope you are following along and that you understand. You can solve many more problems using these results.