Lecture 10
Introduction to Probability

Alice Gao

June 7, 2021

Contents

1 Learning Goals

By the end of the lecture, you should be able to

- Give reasons for modelling uncertainty in the first place.
- Understand terminologies such as random variable, joint probability distribution, prior unconditional probability, and posterior or conditional probability.
- Come up with random variables to model the important elements in a story, and calculate the number of probabilities in the join probability distribution.
- Calculate the probability over a subset of the variables using the sum rule given the joint distribution.
- Calculate a conditional probability using the product rule given a joint distribution.
- Calculate a joint probability over a subset of the variables using the chain rule.
- Calculate a conditional probability using the Bayes' rule.
- Interpret the Bayes' rule in a different way, where the denominator in the equation is thought of as a normalization constant.

2 Introduction to Uncertainty

2.1 Why handle uncertainty?

So far, we've considered a world without uncertainty. In search and supervised learning problems, all the information is available to us, and we can use the information to perform inference and to make decisions.

Why do we need to think about uncertainty in the first place?

First, when an agent is navigating the world, the agent cannot observe everything. The information that is not available to us constitutes a large part of our uncertainty.

Second, the agent, whether it's a robot or a program, is imperfect. When the agent performs an action, an action may not produce the intended consequences. This leads to another possible source of uncertainty.

Even though we are in a world with so much uncertainty, we need to take actions. We often have to make a decision based on missing, imperfect or noisy information. The best thing we can do is quantify our uncertainty about the world and use our uncertainty to perform inference and to make decisions.

A student once asked me a philosophical question about making decisions under uncertainty in real life. The question was: Suppose that I make an important decision and the outcome turns out to be not as good as I hoped. If I could go back in time, would I make a better decision? This type of what-if question can really keep a person up all night. What is your opinion on this?

Here's my personal opinion. I believe that even if I can go back in time, I would still make the same decision. The reason is that, at the time, I did not have as much information as I do now. I made the best decision I could given the limited information that I had at the time. Even if I could go back in time, I would still make the same deicsion. This is just my opinion. I would be curious to see what you think about it.

2.2 Frequentists v.s. Bayesians

Next, let me discuss two views of probabilities.

We live in a world full of uncertainty. Formally, we can use probabilities to measure uncertainty. When it comes to probabilities, there are two camps, called Frequentists and Bayesians.

The Frequentists believe that probabilities are objective — something we can observe in the world. If we want to calculate a probability of an event, we can count the number of times a outcome of the event occurs and use that to calculate the probability.

For example, suppose that we want to know the probability that a coin will turn up heads if we do a coin toss. A Frequentist will do many coin tosses and determine the probability as the fraction of heads observed. Basically, the probability is based on historical observations.

The advantage of this view is that since probabilities are observable, it's something that we can agree on. The disadvantage is that a Frequentist cannot estimate a probability if they haven't observed anything.

In contrast, a Bayesian thinks about probability as something subjective. A probability is just a degree of belief in our head. Bayesians believe that we have some prior belief about a probability based on our previous experience. As we make observations, we will update this belief based on these observations.

Consider the coin flip example again. Before observing any coin flip, two Bayesians may already have different prior beliefs about the probability for the coin to turn up heads. These prior beliefs are based on their different prior experience. In general, different people can have different probabilities in their heads for the same event. As the two Bayesian observe some coin tosses, they will update their beliefs. The more coin flips they observe, the similar their updated beliefs will be.

The advantage of the Bayesian view is that we can have a probability estimate even if we haven't observed anything. The disadvantage is that two people may not agree on what a probability should be if they have different prior experience.

In this course, we are going to adopt the Bayesian view of probability.

2.3 Random Variables

In order to use probabilities to model the world, we'll need to come up with random variables. Each random variable describes an event. It has a domain, which contains all the values that the random variable can take. There is an associated probability distribution, which specifies a probability for each value in the domain.

Here is a simple example. Suppose we are modeling a scenario involving burglary and alarm. If a burglary happens, then the alarm may go off. One random variable could be whether the alarm is going off or not and the domain would contain true and false. The distribution could specify that the alarm does not go off 90% of the time.

In this unit, we will primarily deal with Boolean random variables. Each Boolean random variable can be either true or false. We will use a simplified notation for Boolean random variables. We will use A and not A to denote A = true and A = false respectively.

2.4 Axioms of Probability

There are some common properties that all probability functions satisfy. They are called axioms of probability.

The first axiom says, every probability is between 0 and 1. Interestingly, the choice of 0 and 1 is purely a convention. Any other number range would also work. We need to restrict these values to a range so that they are comparable in different scenarios.

The second axiom says that if something is true, then its probability is 1. If something is false, then its probability is 0.

The third axiom is the inclusion/exclusion principle. We can calculate the joint probability of two events A and B using the given formula. When we add the probabilities of A and B, we double counted the probability that A and B both happen. So we need to subtract that.

All of the probability functions we consider satisfy these axioms.

2.5 Prior and Posterior Probabilities

Let me introduce two more terminologies. Remember that we adopted a Bayesian view of probabilities. As a Bayesian, we have a prior belief, and we will update this prior belief based on observations.

When we have a probability with no evidence, it's a prior or unconditional probability. This is our probability estimate of X when we haven't made any observation about X.

Once we make new observations, for example Y, we will update our belief about X given the observation. $P(X|Y)$ is called a posterior or conditional probability.

These two terminologies go hand in hand. We have a prior (unconditional) belief before observing anything, and we have a posterior (conditional) belief after getting some observations.

2.6 The Holmes Scenario

We are going to use the following story as a running example.

Example: Mr. Holmes lives in a high crime area and therefore has installed a burglar alarm. He relies on his neighbors to phone him when they hear the alarm sound. Mr. Holmes has two neighbors, Dr. Watson and Mrs. Gibbon.

Unfortunately, his neighbors are not entirely reliable. Dr. Watson is known to be a tasteless practical joker, and Mrs. Gibbon, while more reliable in general, has occasional drinking problems.

Mr. Holmes also knows from reading the instruction manual of his alarm system that the device is sensitive to earthquakes and can be triggered by one accidentally. He realizes that if an earthquake has occurred, it would surely be on the radio news.

Problem: Come up with the random variables to describe the Holmes Scenario.

Solution: The following random variables model this story. It is possible to come up with different ones depending on your interpretation.

- B: A burglary is happening.
- A: The alarm is ringing.
- W: Dr. Watson is calling.
- G: Mrs. Gibbon is calling.
- E: Earthquake is happening.
- R: A report of an earthquake is on the radio.

All of these are Boolean random variables, so there are two possible outcomes for each. With more practice, you'll start to get a sense of what are the important elements to model in a particular story.

Problem: How many probabilities are there in the joint probability distribution?

Solution: To count this, we have to count how many combinations of values can we get with 6 Boolean random variables. This is simply 2^6 which is 64.

To describe this story using the joint probability distribution, we need to specify at least 63 numbers since the last probability can be 1 minus the sum of the first 63 probabilities.

Later on, we will construct a Bayesian network to describe the story. You will see that the Bayesian network representation requires fewer probabilities to describe this story.

3 A Review of Probability Theory

The joint probability distribution is something really nice to have because when we have it, we have all the information that we need to calculate the probability that we're interested in, over any subset of variables. Unfortunately, in practice, we usually do not have the joint distribution, because it's very expensive to create.

In theory, it's nice to have a small example where we can look at the joint distribution to discuss different ways of manipulating probabilities. In this section, we're going to look at four different rules that we can use to calculate probabilities: **the sum rule**, **the product rule**, **the chain rule**, and **the Bayes' rule**.

We are going to use the Holmes scenario, but to simplify, we are only going to consider the following three random variables:

- A: The alarm is ringing.
- W: Dr. Watson is calling.
- G: Mrs. Gibbon is calling.

The following table represents the joint probability distribution over these three variables:

A			$\neg A$		
	G	$\neg G$		G	$\neg G$
W	0.032	0.048	W	0.036	0.324
$\neg W$	0.008	0.012	$\neg W$	0.054	0.486

How do we interpret this table?

For example, the number 0.054 is the probability that the alarm is not ringing, Mrs. Gibbon is calling, and Dr. Watson is not calling. You can interpret each of the numbers in the table with similar reasoning.

You should check that all of the numbers in the probability distribution adds up to 1 to confirm that the probability distribution is valid.

3.1 The Sum Rule

Given a joint probability distribution, how do we compute the probability over a subset of the variables? The rule that we can use here is called **the sum rule**. The sum rule says that if we only care about a subset of the variables, we can simply sum out every variable we do not care about.

What does "sum out" mean here?

Example: Suppose that we have a joint distribution over three variables: A, B, and C. These variables are all Boolean random variables. Suppose that we want to calculate the probability that A is true and B is true. Notice here that we're using the shorthand notation where just writing A means A is true and just writing B means B is true. We can calculate $P(A \land B)$ by summing out C. Here is the expression that we can use:

$$P(A \land B) = P(A \land B \land C) + P(A \land B \land \neg C)$$

The probability of A and B is equal to the probability that A, B, and C are all true, plus the probability that A and B are true and C is false. We vary the values of C and add up all the terms of all possible values for C; the values of A and B are fixed since we care about both A and B being true.

Example: Suppose we have the same joint distribution as the previous example, but we only care about A being true, and we don't care about B and C. We need to sum out both B and C. For these who Boolean variables, we have four possible combinations of values.

$$P(A) = P(A \land B \land C) + P(A \land B \land \neg C) + P(A \land \neg B \land C) + P(A \land \neg B \land \neg C)$$

In these four terms, we are fixing the value of A to always be true, but we're varying the values of B and C, and considering all four possibilities. The possibilities are:

- B and C being true
- B being true and C being false
- B being false and C being true
- B and C being false

To summarize, how do we calculate the probability over a subset of the variables using the sum rule?

We need to add up a bunch of probabilities. In all of these probabilities, for those variables that we care about, we are going to fix their values to whatever values that we care about, and for those variables that we do not care about, we are going to vary their values.

In particular, we will consider all possible combinations of values for the variables that we do not care about.

Problem: What is the probability that **the alarm is NOT ringing** and **Dr. Watson is calling**?

(A) 0.36
(B) 0.46
(C) 0.56
(D) 0.66
(E) 0.76

A			$\neg A$		
	G	$\neg G$		G	$\neg G$
W	0.032	0.048	W	0.036	0.324
$\neg W$	0.008	0.012	$\neg W$	0.054	0.486

Solution: We are asked to find $P(\neg A \land W)$. $\neg A$ is on the right side of the table, and W is the top row of the table. Considering both of these, we're looking at two numbers: 0.036 and 0.324. Intuitively, all we need to do is sum up these two numbers:

$$P(\neg A \land W) = P(\neg A \land W \land G) + P(\neg A \land W \land \neg G) = 0.036 + 0.024 = 0.36$$

The correct answer is (A) 0.36.

Problem: What is the probability that **the alarm is ringing** and **Mrs. Gibbon is NOT calling**?

(A) 0.05
(B) 0.06
(C) 0.07
(D) 0.08
(E) 0.09

A			$\neg A$		
	G	$\neg G$		G	$\neg G$
W	0.032	0.048	W	0.036	0.324
$\neg W$	0.008	0.012	$\neg W$	0.054	0.486

Solution: This problem is conceptually the same as the previous problem. We are asked to find $P(A \wedge \neg G)$. A is on the left side of the table, and $\neg G$ is the second column of the second row. Considering both of these, we're looking at two numbers: 0.048 and 0.012. Same as before, we sum up these two numbers:

$$P(A \wedge \neg G) = P(A \wedge \neg G \wedge W) + P(A \wedge \neg G \wedge \neg W) = 0.048 + 0.012 = 0.06$$

The correct answer is (B) 0.06.

Problem: What is the probability that **the alarm is NOT ringing**?

(A) 0.1
(B) 0.3
(C) 0.5
(D) 0.7
(E) 0.9

A			$\neg A$		
	G	$\neg G$		G	$\neg G$
W	0.032	0.048	W	0.036	0.324
$\neg W$	0.008	0.012	$\neg W$	0.054	0.486

Solution: We are asked to find $P(\neg A)$. In this case, we're summing out both W and G. In terms of the table, we're looking at the whole right side of the table, and we're looking at four numbers: 0.036, 0.324, 0.054, and 0.486. We need to sum up all four numbers:

$$\begin{aligned} P(\neg A) &= P(\neg A \wedge W \wedge G) + P(\neg A \wedge W \wedge \neg G) \\ &+ P(\neg A \wedge \neg W \wedge G) + P(\neg A \wedge \neg W \wedge \neg G) \\ &= 0.036 + 0.324 + 0.054 + 0.486 \\ &= 0.9 \end{aligned}$$

The correct answer is (E) 0.9.

3.2 The Product Rule

One way we can use the product rule is to calculate a conditional probability. Conditional probabilities are about the probability of one variable, say A, conditioned on knowing the value of another variable, say B.

Example: Suppose that we have a joint distribution over A, B, and C again, and we want to calculate $P(A|B)$. We can do this using a form of the product rule:

$$P(A|B) = \frac{P(A \wedge B)}{P(B)}.$$

The formula above only applies to the cases where both A and B are true, because we only care about A and B both being true. You can derive similar formulas when one of them is false, or both of them are false.

You may look at this formula and think that this doesn't look like a product, yet it's called the product rule. This is because you can rewrite this formula in a different form which looks more like a product by taking the denominator and moving it to the left, as shown here:

$$P(A \wedge B) = P(A|B)P(B)$$

Notice that in order to calculate the conditional probability, we need the joint probability over A and B, $P(A \wedge B)$, and we also need the prior or unconditional probability over B, $P(B)$ – both are not given directly but can be calculated using the sum rule.

Calculating conditional probability is going to be a more complicated than before, since we have to apply the sum rule first to get the components we need, then calculate the fraction.

How can we interpret this fraction? The denominator means that in order to calculate this conditional probability, we only care about those worlds in which B is true. Dividing by the denominator means we're ruling out all possible worlds where B is false. The numerator means that within all of the worlds in which B is true, we want only those worlds in which A is true as well. The worlds we actually care about have the property that A and B are both true, $P(A \wedge B)$.

Problem: What is the probability that **Dr. Watson is calling** given that **the alarm is NOT ringing**?

(A) 0.2
(B) 0.4
(C) 0.6
(D) 0.8
(E) 1.0

A			$\neg A$		
	G	$\neg G$		G	$\neg G$
W	0.032	0.048	W	0.036	0.324
$\neg W$	0.008	0.012	$\neg W$	0.054	0.486

$P(\neg A \wedge W) = 0.36$,
$P(A \wedge \neg G) = 0.06$,
$P(\neg A) = 0.9$.

Solution: We are asked to find $P(W|\neg A)$. Using the product rule,

$$P(W|\neg A) = \frac{P(W \wedge \neg A)}{P(\neg A)}$$

Normally, we would have to go back to the joint distribution to calculate both the numerator and the denominator, but since the probabilities are already provided, we can use them directly.

$$P(W|\neg A) = \frac{P(W \wedge \neg A)}{P(\neg A)} = \frac{0.36}{0.9} = 0.4$$

The correct answer is (B) 0.4.

Problem: What is the probability that **Mrs. Gibbon is NOT calling** given that **the alarm is ringing**?

(A) 0.2
(B) 0.4
(C) 0.6
(D) 0.8
(E) 1.0

A			$\neg A$		
	G	$\neg G$		G	$\neg G$
W	0.032	0.048	W	0.036	0.324
$\neg W$	0.008	0.012	$\neg W$	0.054	0.486

$P(\neg A \wedge W) = 0.36$,
$P(A \wedge \neg G) = 0.06$,
$P(\neg A) = 0.9$.

Solution: This problem is similar to the last problem. We are asked to find $P(\neg G|A)$. Using the product rule,

$$P(\neg G|A) = \frac{P(\neg G \wedge A)}{P(A)}$$

We have $P(A \wedge \neg G) = 0.06$, but we don't directly have $P(A)$. $P(A)$ is easy to calculate from the information we have, since $P(A) = 1 - P(\neg A)$, which is 0.1. Now that we have both the numerator and denominator, we can plug them into the formula.

$$P(\neg G|A) = \frac{P(\neg G \wedge A)}{P(A)} = \frac{0.06}{1 - 0.9} = \frac{0.06}{0.1} = 0.6$$

The correct answer is (C) 0.6.

3.3 The Chain Rule

Here's an example of some prior and conditional probabilities we can use to describe our Holmes scenario. Remember, we've simplified it so that we're only thinking about three variables: alarm, Watson, and Gibbon.

The prior probabilities:

$P(A) = 0.1$

The conditional probabilities:

$P(W|A) = 0.9$

$P(W|\neg A) = 0.4$

$P(W|A \wedge G) = 0.9$

$P(W|A \wedge \neg G) = 0.9$

$P(W|\neg A \wedge G) = 0.4$

$P(W|\neg A \wedge \neg G) = 0.4$

$P(G|A) = 0.3$

$P(G|\neg A) = 0.1$

$P(G|A \wedge W) = 0.3$

$P(G|A \wedge \neg W) = 0.3$

$P(G|\neg A \wedge W) = 0.1$

$P(G|\neg A \wedge \neg W) = 0.1$

These numbers are carefully constructed with an underlying model in mind, but it may not be apparent at this moment. We will talk more about it when we talk about Bayesian networks.

The chain rule is best explained through examples. We have three examples here: one for two variables, one for three variables, and the general case for any number of variables.

The two-variable case looks exactly like the product rule that we saw earlier:

$$P(A \wedge B) = P(A|B) \cdot P(B)$$

They are in fact equivalent for the simple two variable case. A and B can also be ordered differently to get an equivalent version of the formula:

$$P(A \wedge B) = P(A|B) \cdot P(B) = P(B|A) \cdot P(A)$$

For the three-variable version, we can begin to see the general pattern for how the chain rule works:

$$P(A \wedge B \wedge C) = P(A|B \wedge C) \cdot P(B|C) \cdot P(C)$$

We are taking these the three variables, A, B, and C, and ordering them in some way. You can think about coming up with the expression backwards, where we have the last variable in the order, just the prior or unconditional probability of C, $P(C)$, and then the second-to-last variable, we have the conditional probability of B given C, $P(B|C)$, and finally the first variable, we have the conditional probability of A given B and C, $P(A|B \wedge C)$.

If you try to extract the general pattern, the pattern seems like every variable in the sequence is going to depend on all of the variables that come before in the sequence, given some sort of ordering of all of the variables. This is exactly the general pattern.

For the general case, we have n variables, any you can think about them as ordered from right to left:

$$\begin{aligned}
&P(X_n \wedge X_{n-1} \wedge \cdots \wedge X_2 \wedge X_1) \\
&= \prod_{i=1}^{n} P(X_i | X_{i-1} \wedge \cdots \wedge X_1) \\
&= P(X_n | X_{n-1} \wedge \cdots \wedge X_2 \wedge X_1) \cdot ... \cdot P(X_2|X_1) \cdot P(X_1)
\end{aligned}$$

Given this ordering, we have a giant product. We can decompose this joint probability into a product of prior and conditional probabilities, which is shown on the second line above. The third line above shows what it would look like if the entire expression was written out.

Notice that there are many versions of this expression which depends on how you've ordered the variables.

Problem: What is the probability that **the alarm is ringing**, **Dr. Watson is calling** and **Mrs. Gibbon is NOT calling**?

(A) 0.060	$P(A) = 0.1$
(B) 0.061	$P(W|A) = 0.9$
(C) 0.062	$P(W|A \wedge \neg G) = 0.9$
(D) 0.063	$P(G|A) = 0.3$
(E) 0.064	$P(G|A \wedge W) = 0.3$

Solution: We are asked to find $P(A \wedge W \wedge \neg G)$. From the chain rule, depending on the order in which we choose these variables, we may come up with different expressions. Therefore, when we're applying the chain rule, it's really important to come up with an order that works. With 3 variables, we have $3 \cdot 2 \cdot 1 = 6$ possible different permutations of the variables.

We need the right information for the ordering that we picked. We only have one prior probability: $P(A)$. This means for the ordering, we need to choose A to be the first one. In terms of conditional probability, we have all sorts of possibilities. We have $P(W|A)$, $P(G|A)$, $P(W|A \wedge G)$, and $P(G|A \wedge W)$. The order of G a W does not really matter, but let's try both orderings.

The first ordering is A first, then W, then $\neg G$. We have $P(A)$ first. The second one is $P(W|A)$. The third one is $P(\neg G|A \wedge W)$. We have $P(A) = 0.1$ and $P(W|A) = 0.9$, but we don't have $P(\neg G|A \wedge W)$. However, we do have $P(G|A \wedge W)$, and can use probabilities to find $P(\neg G|A \wedge W)$ from this.

If you remember probabilities, you may realize that $P(G|A \wedge W) + P(\neg G|A \wedge W) = 1$. This is a rule of conditional probabilities. Among all worlds where A and W are both true, there are only two possibilities. One possibility is that G is true, and the other possibility is that G is false. If we add these two probabilities, they must sum to one.

Given this, we can derive the probability of $P(\neg G|A \wedge W)$ as $1 - (G|A \wedge W)$. This leads us to the final result:

$$\begin{aligned} P(A \wedge W \wedge \neg G) &= P(A)P(W|A)P(\neg G|A \wedge W) \\ &= 0.1 \cdot 0.9 \cdot (1 - 0.3) \\ &= 0.063 \end{aligned}$$

The other ordering is A first, then $\neg G$, then W. We have $P(A)$ first. The second one is $P(\neg G|A)$. The third one is $P(W|A \wedge \neg G)$. We have $P(A) = 0.1$ and $P(W|A \wedge \neg G) = 0.9$, but we don't have $P(\neg G|A)$. Similar to above, we can use $P(G|A) = 0.3$ to derive $P(\neg G|A) = 1 - P(G|A)$. This leads us to the final result:

$$\begin{aligned} P(A \wedge W \wedge \neg G) &= P(A)P(\neg G|A)P(W|A \wedge \neg G) \\ &= 0.1 \cdot (1 - 0.3) \cdot 0.9 \\ &= 0.063 \end{aligned}$$

The correct answer is (D) 0.063.

Problem: What is the probability that **the alarm is NOT ringing**, **Dr. Watson is NOT calling** and **Mrs. Gibbon is NOT calling**?

(A) 0.486	$P(A) = 0.1$
(B) 0.586	$P(W|\neg A) = 0.4$
(C) 0.686	$P(W|\neg A \wedge \neg G) = 0.4$
(D) 0.786	$P(G|\neg A) = 0.1$
(E) 0.886	$P(G|\neg A \wedge \neg W) = 0.1$

Solution: This questions is conceptually exactly the same as the previous question, except we have to do a little bit more calculation. Similar to the previous question, there are two possible orderings that we can choose.

The first ordering is $\neg A$ first, then $\neg W$, then $\neg G$. This leads us to the final result:

$$\begin{aligned}P(\neg A \land \neg W \land \neg G) &= P(\neg A)P(\neg W|\neg A)P(\neg G|\neg A \land \neg W)\\ &= (1-0.1)\cdot(1-0.4)\cdot(1-0.1)\\ &= 0.486\end{aligned}$$

The other ordering is $\neg A$ first, then $\neg G$, then $\neg W$. This leads us to the final result:

$$\begin{aligned}P(\neg A \land \neg W \land \neg G) &= P(\neg A)P(\neg G|\neg A)P(\neg W|\neg A \land \neg G)\\ &= (1-0.1)\cdot(1-0.1)\cdot(1-0.4)\\ &= 0.486\end{aligned}$$

The correct answer is (A) 0.486.

3.4 The Bayes' Rule

How do we calculate a conditional probability? In particular, how do we flip a conditional probability? Why would we want to flip a conditional probability? This turns out to be quite a common scenario.

Think about this in real life. Often, we have some knowledge that one thing could cause another. We have some knowledge about the possible causes for a particular phenomenon.

Example: We would know a particular disease is going to cause some particular symptom on a person:

$$P(\text{symptom} \mid \text{disease})$$

Example: We would know that if there's a fire, then the alarm is going to ring:

$$P(\text{alarm} \mid \text{fire})$$

It's great to have this kind of knowledge, but they are often not helpful in practice when we're asking questions. We often do not know whether a person has a disease, or we do not directly observe whether there's a fire or not. Instead, we often observe the evidence, or the symptom.

In the case of medical diagnoses, we often observe a symptom a person has, and then we ask how likely does the person have a particular disease based on the symptoms observed.

In the case of fires, we often observe whether the alarm is ringing, and we want to infer whether there is a fire somewhere.

The knowledge that we have ($P(\text{symptom} \mid \text{disease})$ and $P(\text{alarm} \mid \text{fire})$) does not really match with the kind of reasoning we want to do. We want to be able to flip the conditional probabilities ($P(\text{disease} \mid \text{symptom})$ and $P(\text{fire} \mid \text{alarm})$) and calculate the resulting conditional probabilities.

This is the **Bayes' rule**:

$$P(X|Y) = \frac{P(Y|X) \cdot P(X)}{P(Y)}$$

This allows us to take $P(Y|X)$ and flip it to get $P(X|Y)$. To do this, we need to know $P(Y|X)$ and $P(X)$. Notice that we don't directly need to know $P(Y)$, as the numerator gives us enough information to derive it.

You should not memorize this rule. At any time you should be able to derive it. Here is how you can derive it:

Remember the two different ways of writing out the product rule for two variables. If we have $P(X \wedge Y)$, we can write this out in two ways:

$$P(X \wedge Y) = P(X|Y) \cdot P(Y)$$

where we order Y first, and then X, or

$$P(X \wedge Y) = P(Y|X) \cdot P(X)$$

where we order X first, and then Y, so

$$P(X \wedge Y) = P(X|Y) \cdot P(Y) = P(Y|X) \cdot P(X)$$

Moving the $P(Y)$ in the first expression to the right gives us our version of the Bayes' rule.

$$\begin{aligned} P(X|Y) \cdot P(Y) &= P(Y|X) \cdot P(X) \\ P(X|Y) &= \frac{P(Y|X) \cdot P(X)}{P(Y)} \end{aligned}$$

In order to calculate the probability of X given Y, it's not necessary to directly know the probability of Y. To understand this, some mathematical derivation must be done.

We're going to take the expression of the Bayes' rule and expand the denominator. We will take the denominator and apply the sum rule in reverse. So we expand it by adding X into the expression.

$$P(Y) = P(Y \wedge X) + P(Y \wedge \neg X)$$

Inserting this into the Bayes' rule, we get

$$P(X|Y) = \frac{P(Y|X)P(X)}{P(Y \wedge X) + P(Y \wedge \neg X)}$$

Next, we expand the two terms in the denominator further by using the product rule.

$$\frac{P(Y|X)P(X)}{P(Y \land X) + P(Y \land \neg X)} = \frac{P(Y|X)P(X)}{P(Y|X)P(X) + P(Y|\neg X)P(\neg X)}$$

Notice that the first term in the denominator is the same as the numerator.

Using this formula, we end up calculating a probability distribution, so this equation is only showing one probability in the distribution. The other probability is $P(\neg X|Y)$. We can do a similar derivation, which would result in this formula:

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

Notice that the denominator for the expanded $P(X|Y)$ and $P(\neg X|Y)$ are the same and the numerator of $P(X|Y)$ ends up being the second term in the denominator for $P(\neg X|Y)$.

Let's look at what's happening with these two expressions we just derived. We are trying to calculate the distribution. One probability in the distribution is $P(X|Y)$, and the other one is $P(\neg X|Y)$. This distribution asks: what is the probability that X is true versus the probability that X is false, among all the worlds in which Y is true?

An alternative way to think about it is how we calculate the distribution. We can just calculate the two numerators and normalize them, which means taking the two numbers and dividing each of them by the sum of them. This is what is happening when we divide both of them with the denominator. If you think about this way of calculating this distribution, you will realize that $P(Y)$ is simply a normalization constant that we're using here.

A *normalization* constant basically means we take a bunch of numbers that may not make up a probability distribution since they may not sum up to one, and divide each number by the sum of them to convert it into a valid probability distribution, with the numbers summing up to one. This derivation gives us a new way of thinking about the Bayes' rule, and also gives us a new way of calculating the distribution.

This new procedure has two steps:

1. Calculate $P(Y|X)P(X)$ and $P(Y|\neg X)P(\neg X)$.
2. Normalize these two values so that they sum up to 1.

From this description, we can see what is the minimum set of values we need to calculate this distribution. We do not need to know $P(X)$ to start with. We need to know $P(X)$, $P(Y|X)$, and $P(Y|\neg X)$. With these three quantities, we have enough information to calculate $P(X|Y)$ and $P(\neg X|Y)$. We can derive $P(\neg X)$ from $P(\neg X) = 1 - P(X)$.

Problem: What is the probability that **the alarm is NOT ringing** given that **Dr. Watson is calling**?

(A) 0.6
(B) 0.7
(C) 0.8
(D) 0.9
(E) 1.0

$P(A) = 0.1$
$P(W|A) = 0.9$
$P(W|\neg A) = 0.4$

Solution: We are asked to find $P(\neg A|W)$. We can derive the correct answer in two ways.
The first way is to apply the Bayes' rule straight away. By applying the Bayes' rule, the numerator will have the conditional probability flipped, $P(W|\neg A)$, and $P(\neg A)$, and the denominator will be $P(W)$. We have $P(W|\neg A) = 0.4$ directly. We have $P(\neg A) = 1 - P(A) = 1 - 0.1 = 0.9$ indirectly.

We need to calculate the probability of W by applying the sum rule in reverse and then applying the product rule. The combination of those two steps will give you

$$\begin{aligned} P(W) &= P(W|A)P(A) + P(W|\neg A)P(\neg A) \\ &= 0.9 \cdot 0.1 + 0.4 \cdot 0.9 \\ &= 0.09 + 0.46 \\ &= 0.45 \end{aligned}$$

This leads us to the final result:

$$P(\neg A|W) = \frac{P(W|\neg A)P(\neg A)}{P(W)} = \frac{0.4 \cdot (1 - 0.1)}{0.45} = 0.8$$

It takes some effort to calculate the denominator which is important for applying the Bayes' rule. The other approach shortens the process by calculating in a different way. For the other approach, we are calculating both $P(\neg A|W)$ and $P(A|W)$. In order to calculate these probabilities, we only need to calculate the corresponding numerators to the probability in the Bayes' rule equation. This leads us to get

$$P(W|A)P(A) = 0.9 \cdot 0.1 = 0.09$$

and

$$P(W|\neg A)P(\neg A) = 0.4 \cdot 0.9 = 0.36.$$

The second step is to normalize both values by calculating their sum and then dividing each value by their sum. First we calculate their sum,

$$0.09 + 0.36 = 0.45$$

then we take each value and divide it by the sum that we just calculated:

$$P(A|W) = \frac{0.09}{0.45} = 0.2$$

$$P(\neg A|W) = \frac{0.36}{0.45} = 0.8$$

You can see the effect of this, since previously these two values, 0.09 and 0.36 did not sum up 1. After normalization, we get 0.2 and 0.8 which sum up to 1.

The correct answer is (C) 0.8.

Problem: What is the probability that **the alarm is ringing** given that **Mrs. Gibbon is NOT calling**?

(A) 0.04
(B) 0.05
(C) 0.06
(D) 0.07
(E) 0.08

$P(A) = 0.1$
$P(G|A) = 0.3$
$P(G|\neg A) = 0.1$

Solution: This problem is conceptually exactly the same as the previous question. We are asked to find $P(A|\neg G)$. Directly applying the Bayes' rule leads us to the following result:

$$\begin{aligned}
P(A|\neg G) &= \frac{P(\neg G|A)P(A)}{P(\neg G|A)P(A) + P(\neg G|\neg A)P(\neg A)} \\
&= \frac{(1-0.3)\cdot 0.1}{(1-0.3)\cdot 0.1 + (1-0.1)\cdot(1-0.1)} \\
&= \frac{0.07}{0.08} \\
&= 0.08
\end{aligned}$$

For the other approach, we follow the steps similar to the ones followed in the previous problem which leads us to the following result.

$$\begin{aligned}
P(\neg G|A)P(A) &= (1-0.3)\cdot 0.1 = 0.7 \\
P(\neg G|\neg A)P(\neg A) &= (1-0.1)\cdot(1-0.1) = 0.81 \\
0.07 + 0.81 &= 0.88 \\
P(A|\neg G) &= \frac{0.07}{0.88} \\
&= 0.08
\end{aligned}$$

The correct answer is (E) 0.08.