Combinatorics and Probability - Introduction

Introduction

In this section, you'll learn about the foundation of statistics: probability!

Combinatorics and Probability

In this section, we'll take a little time to "get our math on" with some basic probability. You're going to start with some basic set theory and look at how to operate on related sets using Python.

From there, we're going to use what we learned about sets to start to learn and apply some of the basic rules of probability.

Factorials and Permutations

Next we're going to dig into factorials, and how they can be used to calculate various permutations.

Combinations

We're then going to examine the difference between permutations and combinations. We'll get some practice calculating combinations for everything from drawing letters from a bag to creating soccer teams for a tournament!

Conditional Probability

We start the section off with an introduction to conditional probability. We look at the difference between dependent and independent events, look at how to calculate dependent probabilities, and then introduce some key theorems related to conditional probabilities: the product rule, the chain rule, and Bayes theorem.

Partitioning and the Law of Total Probabilities

From there, we introduce the concept of partitioning a sample space, explain the law of total probabilities, and then introduce the idea of conditional independence.

Summary

This is a mathematics heavy section. Some of the discrete problems you'll solve may not seem to be particularly relevant to machine learning, but we deliberately introduce them so that you have the foundations required to make thoughtful choices as we introduce you to a range of new machine learning models in the later sections.

What is Probability?

Introduction

As an aspiring data scientist, it's important to know the foundations of probability and combinatorics, as these areas form the backbone of many concepts in data science. In the following lessons and labs, you'll get a gentle introduction to several concepts that are related to probability, such as sets, combinations, and permutations.

Objectives

You will be able to:

- Define probability

What is probability, and how does it relate to data science?

Probability is the chance that a certain event will happen, in other words,how "likely" it is that some event will happen.

As data science often uses *statistical inference* to analyze or predict certain events or trends, knowing probability and its applications is important because statistical inference uses probability distributions of the data. Although it might take a little more time for you to understand just how important the foundations of probability are for data science, by the end of the first part of the probability section, you'll be able to answer questions like:

- How likely is it to end up with heads when flipping a coin once? (the answer here is 50% - not very surprising)
- How likely is it to end up with exactly 2 x heads and 3 x tails when flipping a coin 5 times?
- How likely is it to throw tails first, then heads, then tails, then heads, then tails when flipping a coin 5 times?
- If you throw 5 dice, what is the probability of throwing a "full house"?
- What is the probability of drawing 2 consecutive aces from a standard deck of cards?

Summary

Now, let's dive deeper into the understanding of sets. Getting these concepts will make calculating your first probabilities much easier!

Introduction to Sets

Introduction

You have definitely heard of sets before. In this section, however, you will learn about the formal definition of sets, which will serve as a foundation for everything related to probability and combinatorics!

Objectives

You will be able to:

- Define a set in the context of probability theory
- Define a universal set and subsets
- Describe the process of making unions, intersections, and complements
- Use Venn Diagrams to visually demonstrate set operations
- Describe the inclusion-exclusion principle

What is a Set?

In probability theory, a set is defined as a *well-defined collection of objects*.

Mathematically, you can denote a set by S. If an element x belongs to a set S, then you'd write $x \in S$. On the other hand, if x does not belong to a set S, then you'd write $x \notin S$.

Example: If S is defined as the set of even numbers, then:

- If $x = 2$, $x \in S$ because x is an even number.
- If $x = 9$, $x \notin S$ because x is not an even number.

Subsets

Set T is a subset of set S if *every element* in set T is also in set S. The mathematical notation for a subset is $T \subseteq S$.

Typically, you'll be more interested in *proper subsets*. All proper subsets are subsets. The only difference between subsets and proper subsets is that a subset can technically be the entire set. In other words, if A = {1,2,3} and B = {1,2,3} A is subset of B. If C = {1,2} then C is both a subset and proper subset of A. C is also a subset and proper subset of B. The mathematical notation for proper subsets is : $C \subset A$

Example: If S is the set of even numbers, set $T = \{2, 6, 22\}$ is a proper subset of S. Formally, you can write this as $T \subset S$. $T \subseteq S$ is also correct in this case!

Universal Sets

The collection of all possible outcomes in a certain context or universe is called the **universal set**. A universal set is often denoted by Ω.

Example of a universal set: All the possible outcomes when rolling a dice.

$\Omega = \{1, 2, 3, 4, 5, 6\}$

Remember that a universal set is not necessarily all the possible things that have ever existed. Typically, a universal set is just all the possible elements within certain bounds, e.g., the set of all countries in the world, the set of all the animal species in the Bronx Zoo, etc.

A universal set can have an infinite number of elements, for example, the set of all real numbers!

Elementary Set Operations

Next, let's talk about set operations. Imagine you have two sets of numbers, say the first 4 multiples of 3 in set S:

$S = \{3, 6, 9, 12\}$

and the first 4 multiples of 2 in set T:

$T = \{2, 4, 6, 8\}$

$T = \{2, 4, 6, 8\}$.

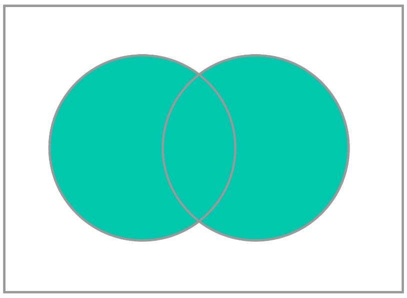
a) Union of Two Sets

The union of two sets S and T is the set of elements of either S or T, or in both.

Applied to our example, the union of S and T is given by the elements $\{2, 3, 4, 6, 8, 9, 12\}$.

In mathematical terms, the union of S and T is denoted as $S \cup T$.

A popular way to represent sets and their relationships is through Venn Diagrams, (https://en.wikipedia.org/wiki/Venn_diagram), see picture below!

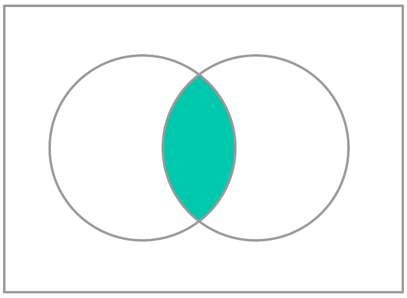


b) Intersection of Two Sets

The intersection of two sets S and T is the set that contains all elements of S that also belong to T.

Applied to our example, the intersection of S and T is given by {6}, so it contains the elements that are multiples of both 2 AND 3.

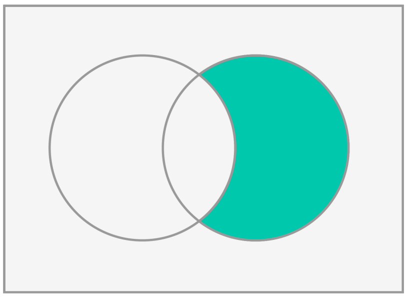
In mathematical terms, the intersection of S and T is denoted as $S \cap T$.



c) Relative Complement or the Difference

If you have S and T, the relative complement of S contains all the elements of T that are NOT in S. This is also sometimes referred to as the *difference*. The difference is denoted by $T \backslash S$ or $T - S$.

In this case, the relative complement of S (or $T \backslash S$) is $\{2, 4, 8\}$. The relative complement of T (or $S \backslash T$) is $\{3, 9, 12\}$.



d) Absolute Complement

There is another definition of the complement when considering universal sets Ω as well. In this context, we're talking about the *absolute complement*.

The absolute complement of S, with respect to the Universal set Ω, is the collection of the objects in Ω that don't belong to S.

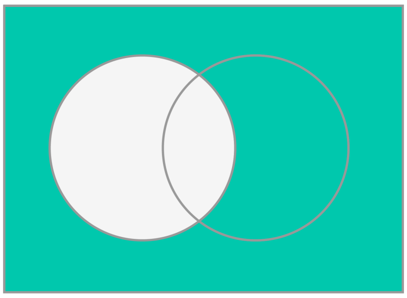
Note how the definition of Ω is very important here. Imagine a set $S = \{\text{elephant, alligator, tiger, bear}\}$. The complement of this set will depend on how the universal set is defined: Is Ω equal to *the animals in the Bronx Zoo*, or *the 20 most deadly animals in the world*?

Mathematically, the absolute complement of S is denoted as S' or S^c.

Let's reconsider S and T as defined previously.

Let's define Ω, the universal set (denoted by the box around the two Venn diagrams), as the set that contains the multiples of both 2 and 3 until 20. Then the elements of Ω are $\{2, 3, 4, 6, 8, 9, 10, 12, 14, 15, 16, 18, 20\}$.

The absolute complement of S (so, S' or S^c) is then given by $\{2, 4, 8, 10, 14, 15, 16, 18, 20\}$.



Inclusion-Exclusion Principle

Note that if you want to know how many elements are in set S versus T, you can't simply sum up the elements, because they have elements in common.

In combinational mathematics, the inclusion-exclusion principle is a counting technique that solves this problem.

When having two sets, the method for counting the number of elements in the union of two finite sets is given by:

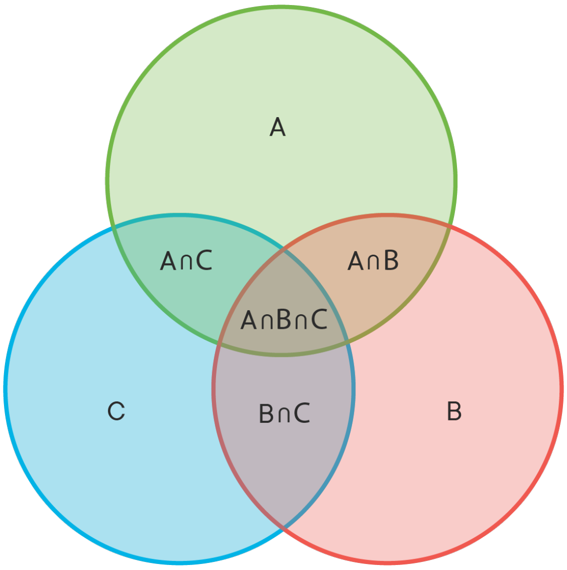
$|S \cup T| = |S| + |T| - |S \cap T|$,

where the horizontal lines denote the *cardinality* of a set, which is the number of elements in the set, considering a set with a finite number of elements.

The formula expresses the fact that the sum of the sizes of the two sets may be too large since some elements may be counted twice. For the double-counted elements, one is subtracted again.

This formula can be extended to three sets, four sets, etc. For example, imagine you have a third set R. The number of elements in the union of three finite sets is given by:

$|S \cup T \cup R| = |S| + |T| + |R| - |S \cap T| - |S \cap R| - |R \cap T| + |S \cap T \cap R|$



Empty Sets

When there are no elements in a certain set, this set is **empty**, denoted by $\varnothing$ or simply $\{\}$

Sets in Python

Some things to bear in mind when working with sets in Python:

- Sets are unordered collections of unique elements.
- Sets are iterable.
- Sets are collections of lower level python objects (just like lists or dictionaries).

Documentation for sets in Python can be found here: Sets

Sets and Set Operations: A Summative Example

To put this all together, let's consider an example with restaurants:

Think about a *set A* with all the restaurants that serve Italian food. Next, there is a *set B* with all the restaurants that serve burgers.

The **union** of these sets, *set C*, contains the set of restaurants that either serve Italian food, burgers or both.

You could say that the **universal set** here, *set U*, contains all the restaurants in the world (with any type of food). Then *set C* is a **subset** of *set U*.

The **intersection** of *A* and *B* contains the restaurants that serve *both* Italian food and burgers.

The **relative complement** of *set A* contains the restaurants that serve burgers but *not* Italian food.

The **absolute complement** of *set A* contains the restaurants that serve *any food* but *no* Italian food.

Summary

In this section, you learned about sets, subsets, and universal sets. Next, you were introduced to some elementary set operations such as unions, intersections, and complements. After that, all this information was tied together through the inclusion-exclusion principle. Next, you saw how sets translate into Python. You'll start exploring this in further detail in the next lab!

Introduction to Probability

Introduction

Now that you understand the basics of sets, you'll learn how this knowledge can be used to calculate your first probabilities! In this section, you'll learn how to use sets to create probabilities and you'll learn about the very foundations of probability through the three probability axioms.

Objectives

You will be able to:

- Compare experiments, outcomes, and the event space
- Calculate probabilities by using relative frequency of outcomes to event space
- Describe the three axioms of probability
- Describe the addition law of probability

Experiment and outcomes

Previously, we defined sets and related concepts. Now let's look at the set

$S = \{1, 2, 3, 4, 5, 6\}$, which contains all possible outcomes when throwing a dice.

When you throw a dice once, you can consider this a *random experiment*. The result of this "experiment" is the *outcome*.

You can then say that:

- S defines all the **possible outcomes** when throwing the dice once
- S is our Universal set Ω, as seen before

When conducting experiments, you say that your universal set is your **sample space**: it is the universe in which your possible outcomes are listed as elements.

Other examples of sample spaces:

- The number of text messages you send each day: in this case, S is equal to some number x, with x being a **positive integer**, or mathematically: $S = \{x \mid x \in \mathbb{Z}, x \geq 0\}$
- The number of hours someone watches TV each day: $S = \{x \mid x \in \mathbb{R}, 0 \leq x \leq 24\}$

Event space

Next, let's define event space. The **event space** is a subset of the sample space, $E \subseteq S$

For example, the event "throwing a number higher than 4" when throwing a dice would result in an event space $E = \{5, 6\}$. Throwing an odd number would lead to an event space $E = \{1, 3, 5\}$.

Summarized, the event space is a collection of events that we *care* about. We say that event E happened if the actual outcome after rolling the dice belongs to the predefined event space E.

With **sample space** and **event space**, you now understand the two foundational concepts of **probability**.

Other examples of event spaces based on previously defined sample spaces:

- If you define that the event "low daily number of text messages sent" means 20 or fewer text messages, the event space is defined as: $E = \{x \mid x \in \mathbb{Z}, 0 \leq x \leq 20\}$
- Binge-watch day: $E = \{x \mid x \in \mathbb{R}, x \geq 6\}$

Introduction to probability

The law of relative frequency

While conducting an endless stream of experiments, the relative frequency by which an event will happen becomes a fixed number.

Let's denote an event by E, and the *probability* of the event E occurring by $P(E)$. Next, let n be the number of conducted experiments, and $S(n)$ the count of "successful" experiments (i.e. the times that event E happened). The formal definition of probability as a relative frequency is given by:

$$P(E) = \lim_{n\to\infty} \frac{S(n)}{n}$$

This is the basis of a frequentist statistical interpretation: an event's probability is the ratio of the positive trials to the total number of trials as we repeat the process infinitely.

For example, the probability of rolling a 5 on a 6 sided dice is the limit of the successes to trials as the number of trials goes to infinity.

Probability axioms

In the early 20th century, Kolmogorov and Von Mises came up with three axioms that further expand on the idea of probability. The three axioms are:

1. Positivity

A probability is always bigger than or equal to 0, or $0 \leq P(E) \leq 1$

2. Probability of a certain event

If the event of interest is the sample space, we say that the outcome is a certain event, or $P(S) = 1$

3. Additivity

The probability of the union of two exclusive events is equal to the sum of the probabilities of the individual events happening.

If $A \cap B = \varnothing$, then $P(A \cup B) = P(A) + P(B)$

Addition law of probability

The additivity axiom is great, but most of the time events are not exclusive. A very important property is the **addition law or probability** or the **sum rule**.

$P(A \cup B) = P(A) + P(B) - P(A \cap B)$

Put in words, the probability that A or B will happen is the sum of the probabilities that A will happen and that B will happen, minus the probability that *both* A and B will happen.

Examples

Let's reconsider the dice example to explain what was explained before:

Additivity of exclusive events

Let's consider two events: event M means throwing a 6, event N means that you throw an odd number $N = 1, 3, 5$. These events are exclusive, and you can use the additivity rule if you want to know the answer to the question:

"what is the probability that your outcome will be a 6, or an odd number?"

$P(M \cup N) = P(M) + P(N) = \frac{1}{6} + \frac{3}{6} = \frac{4}{6}$

Addition law of probability

Now, let's consider the same event $N = 1, 3, 5$ and another event $Q = 4, 5$. These events are *not* mutually exclusive, so if you want to know the probability that N or Q will happen, you need to use the addition law of probability.

Note that $(N \cap Q)$ is equal to getting an outcome of 5, as that is the "common" element in the respective event spaces of N and Q.

This means that $P(N \cap Q) = \frac{1}{6}$

$P(N \cup Q) = P(N) + P(Q) - P(N \cap Q) = \frac{3}{6} + \frac{2}{6} - \frac{1}{6} = \frac{4}{6}$

Final Note

In the previous examples, you noticed that for our dice example, it is easy to use these fairly straightforward probability formulas to calculate probabilities of certain outcomes.

However, if you think about our text message example, things are less straightforward, e.g.:

"What is the probability of sending less than 20 text messages in a day?"

This is where the probability concepts introduced here fall short. The probability of throwing any number between 1 and 6 with a die is always exactly $\frac{1}{6}$, but we can't simply count our messages event space. In words, the probability of sending 20 messages is likely different than the probability of sending, say, 5 messages, and will be different for any number of messages sent. You'll learn about tools to solve problems like these later on.

Summary

Well done! In this section, you learned how to use sets to get to probabilities. You learned about experiments, event spaces, and outcomes. Next, you learned about the law of relative frequency and how it can be used to calculate probabilities, along with the three probability axioms.

Permutations and Factorials

Introduction

In the previous lab, you defined a few sample spaces by counting the total number of possible outcomes. This is not very practical when sample spaces grow. In this lab, you'll be introduced to *permutations*, which will provide a structured way to help you define sample space sizes!

Objectives

You will be able to:

- Describe how factorials are related to permutations
- Mathematically derive how many permutations there are for large sets
- Calculate permutations of a subset
- Calculate permutations with repetition and replacement

Defining the Sample Space by Counting

Let's consider the following example.

The Beyoncé tribute band "The Single Ladies" is playing a free mini gig in your local park next week. They have selected three all-time classics: "Drunk in Love", "Crazy in Love" and "Formation", but still have to decide the order they will play the songs in. Knowing this, how many playlists are possible?

It is easy and fairly quick to write down possible orders here:

"Drunk in Love", "Crazy in Love", "Formation"

"Drunk in Love", "Formation", "Crazy in Love"

"Crazy in Love", "Drunk in Love", "Formation"

"Crazy in Love", "Formation", "Drunk in Love"

"Formation", "Drunk in Love", "Crazy in Love"

"Formation", "Crazy in Love", "Drunk in Love"

That's it! When we count the possible outcomes, we get to 6 elements in the sample set. Now what if "The Single Ladies" plays a setlist of 4 songs? or 5? That's where the notion of *permutations* comes in handy.

Permutations

The problem setting, in general, is that there are n objects and we want to know how many *permutations* are possible.

This is a way how you can tackle this. You're the lead singer and have to decide which song to play first. You have 3 songs to choose from, so 3 ways of choosing a first song. Then, you move on to the second song. You've chosen the first one, so you have 2 songs to choose from now, etc. Mathematically, this boils down to:

$\text{\# Beyoncé permutations} = 3 * 2 * 1 = 3! = 6$

Generalizing this to n, this means that the number of permutations with n distinct objects is $n!$, or the factorial of n.

Permutations of a Subset

Now, let's consider another example. "The Single Ladies" are still playing a concert at central park, but they disagree on the final three songs that they will play. They only get a 12 min gig slot, so they really can't play more than 3, yet they have a shortlist of 8 they need to pick from. How many final song selections are possible given this info? As for the first example, the order of the songs played is still important.

When the band members decide on the first song, they have 8 possible songs to choose from. When choosing the second song, they have 7 to choose from. Then for the third song, they have 6 left.

$\text{\# Beyoncé k-permutations} = 8 * 7 * 6 = 336$

formalizing this, the question is how many ways we can select k elements out of a pool of n objects. The answer is

$n * (n-1) * \ldots * (n-k+1)$ or in other words, $P_k^n = \frac{n!}{(n-k)!}$

This is known as a k-permutation of n.

The idea is here that we only "care" about the order of the first k objects. The order of the other $(n-k)$ objects doesn't matter, hence they're left out of the equation.

Permutations with Replacement

When talking about setlists, it makes total sense to assume that songs will not be played twice. This is not always how it works though. Imagine a bag with three marbles in it: a green one, a red one, and a blue one. Now we'll draw marbles three times in a row, but each time, we'll write down the marble color and *put it back in the bag* before drawing again.

Now the number of possible outcomes is $3 * 3 * 3$.

Generalizing this to n, this means that the number of permutations with replacement when having n distinct objects is equal to n^j where j is the number of "draws".

Permutations with Repetition

When using permutations, some elements may be *repeated*.

A classic example is using permutations on words. Let's say you have the letters of the word "TENNESSEE". How many different words can you create using these letters?

Simply saying that there are 9 letters so the answer is $9!$ does not give you the correct answer. Looking at the word TENNESSEE by itself, you can swap the 3rd and the 4th letter and have the same word. So the total number is less than $9!$.

The solution is to divide $9!$ by the factorials for each letter that is repeated!

The answer here is then (9 letters, 4 x E, 2 x N, 2 x S)

$$\frac{9!}{4!2!2!} = 3780$$

The general formula can be written as:

$$\frac{n!}{n_1!n_2!\ldots n_k!}$$

where n_j stands for identical objects of type j (the distinct letters in our TENNESSEE example).

Level-Up: Factorials and Recursion

At the start of this lesson, when discussing the number of possible permutations we can obtain for n distinct objects, we mentioned the concept of the factorial of n, denoted by $n!$.

In the example presented to you, we wanted to count all possible ways in which three different Beyoncé songs could be played by the Beyoncé tribute band "The Single Ladies". There were 3 possible ways of choosing a first song, 2 possible ways of choosing a second song, and only 1 way of choosing a third and final song, for $3 * 2 * 1 = 6$ different ways in which the three different songs could be played. This number, 6, is equal to the factorial of 3, $3!$, the number of permutations of 3 distinct objects.

Here, $3! = (3 * 2 * 1) = 6$. Notice that this is the same as writing $3 * 2! = 3 * (2 * 1)$ and $3 * 2 * 1! = 3 * 2 * (1)$. (By definition, the factorial of 1, $1!$, is equal to 1. The factorial of 0, $0!$ is also defined to be equal to 1.)

We can generalize this to the case of computing the factorial of an integer n, $n!$. The factorial of n, $n!$, can be written as $n * (n-1)!$, which itself can be written as $n * (n-1) * (n-2)!$. That is, we can define the factorial of n in terms of the product of n and the factorial of $(n-1)$, and the factorial $(n-1)$ can be defined in terms of the product of $(n-1)$ and the factorial of $(n-2)$, and so on and so forth, as seen in the equation below, until we get to $1!$, which is defined to be equal to 1:

$$n! = n * (n-1)! = n * (n-1) * (n-2)! = \ldots = n * (n-1) * (n-2) * \ldots * 2! = n * (n-1) * (n-2) * \ldots * 2 * 1!$$

Recursion

When we define a function in terms of itself, in this case, the factorial of n in terms of the factorial of (n-1), we are using **recursion**. Recursive functions are functions that can call themselves in order to loop until a condition is met. In the next lab, you'll get a glimpse on how to write a recursive function in Python, but in the Appendix to this Module, we go over recursive functions in Python in much more detail.

Summary

Now you're well on your way to calculate all sorts of permutations using factorials - both for understanding the sample space, subsets, etc! Let's move on for some practice!

Combinations

Introduction

In the previous section, you learned about how to apply permutations. Permutations come in handy when we want to know how many ways we can order sets. Now, what if order is not important? That's where *combinations* come in.

Objectives

You will be able to:

- Describe how combinations are used when order is not important

Why combinations?

In some settings, the order of the selection is not important.

Let's go back to our example of a coverband creating a setlist. Imagine that the band is playing 3 songs out of their 8 song repertoire. How many ways can they select songs, assuming that the **order of the chosen songs is not important**? Here, we just want to know *which* three songs they play, and not which song goes first, second and last.

You can use a backward rationale here. You know that when order *did* matter, our answer was $8*7*6$. When having three elements, there are 6 possible orders (ABC, ACB, CAB, CBA, BAC, BCA), so the answer can be obtained by dividing our previous answer by 6.

This type of problem can be solved by using *combinations*. In general, combinations answer the question: "How many ways can we create a subset k out of n objects?". The subset is not ordered.

$$\binom{n}{k} = \frac{P^n_k}{k!} = \frac{\frac{n!}{(n-k)!}}{k!} = \frac{n!}{(n-k)!k!}$$

Applied to our example, this means that there are

$$\frac{8!}{(8-3)!3!} = \frac{8!}{(8-3)!3!} = \frac{8*7*6}{6} = 56$$

.

so there are 56 ways to choose 3 songs out of an 8 song repertoire.

Summary

In this section, you learned what combinations are and how to use them. Let's put this knowledge into practice!

In [ ]:

```
```

Conditional Probability

Introduction

In the previous lessons and labs, you learned about some fundamentals of probability theory, along with basic combinatorics such as permutations and combinations. You'll now extend your knowledge of probability by learning about **Conditional Probability**. You'll see how Conditional Probability is extremely important in Statistics, and the foundation of many applications. Understanding conditional probability is essential when exploring fields in Machine Learning and Artificial Intelligence.

In this lesson, you'll learn about conditional probability, what it is, and how and when to use it. Later on, you'll see how this simple idea becomes a key component in most statistical machine learning algorithms.

Objectives

You will be able to:

- Differentiate between independent and dependent events
- Use the multiplication rule to find the probability of the intersection of two events
- Use conditional probability to explain the Product Rule, Chain Rule, and Bayes Theorem

Events and Sample Space

Before introducing you to specific event types, let's do a quick recap of the notion of event and sample space.

An **event** is the outcome of an experiment, for example, obtaining heads when tossing of a coin or getting 3 after a dice roll. Note: an event can also be a collection of different events grouped together (or a so-called **compound** event), e.g. getting a 3 twice when rolling a dice twice.

A **sample space** is a collection of every single possible outcome in a trial, generally represented by Ω. The sample space for 1 random dice throw is {1,2,3,4,5,6}.

As you remember for the previous lessons, we can combine events and sample space to compute event probability.

You'll learn about 3 important event types: **independent**, **disjoint**, and **dependent** events.

Independent Events

Events A and B are independent when the occurrence of A has no effect on whether B will occur (or not).

Consider the following independent events

- Getting heads after flipping a coin **and** getting a 5 after throwing a fair dice
- Choosing a marble from a container **and** getting heads after flipping a coin

Two independent events

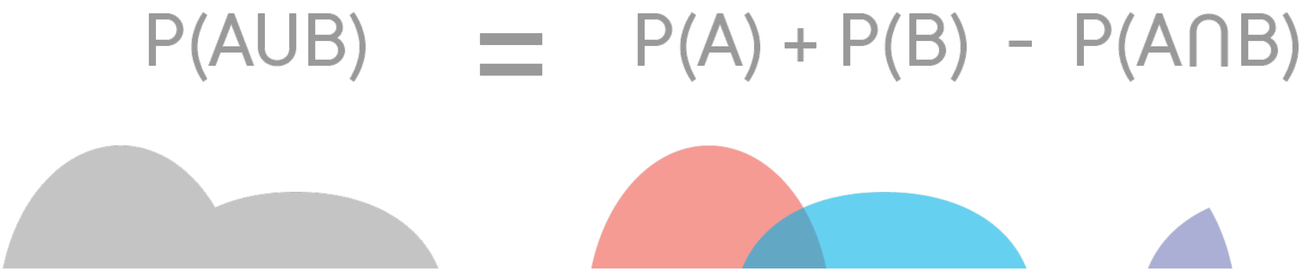
Formally, events A and B are independent if

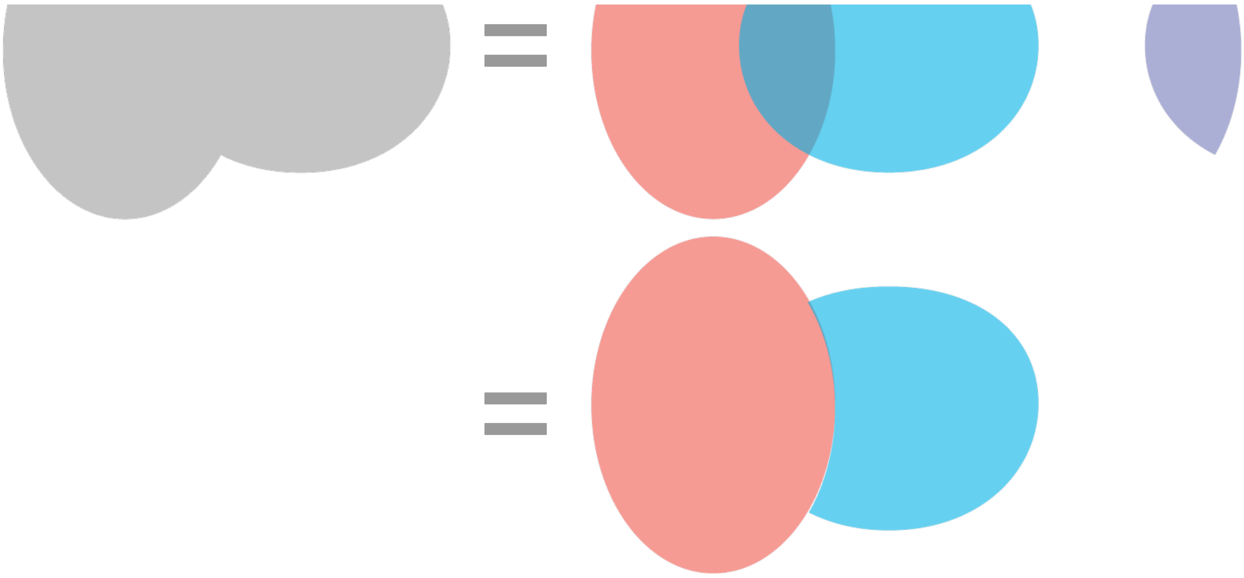
- $P(A \cap B) = P(A)P(B)$

The probability of A or B occurring, $P(A \cup B)$, is given by the addition rule of probability:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B).$$

We subtract the intersection of the two events to avoid over-counting. See the diagram below for some intuition:





Thus, in the case of two independent events, by substitution,

$$P(A \cup B) = P(A) + P(B) - P(A)P(B).$$

Three independent events

Three events A, B and C if

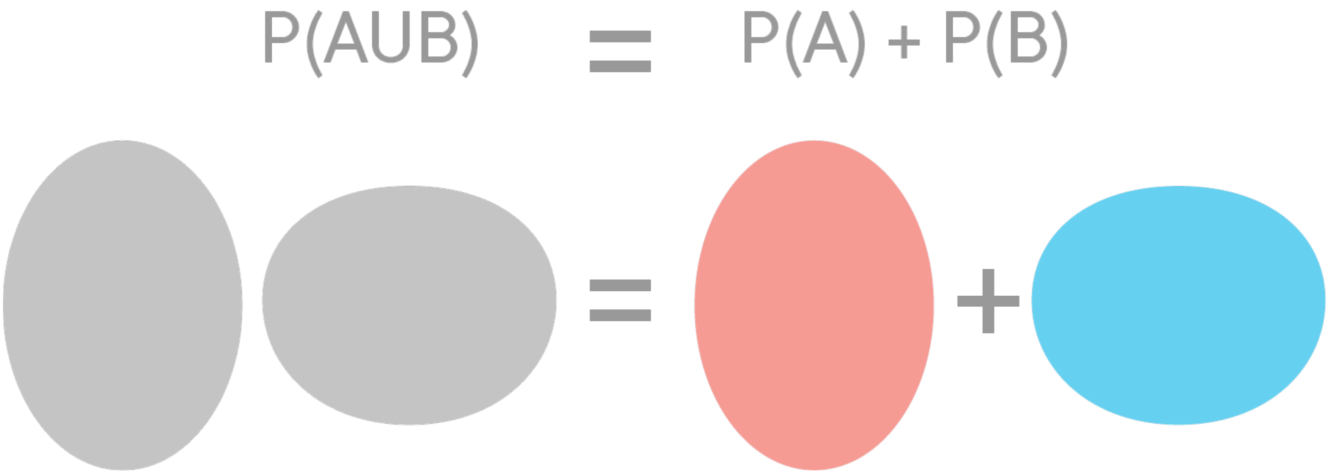
- $P(A \cap B) = P(A)P(B)$
- $P(A \cap C) = P(A)P(C)$
- $P(B \cap C) = P(B)P(C)$
- $P(A \cap B \cap C) = P(A)P(B)P(C)$

So you need both *pairwise independence* and *three-way independence*

Disjoint Events

Events A and B are disjoint if A occurring means that B cannot occur.

Disjoint events are **mutually exclusive**. $P(A \cap B)$ is **empty**.

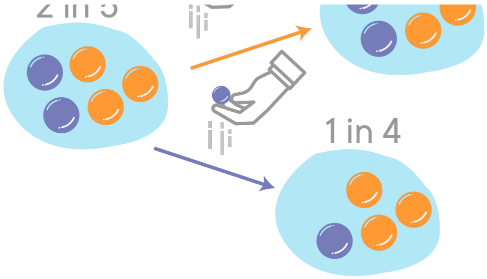


Dependent Events

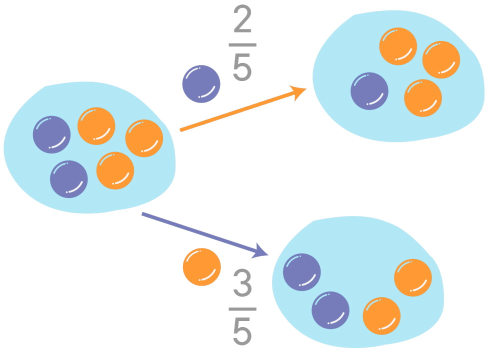
Events A and B are dependent when the occurrence of A somehow has an effect on whether B will occur (or not).

Now things start getting a bit more interesting.

Let's look at an example. Let's say event A is taking an orange or purple marble out of a jar. The jar contains 3 orange and 2 purple marbles.



The probability of getting a purple marble is $\frac{2}{5}$ and getting an orange marble is $\frac{3}{5}$.



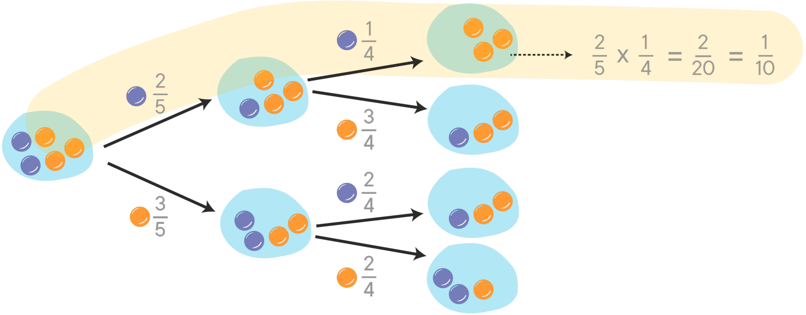
At that point, one marble is taken out and we now take another marble from the jar (event B).

Here you can see that our second event is dependent on the outcome of the first draw.

- If we drew an orange marble first, the probability of getting a purple marble for event B is $\frac{2}{4}$.
- If we saw a purple marble first, however, the probability of seeing a purple in the second trial is $\frac{1}{4}$.

In simple terms, the probability of seeing an event B in the second trial depends on the outcome A of the first trial. We say that $P(B)$ is **conditional** on $P(A)$.

A **tree diagram** can be used to explore all possible events.



Conditional Probability

Conditional probability emerges when the outcome a trial may influence the results of the upcoming trials.

While calculating the probability of the second event (event B) given that the primary event (event A) has just happened, we say that the probability of event B relies on the occurrence of event A.

Here are some more examples:

- Drawing a 2nd Ace from a deck of cards given that the first card you drew was an Ace.
- Finding the probability of liking "The Matrix" given that you know this person likes science fiction.

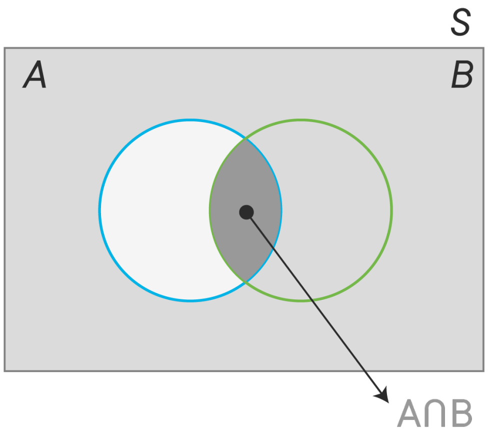
Let's say that $P(A)$ is the event we are interested in, and this event depends on a certain event B that has happened.

The conditional probability (Probability of A **given** B) can be written as:

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$

$P(A \mid B)$ is the probability A **given** that B has just happened

$P(A|B)$, is the probability A **given** that B has just happened.



Understanding this formula may be easier if you look at two simple Venn Diagrams and use the multiplication rule. Here's how to derive this formula:

Step 1: Write out the multiplication rule:

- $P(A \cap B) = P(B) * P(A \mid B)$

Step 2: Divide both sides of the equation by P(B):

- $\frac{P(A \cap B)}{P(B)} = \frac{P(B) * P(A \mid B)}{P(B)}$

Step 3: Cancel P(B) on the right side of the equation:

- $\frac{P(A \cap B)}{P(B)} = P(A \mid B)$

Step 4: This is of course equal to:

- $P(A \mid B) = \frac{P(A \cap B)}{P(B)}$

And this is our conditional probability formula.

There are a few variations and theorems that are related to and/or results of this conditional probability formula. The most important ones are: the **product rule**, the **chain rule** and **Bayes Theorem**

Theorem 1 - Product Rule

The **product rule** was used to derive the conditional probability formula above, but is often used in situations where the conditional probability is easy to compute, but the probability of intersections of events isn't.

The intersection of events A and B can be given by

$$P(A \cap B) = P(B)P(A \mid B) = P(A)P(B \mid A)$$

Remember that if A and B are independent, then conditioning on B means nothing (and vice-versa) so $P(A|B) = P(A)$, and $P(A \cap B) = P(A)P(B)$.

Theorem 2 - Chain Rule

The **chain rule** (also called the **general product rule**) permits the calculation of any member of the joint distribution of a set of random variables using only conditional probabilities.

Recall the product rule:

$P(A \cap B) = P(A \mid B)P(B)$

When you extend this for three variables:

$P(A \cap B \cap C) = P(A \cap (B \cap C)) = P(A \mid B \cap C)P(B \cap C) = P(A \mid B \cap C)P(B \mid C)P(C)$

And you can keep extending this to n variables:

$$P(A_1 \cap A_2 \cap \ldots \cap A_n) = P(A_1 \mid A_2 \cap \ldots \cap A_n)P(A_2 \mid A_3 \cap \ldots \cap \; A_n)P(A_{n-1}|A_n)P(A_n)$$

This idea is known as the **chain rule**.

If on the other hand you have disjoint events $C_1, C_2, \ldots, C_m$ such that $C_1 \cup C_2 \cup \cdots \cup C_m = \Omega$, the probability of any event can be decomposed as:

$$P(A) = P(A \mid C_1)P(C_1) + P(A \mid C_2)P(C_2) + \ldots + P(A \mid C_m)P(C_m)$$

Theorem 3 - Bayes Theorem

The **Bayes theorem**, which is the outcome of this section. Below is the formula that we will dig deeper into in upcoming lessons.

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \text{ - this follows from Theorem 1}$$

Additional note: the complement of an event

You learned about (absolute and relative) complements before, but the complement of an event is also applicable to conditional probabilities.

The basic rule is:

$P(A) + P(A') = 1$

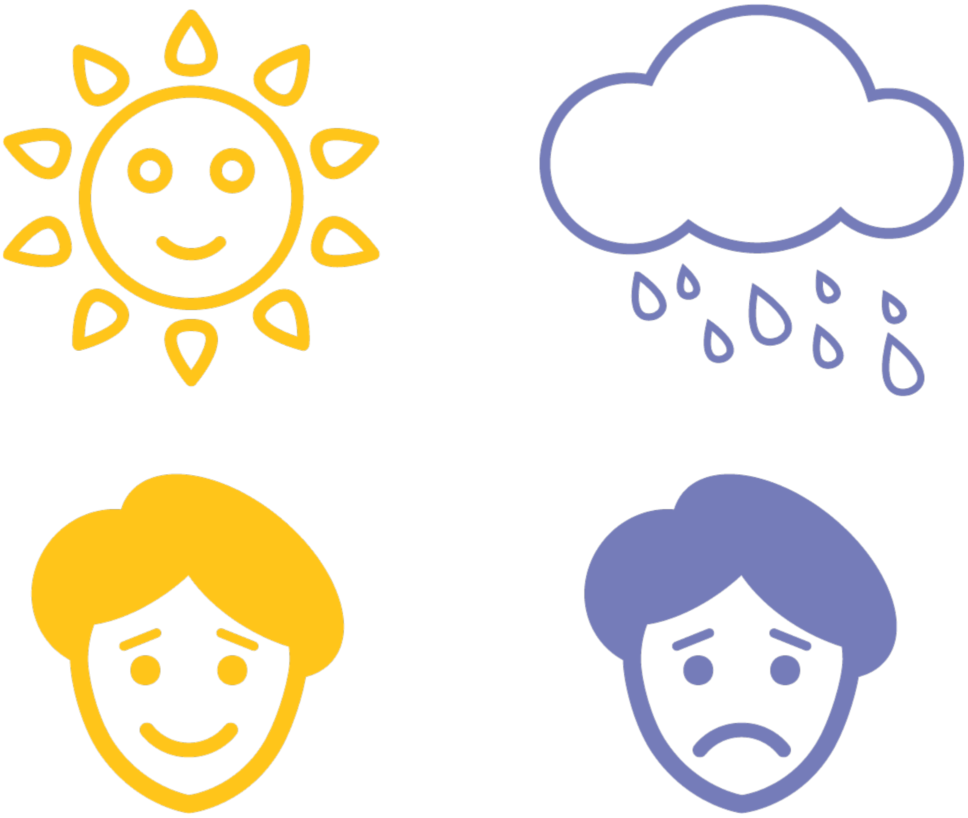
with A' being the complement of A.

Similarly, extending this to conditional probabilities:

$P(A|B) + P(A'|B) = 1$

Example : An Aspiring Data Scientist's Dilemma

Let's see a very simple use of the conditional probability formula. A data scientist comes across the following infographic:



Curious as data scientists are, he starts collecting data about weather conditions and his own mood.

Consider the data in the following table, recorded over a month with 50 days by our data scientist. On each day he recorded whether it was sunny or Cloudy, and whether his mood was good or not.

	Sunny weather	Cloudy weather
Good mood	14	11
Bad mood	2	23

He wants to now know if his mood had anything to do with the weather on a particular day and how he can calculate the probability of having a good mood given the weather conditions.

If he picked a day at random from the 50 days on record, what is the probability that he was in a good mood on that day, $P(G)$?

- The sample space is 50 days here
- The event space is "good mood", so $14 + 11 = 25$.
- $P(G) = \frac{25}{50} = 0.5$

What is the probability that the day chosen was a Sunny day, $P(S)$?

- The sample space is still 50 days
- It was sunny on $14 + 2 = 16$
- $P(S) = \frac{16}{50} = 0.32$

What is the probability of having a good mood given it's a sunny day $P(G \mid S)$?

- $P(G \mid S) = \frac{P(G \cap S)}{P(S)}$, so we need to calculate $P(G \cap S)$ first.
- $P(G \cap S)$ consists of sunny days in which he is in a good mood. There were 14 of them, so $P(G \cap S) = \frac{14}{50}$
- Therefore $P(G \mid S) = \frac{\frac{14}{50}}{\frac{16}{50}} = 0.875$

The infographic had some truth in it indeed. There's a 87.5 chance that our curious data scientist would be in good mood on a sunny day.

The data scientist is satisfied and thinks the outcome is comforting.

Surfing the Internet, however, he comes across a Garth Stein quote. Although not very scientific, this raises his curiosity further. The quote goes as follows:

> "That which is around me does not affect my mood; my mood affects that which is around me"

What if...?

Now the data scientist wants to know if his mood had any impact on the weather. What is $P(S \mid G)$

$$P(S \mid G) = \frac{P(G \cap S)}{P(G)} = \frac{\frac{14}{50}}{\frac{25}{50}} = 0.56$$

He finds that the probability is slightly higher than random chance (50%). In other words, there's a 56% chance that it will be all nice and sunny given that he is in a good mood.

He also realizes that $P(G \mid S)$ is not equal to $P(S \mid G)$. So does this mean that weather has a higher impact on his mood than his mood has on the weather...?

This doesn't really make sense. Our mood doesn't *cause* the weather, so there is no cause-effect relationship. In the example above, the weather and other such external conditions can have a positive effect on human mood and behavior, and this can be said with reference to literature. However, it is unlikely that mood has any effect on weather. There is no scientific evidence to support this notion (and it's very unlikely that there will ever be). What is clear, however, is that there is a relationship between weather and mood.

Say Hello to Reverend Thomas Bayes

Bayes theorem is a very foundational theorem that uses the fact that $P(A \cap B) = P(B)P(A \mid B) = P(A)P(B \mid A)$. Note that, using Bayes theorem, you can compute conditional probabilities without explicitly needing to know $P(A \cap B)$!

This theorem is extremely important in many machine learning algorithms.

Our data scientist realizes that he needs to learn a bit of Bayesian reasoning in order to get more meaningful results. And that is exactly what we will discuss further. First, we need to cover a few topics to fully understand how this simple equation lets you do some serious predictive analysis.

You'll do a few exercises next to get a good grip on conditional probability calculations.

Additional Resources

You are strongly advised to visit the following links to get an in-depth understanding with examples and proofs for explaining the formulas highlighted in this lesson.

Conditional probability, Independence and Bayes rule - A deeper mathematical explanation around Independence and theorems we have seen above (and some we shall cover in upcoming lessons)

Tree Diagrams - Drawing tree diagrams to calculate conditional probability

Conditional Probability, Examples and simple exercises - Practice with probability calculations

Conditional probability: A visual explanation - A great little interactive animation to explain how conditional probability works

Summary

In this lesson, you learned about disjoint, independent, and dependent events, and how to use the addition and multiplication rule to find the probability of the union and intersection of two events, respectively. You also learned how to compute conditional probabilities in case you have dependent events in your sample space, with a step-by-step derivation of the formula used to compute conditional probabilities. You also worked through an example to see this concept in action. Finally, Bayes' theorem was discussed. Later in the course, you'll build further on these ideas towards having a clear understanding of Bayesian Logic and its role in machine learning. Next up, you'll practice solving problems with conditional probability!

Partitioning and the Law of Total Probability

Introduction

In this lesson, we'll look at the law of total probability. In probability theory, the law (or formula) of total probability is a fundamental rule relating **marginal probabilities** to conditional probabilities. It expresses the total probability of an outcome that can be realized via several distinct events.

Objectives

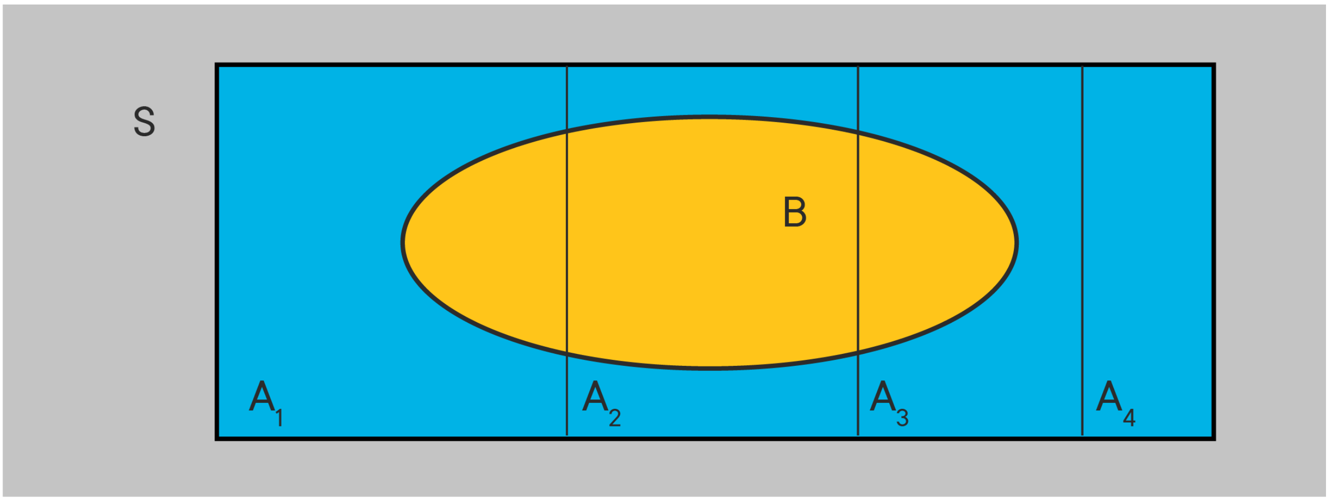
You will be able to:

- State the law of total probabilities based on a partitioned event space
- Explain the concept of event space and partitioning
- Describe conditional independence
- Perform partitioning based on known and unknown probabilities to solve a problem

Partitioning a Sample Space

the Law of Total Probability can be used to calculate $P(B)$. The law requires that you have a set of disjoint events A_i that collectively "cover" the event B. Then, instead of calculating $P(B)$ directly, you add up the intersection of B with each of the events A_i. Let's see this graphically below:

Let $A_1, A_2, \ldots, A_n$ partition sample space S into disjoint regions that sum up to S. In the example, the four regions A_1, A_2, A_3 and A_4 sum up to sample space S.



The probability of a random event B (orange area) can be written down as:

$$\begin{aligned} P(B) &= P(B \cap A_1) + P(B \cap A_2) + P(B \cap A_3) + P(B \cap A_4) \\ &= P(B \mid A_1)P(A_1) + P(B \mid A_2)P(A_2) + P(B \mid A_3)P(A_3) + P(B \mid A_4)P(A_4) \end{aligned}$$

Here we use the first theorem mentioned in the previous lesson to find the combined probabilities.

Example

Let's use a simple example to clarify the image above! The example is created to match the image.

In a certain country, there are four provinces (eg. disjoint regions) A_1, A_2, A_3 and A_4.

You are interested in the total forest area, B, in the country.

Suppose that you know that the forest area in A_1, A_2, and A_3 are 100 km^2, 50 km^2, and 150 km^2, and 0 km^2 respectively. What is the total forest area in the country?

100km^2 + 50km^2 + 150km^2+ 0 km^2 = 300 km^2

$100km^2 + 50km^2 + 150km^2 + 0\ km^2 = 300\ km^2$

We can simply add forest areas in each province to obtain the forest area in the whole country.

This is the idea behind the law of total probability, in which the area of forest is replaced by probability of an event B. In particular, if you want to find $P(B)$, you can look at a partition of S (our sample space composed of $A_1, \ldots, A_4$), and add the amount of probability of A that falls in each partition.

Two Events

In general, we can say that for any two events A and B:

$P(A) = P(A \cap B) + P(A \cap B')$

and using the definition of conditional probability, $P(A \cap B) = P(A \mid B)P(B)$, we can write

$P(A) = P(A \mid B)P(B) + P(A \mid B')P(B')$

The law of total probability is basically a general version of this.

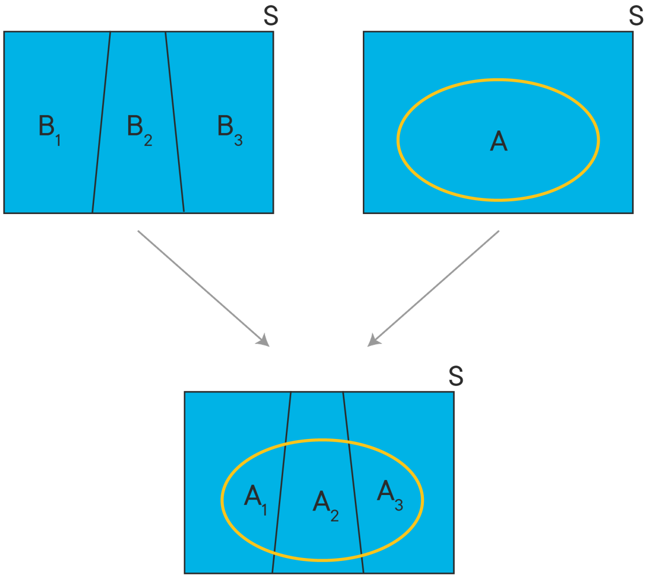
Law of Total Probability

If B_1,B_2,B_3,… is a partition of the sample space S, then for any event A we have

$$P(A) = \sum_i P(A \cap B_i) = \sum_i P(A \mid B_i)P(B_i)$$

Using a Venn diagram, we can pictorially see the idea behind the law of total probability. In the figure below, we have

- $A_1 = A \cap B_1$
- $A_2 = A \cap B_2$
- $A_3 = A \cap B_3$



As it can be seen from the figure, A_1, A_2, and A_3 form a partition of the set A, and thus

$P(A) = P(A_1) + P(A_2) + P(A_3)$

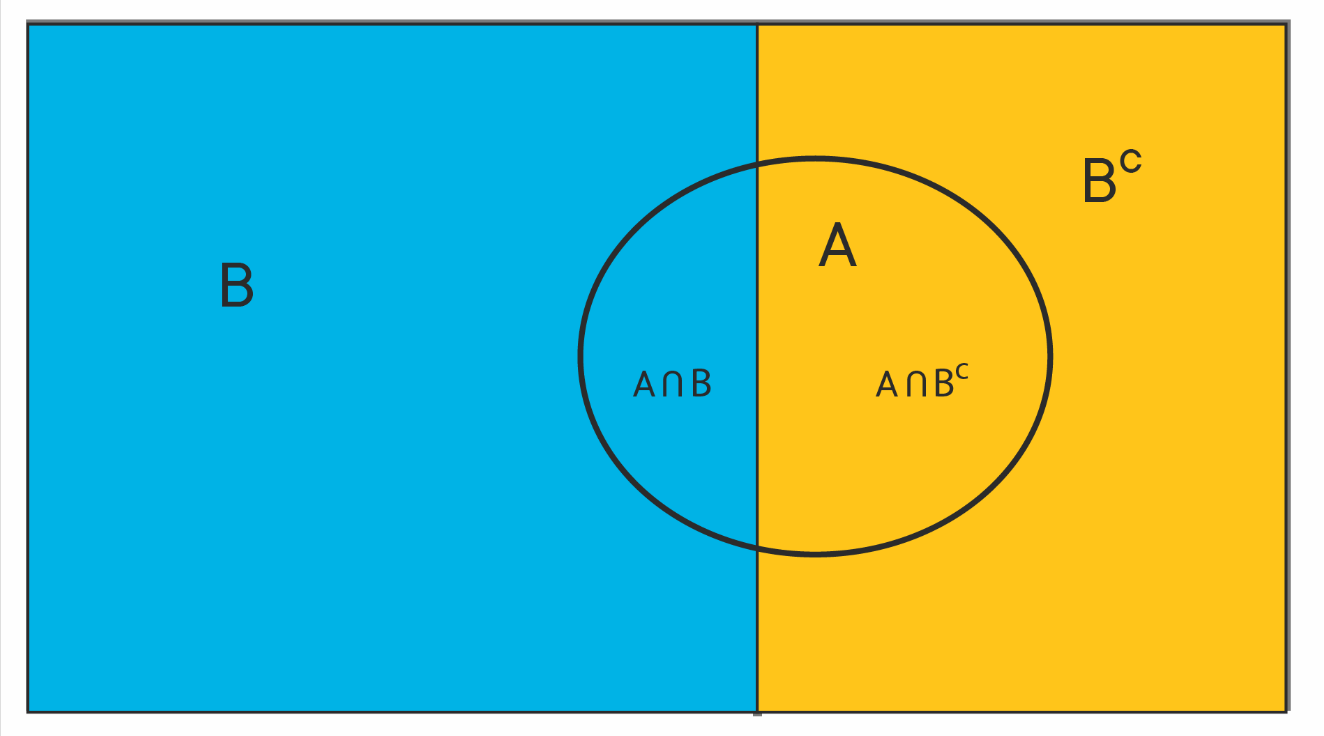
Here is a typical scenario in which we use the law of total probability. We are interested in finding the probability of an event A, but we don't know how to find P(A) directly. Instead, we know the conditional probability of A given some events B_i, where the B_i's form a partition of the sample space. This way, you can use $P(A)$ using the law of total probability

$P(A) = \sum_i P(A \mid B_i)P(B_i)$

More on Partitions

- The natural numbers N can be partitioned into even and odd numbers.
- The set of animal species in the world can be partitioned into subsets where a subset reflects a continent and each species is positioned in a subset depending on which continent they originated from.

In statistics, choosing the right partitioning is key as bad choices of partitions may results in many sub-problems that are even more difficult to solve.



The probability of A can be written as sums of event B (note that B^c is another way of writing B') The total probability rule is:

$P(A) = P(A\cap B) + P(A\cap B^c)$

An alternate version of the total probability rule (found with the multiplication rule) can be used when the necessary probabilities are known:

$P(A) = P(A \mid B)P(B) + P(A \mid B^c)P(B^c)$

You need to be careful when dealing with conditional probabilities and conditioning. Let's look at a few examples to see this idea in action.

Example 1

In a certain county in the United States, 60% of registered voters are Republicans, 30% are Democrats and 10% are Independents.

When those voters were asked about increasing military spending

- 40% of Republicans opposed it
- 65% of the Democrats opposed it
- 55% of the Independents opposed it.

What is the probability that a randomly selected voter in this county opposes increased military spending?

You know that:

- Ω = {registered voters in the county}
- R = {registered republicans}, $P(R)$ = 0.6
- D = {registered democrats}, $P(D)$ = 0.3
- I= {registered independents}, $P(I)$ = 0.1
- B = {registered voters opposing increased military spending}

You also know that:

- $P(B \mid R) = 0.4$
- $P(B \mid D) = 0.65$
- $P(B \mid I) = 0.55$

By the total probability theorem:

$Pr(B) = Pr(B \mid R)Pr(R) + Pr(B \mid D)Pr(D) + Pr(B \mid I)Pr(I)$

$= (0.4 * 0.6) + (0.65 * 0.3) + (0.55 * 0.1) = 0.49$

Example 2

Let's consider a 2-card hand drawn from a standard playing deck. What is the probability of drawing 2 aces, given that we know one of the cards is an ace?

$$P(\text{both are aces | one is ace}) = \dfrac{P(\text{both are aces})}{P(\text{one is ace})} = \dfrac{P(\text{both are aces})}{1 - P(\text{neither is ace})} =\dfrac{\binom{4}{2}/\binom{52}{2}}{1 - \binom{48}{2}/\binom{52}{2}}=\dfrac{1}{33}$$

But now think about this: What is the probability of drawing 2 aces, knowing that one of the cards **is the ace of spades**?

$$P(\text{both are aces | ace of spades}) = P(\text{other card is also an ace}) = \dfrac{3}{51}= \dfrac{1}{17}$$

Notice how the fact that we know we have the ace of spades nearly doubles the probability of having 2 aces

Example 3

Suppose there is a test for a disease, and this test is said to be "95% accurate". The disease in question afflicts 1% of the population. Now say that there is a patient who tests positive for this disease under this test.

First, we define the events in question:

Let D be the event that the patient actually has the disease.

Let T be the event that the patient tests positive.

Since that phrase "95% accurate" is ambiguous, we need to clarify that.

$$P(T|D) = P(T^c|D^c) = 0.95$$

In other words, **conditioning on whether or not the patient has the disease**, we will assume that the test is 95% accurate.

What exactly are we trying to find?

What the patient really wants to know is not $P(T|D)$, which is the accuracy of the test; but rather $P(D|T)$, or the probability she has the disease given that the test returns positive. Fortunately, we know how $P(T|D)$ relates to $P(D|T)$.

$$\begin{align} P(D|T) &= \frac{P(T|D)P(D)}{P(T)} ~~~~ & &\text{... Bayes Rule} \\ &= \frac{P(T|D)P(D)}{P(T|D)P(D) + P(T|D^c)P(D^c)} ~~~~ & & \text{... by the Law of Total Probability} \\ &= \frac{(0.95)(0.01)}{(0.95)(0.01) + (0.05)(0.99)} ~~~~ & & \text{... the rarity of the disease competes with the rarity of true negatives}\\ &\approx 0.16 \end{align}$$

Common Pitfalls

- Mistaking $P(A|B)$ for $P(B|A)$.

This is also known as the Prosecutor's Fallacy, where instead of asking about the *probability of guilt (or innocence) given all the evidence*, we make the mistake of concerning ourselves with the *probability of the evidence given guilt*.

- Confusing *prior* $P(A)$ with *posterior* $P(A \mid B)$.

Observing that event A occurred does **not** mean that $P(A) = 1$. But $P(A \mid A) = 1$ and $P(A) \neq 1$.

- Confusing *independence* with **conditional independence**.

This is more subtle than the other two. Let's look at this in a bit more detail

Conditional Independence

Events A and B are **conditionally independent** given event C, if

$$P(A \cap B \mid C) = P(A \mid C)P(B \mid C)$$

i.e. conditioning on event C does not give us any additional information on A or B.

Conditional independence given C DOES NOT imply unconditional independence

Consider playing a series of 5 games against a chess opponent of unknown strength. Winning all five games would give you a good idea that you are a better player. So winning each successive game is actually providing us with information about the strength of our opponent. If you have prior knowledge about the strength of your opponent, you condition on the strength of our opponent i.e. winning one game would not provide us with any additional information on the probability of winning the next. Having no prior knowledge of your opponent and winning a string a games will give you information about the probability of winning the next game.

The games are conditionally independent given the strength of our opponent, but **not** independent unconditionally.

Unconditional independence DOES NOT imply conditional independence given C

For example, let A be the event of the fire alarm going off, F be the event of a fire, and C be the event of someone making popcorn. Suppose that either F or C will result in A and the fire alarm going off. Now if F and C are independent: knowing that there's a fire F doesn't tell you anything about anyone making popcorn C, and vice versa. But the probability of a fire given that the alarm goes off **and** no one is making any popcorn is given by $P(F \mid A,C^c) = 1$. After all, if the fire alarm goes off and no one is making popcorn, there can only be one explanation: *there must be a fire*.

So F and C may be independent, but they are not *conditionally independent* when we condition on event A. Knowing that nobody is making any popcorn when the alarm goes off can only mean that there is a fire.

Additional Resources

You are strongly advised to visit following links to get an indepth understanding with examples and proofs for formulas highlighted in this lesson.

The law of total probability - concept and proof - Excellent YouTube video by Phil Chan.

Conditional (Partitioned) Probability — A Primer - Deep dive into partitions (A Must Read)

Law of Total Probability - More examples for a deeper understanding around partitioning

Summary

In this lesson, you further learned about the ideas of conditional probability covered in the previous lessons to explain the law of total probability using partitioning of the sample space. You learned how you can partition probabilities with respect to some other event, when the direct probabilities are not known. Let's move on to some practice!