1 Probabilistic Foundations for Machine Learning

1.1 Introduction to Probability

Probability provides a mathematical framework for reasoning about uncertainty. In machine learning, uncertainty arises from noisy data, incomplete information, and randomness in data-generating processes. Before exploring probability distributions and Bayesian reasoning, we begin with the basic idea of probability itself.

Probability measures how likely an event is to occur. For example, if a fair coin is flipped, the probability of obtaining heads is

$$P(\text{Heads}) = \frac{1}{2}.$$

Similarly, when a fair six-sided die is rolled, the probability of landing on the number 4 is

$$P(4) = \frac{1}{6}.$$

Consider a school with 10 children, where 3 of them play soccer and 7 do not. If a child is selected at random, we want to find the probability that the child plays soccer. Denote this event by Soccer. Using the classical definition of probability,

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

The numerator represents the event of interest, while the denominator represents the sample space. This simple example demonstrates how probability quantifies uncertainty using ratios of favorable outcomes to total possible outcomes.

Now consider a simple random experiment: flipping a fair coin. Since the outcome is uncertain, it qualifies as an experiment in probability theory. A fair coin has two equally likely outcomes:

$$P(\text{Heads}) = \frac{1}{2}, \quad P(\text{Tails}) = \frac{1}{2}.$$

If we flip two fair coins, the total number of outcomes is 4. The possible outcomes are:

$$\{\text{HH},\ \text{HT},\ \text{TH},\ \text{TT}\}.$$

Only one outcome contains two heads. Therefore,

$$P(\text{HH}) = \frac{1}{4}.$$

If three fair coins are flipped, each coin has two possible outcomes, giving a total of $2^3 = 8$ outcomes:

$$\{\text{HHH},\ \text{HHT},\ \text{HTH},\ \text{HTT},\ \text{THH},\ \text{THT},\ \text{TTH},\ \text{TTT}\}.$$

Only one outcome contains three heads, so the probability is

$$P(\text{HHH}) = \frac{1}{8}.$$

These examples illustrate how probabilities are computed by counting favorable outcomes and dividing by the total number of possible outcomes. This foundational understanding enables deeper study of probability distributions, conditional probability, and Bayesian inference—all central concepts in machine learning.

1.2 Reinforcing Probability with Dice Experiments

To further strengthen the understanding of probability, consider another common random experiment: rolling a fair six-sided die. Each roll of the die produces one outcome from the set $\{1, 2, 3, 4, 5, 6\}$, and all outcomes are equally likely.

Suppose we want to determine the probability of rolling a six. Since there are six possible outcomes and only one of them is favorable, the probability is

$$P(6) = \frac{1}{6}.$$

Now consider rolling two fair dice simultaneously. Each die has six possible outcomes, and because the two dice are independent, the total number of combined outcomes is

$$6 \times 6 = 36.$$

These outcomes can be listed as ordered pairs (i, j), where the first number represents the outcome of Die 1 and the second number represents the outcome of Die 2. The complete sample space is shown in Table 1.

Die 1 \ Die 2	**1**	**2**	**3**	**4**	**5**	**6**
1	(1,1)	(1,2)	(1,3)	(1,4)	(1,5)	(1,6)
2	(2,1)	(2,2)	(2,3)	(2,4)	(2,5)	(2,6)
3	(3,1)	(3,2)	(3,3)	(3,4)	(3,5)	(3,6)
4	(4,1)	(4,2)	(4,3)	(4,4)	(4,5)	(4,6)
5	(5,1)	(5,2)	(5,3)	(5,4)	(5,5)	(5,6)
6	(6,1)	(6,2)	(6,3)	(6,4)	(6,5)	(6,6)

Table 1: Sample space of all 36 possible outcomes when rolling two fair six-sided dice.

From the table, it is clear that only one outcome corresponds to both dice showing a six, namely $(6, 6)$. Therefore, the probability of this event is

$$P(6, 6) = \frac{1}{36}.$$

Dice experiments provide an intuitive setting to explore how probabilities behave when multiple independent events occur. They also illustrate how sample spaces grow multiplicatively as more random components are introduced.

In practice, repeated rolling of dice (or flipping coins) shows that the observed relative frequencies gradually approach the theoretical probabilities as the number of trials increases. This behavior reflects the law of large numbers, a foundational idea that connects probability theory with empirical observations.

1.3 The Complement of an Event

In the previous discussion, we examined how to compute the probability of an event. In this subsection, we explore an important related concept: the probability that an event does not occur. This is known as the *complement* of the event.

If an event has a probability of occurring equal to 0.75, then the probability that it does not occur is simply 0.25. These two probabilities always add up to 1 because one of the two possibilities must happen: either the event occurs or it does not.

To make this concrete, consider again the example of a school with 10 children, where 3 of them play soccer and 7 do not. Suppose we want the probability that a randomly selected child does not play soccer. Denoting this event by Not Soccer, we compute:

$$P(\text{Not Soccer}) = \frac{7}{10} = 0.7.$$

Notice that this value is closely related to the probability of the event Soccer, which we previously found to be 0.3. Indeed,

$$0.7 + 0.3 = 1.$$

This relationship is not a coincidence. It is a fundamental result known as the **complement rule**. The complement rule states that the probability of an event A not occurring is equal to one minus the probability that A does occur:

$$P(A') = 1 - P(A).$$

Here, A' represents the complement of the event A.

Using this rule, we can compute the probability of an event not happening in a quick and efficient way. For example, consider the experiment of flipping three coins. We previously determined that the probability of obtaining three heads is

$$P(\text{HHH}) = \frac{1}{8}.$$

Therefore, the probability of *not* obtaining three heads is

$$P(\text{Not HHH}) = 1 - \frac{1}{8} = \frac{7}{8}.$$

This result can also be interpreted by counting: out of the 8 total possible outcomes of flipping three coins, exactly 7 outcomes do not consist entirely of heads.

We now apply the complement rule to the dice experiment. Suppose we roll a single fair die. What is the probability of obtaining any number other than 6? Since the probability of rolling a 6 is

$$P(6) = \frac{1}{6},$$

the complement rule gives

$$P(\text{Not } 6) = 1 - \frac{1}{6} = \frac{5}{6}.$$

Thus, the complement rule provides a powerful method for calculating the probability that an event does not occur, especially in situations where computing the complementary event is simpler than analyzing the event directly.

1.4 Sum of Probabilities

The sum of probabilities rule helps us compute the probability that at least one of several events occurs. However, the way we apply this rule depends on whether the events overlap. We begin by examining the simpler case of disjoint events, followed by the general rule for events that may overlap.

1.4.1 Disjoint (Mutually Exclusive) Events

Two events are called *disjoint* or *mutually exclusive* if they cannot occur at the same time. In this case, the probability that either event occurs is simply the sum of their individual probabilities.

Consider rolling a fair six-sided die. The probability of obtaining a 2 is:

$$P(2) = \frac{1}{6},$$

and the probability of obtaining a 3 is:

$$P(3) = \frac{1}{6}.$$

Since a single roll cannot produce both outcomes simultaneously, these events are disjoint. Thus,

$$P(2 \text{ or } 3) = P(2) + P(3) = \frac{1}{6} + \frac{1}{6} = \frac{2}{6}.$$

A second example involves a school where each child can play only one sport. Suppose the probability that a child plays soccer is 0.3, and the probability that a child plays basketball is 0.4. Because students cannot play both sports, the probability that a child plays soccer or basketball is:

$$P(\text{Soccer or Basketball}) = 0.3 + 0.4 = 0.7.$$

We can apply this rule to dice as well. If we roll one fair die, let

$$A = \text{rolling an even number}, \qquad B = \text{rolling a five}.$$

There are three even outcomes (2, 4, 6) and one outcome of 5. Because these outcomes do not overlap,

$$P(A \text{ or } B) = \frac{3}{6} + \frac{1}{6} = \frac{4}{6} = \frac{2}{3}.$$

Another example involves the sum of two dice. Out of 36 possible outcomes, six produce a sum of seven, and three produce a sum of ten. Since no outcome can produce both a sum of seven and a sum of ten, these two events are disjoint. Therefore,

$$P(\text{Sum 7 or Sum 10}) = \frac{6}{36} + \frac{3}{36} = \frac{9}{36} = \frac{1}{4}.$$

As another illustration, consider the probability that the difference between the dice is 2 or the difference is 1. There are eight outcomes with difference 2 and ten outcomes with difference 1, and these sets do not overlap. Thus,

$$P(\text{Diff 2 or Diff 1}) = \frac{8}{36} + \frac{10}{36} = \frac{18}{36} = \frac{1}{2}.$$

These examples demonstrate that when two events cannot occur simultaneously, the probability of their union is found by simple addition.

1.4.2 Joint (Non–Mutually Exclusive) Events

Many real-world events are not disjoint; they may occur simultaneously. In such cases, the simple addition rule fails because it double-counts the outcomes in which both events occur. Instead, we use the general sum rule:

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

To see why this is necessary, consider the probability of rain and wind. Suppose:

$$P(\text{Rain}) = 0.8, \qquad P(\text{Wind}) = 0.7.$$

If we simply add these, we get 1.5, which is impossible because probabilities cannot exceed 1. The problem arises because rain and wind may occur together, so the overlap must be subtracted.

Now consider a school where children may play multiple sports. Suppose:

$$P(\text{Soccer}) = 0.6, \qquad P(\text{Basketball}) = 0.5.$$

If we added these directly, we would get 1.1, again impossible. We need to know how many children play both sports, that is, the value of $P(\text{Soccer} \cap \text{Basketball})$.

Suppose the following data are known: - Six children play soccer, - Five children play basketball, - Three children play both.

Then the number of children who play soccer or basketball is:

$$6 + 5 - 3 = 8.$$

In probability terms (dividing by the total number of children), the corresponding rule is:

$$P(\text{Soccer} \cup \text{Basketball}) = P(\text{Soccer}) + P(\text{Basketball}) - P(\text{Soccer} \cap \text{Basketball}).$$

Using the probabilities:

$$P(\text{Soccer} \cup \text{Basketball}) = 0.6 + 0.5 - 0.3 = 0.8.$$

We now apply the joint rule to dice. Consider the events:

$$A = \text{sum of the dice is } 7, \qquad B = \text{difference of the dice is } 1.$$

There are six outcomes with sum 7 and ten outcomes with difference 1. However, two outcomes satisfy both conditions: $(3, 4)$ and $(4, 3)$. Therefore,

$$P(A \cup B) = \frac{6}{36} + \frac{10}{36} - \frac{2}{36} = \frac{14}{36} = \frac{7}{18}.$$

This example illustrates why the general sum rule is essential for calculating probabilities of events that may overlap. Subtracting the intersection prevents double-counting, ensuring the probability remains accurate.

1.5 Independent Events

Two events are said to be *independent* if the occurrence of one event does not affect the probability of the occurrence of the other. Independence is a fundamental idea in probability theory and plays a critical role in many machine learning models, where assumptions of independence help simplify both calculations and model structure.

A classic example of independence occurs when tossing a fair coin multiple times. The result of the first toss has no influence on what happens in the second toss. In contrast, events such as consecutive moves in a game of chess are not independent—each move affects the next.

Independence allows us to compute probabilities using the **product rule**. If events A and B are independent, then

$$P(A \cap B) = P(A) \times P(B).$$

Example 1: Coin Tossing

Consider tossing a fair coin five times. Each toss is independent of the others. The probability of obtaining heads on a single toss is

$$P(\text{Heads}) = \frac{1}{2}.$$

Therefore, the probability of obtaining heads on all five tosses is

$$\left(\frac{1}{2}\right)^5 = \frac{1}{32}.$$

This illustrates that when several independent events must all occur, we multiply their individual probabilities.

Example 2: Rolling Dice

When rolling two fair dice, the outcome of one die does not influence the other. Therefore, the probability of obtaining a six on both dice is

$$P(6 \text{ on die 1 and } 6 \text{ on die 2}) = \frac{1}{6} \times \frac{1}{6} = \frac{1}{36}.$$

This principle generalizes. If we roll ten fair dice, the probability that all ten land on six is

$$\left(\frac{1}{6}\right)^{10},$$

a very small number. The independence of the dice rolls allows us to compute this probability easily using repeated multiplication.

Example 3: Drawing Cards With Replacement

Suppose a standard deck of 52 cards is shuffled. You draw one card, replace it, shuffle the deck again, and draw a second card. Because the card is replaced, the second draw is made from the same full deck, so the two draws do not influence each other.

Let event A be "drawing a red card" and event B be "drawing a red card on the second draw." Since half the cards in the deck are red,

$$P(A) = \frac{26}{52} = \frac{1}{2}, \qquad P(B) = \frac{1}{2}.$$

The probability that both drawn cards are red is

$$P(A \cap B) = P(A) \times P(B) = \frac{1}{2} \times \frac{1}{2} = \frac{1}{4}.$$

These examples highlight the **product rule for independent events**:

$$P(A \cap B) = P(A)\, P(B).$$

This rule extends to any number of independent events. If $A_1, A_2, \ldots, A_n$ are independent, then

$$P(A_1 \cap A_2 \cap \cdots \cap A_n) = P(A_1)\, P(A_2) \cdots P(A_n).$$

Example 4: Multiple Independent Events

Consider a short multiple-choice quiz with five questions. Each question has four answer choices, and a student selects answers randomly without any pattern or prior knowledge. Since each question is answered independently, the probability of choosing the correct answer on one question is

$$P(\text{Correct}) = \frac{1}{4}.$$

Because the questions are independent, the probability of getting all five questions correct purely by guessing is

$$P(\text{All Correct}) = \left(\frac{1}{4}\right)^5 = \frac{1}{1024}.$$

This example shows how quickly probabilities become small when many independent events must all occur at the same time.

Independence is a powerful simplification tool. By recognizing when events do not influence one another, we can compute probabilities more efficiently and understand the structure of complex probabilistic systems.

1.6 Conditional Probability

So far, we have worked with probabilities where all outcomes in the sample space were considered equally when computing an event's likelihood. In many real situations, however, we often have additional information about what has already happened. This extra information changes the sample space and therefore changes the probability of certain outcomes. This leads us to the idea of *conditional probability*.

Conditional probability refers to the probability of an event occurring given that another event has already occurred. For example, the probability that today is humid may change if we are told that yesterday it rained. The extra information modifies our expectations.

Coin Example

Recall the sample space when tossing two fair coins:

$$\{HH, HT, TH, TT\}.$$

The probability of obtaining two heads is

$$P(HH) = \frac{1}{4}.$$

Now suppose we are told that the first coin landed on heads. This additional information removes the outcomes starting with T. The new sample space becomes:

$$\{HH, HT\}.$$

Only one of these outcomes has two heads, so the conditional probability is

$$P(HH \mid \text{first is H}) = \frac{1}{2}.$$

Similarly, if we are told that the first coin landed on tails, then the new sample space is

$$\{TH, TT\},$$

and there is no way to obtain two heads. Thus,

$$P(HH \mid \text{first is T}) = 0.$$

This shows how a condition can dramatically change the probability of an event.

Connecting Conditional Probability to the Product Rule

Earlier, we learned the product rule for independent events:

$$P(A \cap B) = P(A)\,P(B).$$

But this only holds when A and B are independent.

When events are not independent, we must account for how one event affects the other. Consider rolling two dice and asking:

What is the probability that the first die shows a 6 and the sum of the two dice is 10?

The only favorable outcome is $(6, 4)$, so

$$P(\text{first is 6 and sum is 10}) = \frac{1}{36}.$$

Now observe how this relates to conditional probability. The probability that the first die is 6 is

$$P(\text{first is 6}) = \frac{6}{36}.$$

Among the six outcomes where the first die is 6, only one outcome produces a sum of 10. Therefore,

$$P(\text{sum is 10} \mid \text{first is 6}) = \frac{1}{6}.$$

Multiplying these two numbers gives

$$\frac{6}{36} \times \frac{1}{6} = \frac{1}{36},$$

which matches our earlier answer.

This leads to the general formula:

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

When A and B are independent, $P(B \mid A) = P(B)$, restoring the earlier product rule.

Dice Example with Conditions

Consider again the event of obtaining a sum of 10 when rolling two dice. Without any conditions,

$$P(\text{sum is } 10) = \frac{3}{36} = \frac{1}{12}.$$

Now suppose we are told that the first die shows a 6. The new sample space is:

$$\{(6,1),(6,2),(6,3),(6,4),(6,5),(6,6)\}.$$

Only one of these six outcomes produces a sum of 10, so

$$P(\text{sum is } 10 \mid \text{first is } 6) = \frac{1}{6}.$$

On the other hand, if we are told the first die shows a 1, the possible outcomes are:

$$\{(1,1),(1,2),(1,3),(1,4),(1,5),(1,6)\}.$$

None of these produce a sum of 10, so

$$P(\text{sum is } 10 \mid \text{first is } 1) = 0.$$

These examples clearly show how a condition reshapes the sample space and alters the resulting probability. Conditional probability is a key tool in statistical reasoning, machine learning, and inference, especially when information arrives sequentially or when decisions depend on observed events.

Example: Conditional Probability in Email Spam Classification

Conditional probability appears frequently in machine learning models that operate under uncertainty. A common example is spam detection. Imagine a simple spam classifier that uses the presence of a specific keyword—say, the word "offer"—as one of its features.

Suppose we know the following information from historical email data:

$$P(\text{Spam}) = 0.3, \qquad P(\text{“offer”} \mid \text{Spam}) = 0.6, \qquad P(\text{“offer”}) = 0.2.$$

A useful conditional probability here is the probability that an email is spam *given* that it contains the word "offer." Using the definition of conditional probability:

$$P(\text{Spam} \mid \text{“offer”}) = \frac{P(\text{Spam} \cap \text{“offer”})}{P(\text{“offer”})}.$$

We compute the numerator using the product rule:

$$P(\text{Spam} \cap \text{“offer”}) = P(\text{Spam})\, P(\text{“offer”} \mid \text{Spam}) = 0.3 \times 0.6 = 0.18.$$

Thus,

$$P(\text{Spam} \mid \text{“offer”}) = \frac{0.18}{0.2} = 0.9.$$

This example shows how additional information can change a probability, which is exactly what conditional probability captures.

1.7 Bayes' Theorem

Conditional probability allows us to update the likelihood of an event when new information becomes available. Bayes' theorem builds on this idea and provides a precise mathematical rule for reversing conditional probabilities. It is widely used in areas such as medical diagnosis, spam filtering, recommendation systems, and many probabilistic models in machine learning.

A Motivating Example: Medical Testing

Consider a rare disease in a population of one million people. Suppose the disease affects only 1 out of every 10,000 individuals. This means:

$$P(\text{Sick}) = \frac{1}{10{,}000} = 0.0001, \qquad P(\text{Healthy}) = 0.9999.$$

Now imagine we have a medical test that is 99% accurate:

$$P(\text{Positive} \mid \text{Sick}) = 0.99, \qquad P(\text{Negative} \mid \text{Healthy}) = 0.99.$$

This also implies:

$$P(\text{Positive} \mid \text{Healthy}) = 0.01,$$

because 1% of healthy individuals will incorrectly test positive.

Suppose you take the test and receive a *positive* result. The key question is:

What is the probability that you are actually sick, given that you tested positive?

To answer this, it is helpful to imagine testing all one million people.

Population Breakdown

- Sick individuals: 100
- Healthy individuals: 999,900

Apply the test:

- Among the 100 sick individuals, 99 will correctly test positive. - Among the 999,900 healthy individuals, 1% will test positive, which is 9,999 people.

Thus, the total number of people who test positive is:

$$99 + 9{,}999 = 10{,}098.$$

Only 99 of these are actually sick, so:

$$P(\text{Sick} \mid \text{Positive}) = \frac{99}{10{,}098} \approx 0.0098.$$

Even though the test is highly accurate, the probability that a person is sick *after a positive test* is less than 1%, mainly because the disease is extremely rare. This example highlights why conditional probability must be handled carefully in medical and ML contexts.

Deriving Bayes' Theorem

Let

$$A = \text{Sick}, \qquad B = \text{Test Positive}.$$

We wish to compute $P(A \mid B)$. From the definition of conditional probability:

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

To compute $P(A \cap B)$, we apply the product rule:

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

The term $P(B)$ can be computed by considering all ways a person can test positive:

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

Using the product rule again:

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

where A' denotes "not sick."

Substituting this into the conditional probability formula gives Bayes' theorem:

$$P(A \mid B) = \frac{P(A)P(B \mid A)}{P(A)P(B \mid A) + P(A')P(B \mid A')}.$$

Applying Bayes' Theorem to the Example

Plugging in the numbers:

$$P(A) = 0.0001, \qquad P(A') = 0.9999,$$

$$P(B \mid A) = 0.99, \qquad P(B \mid A') = 0.01.$$

Thus,

$$P(A \mid B) = \frac{0.0001 \times 0.99}{0.0001 \times 0.99 + 0.9999 \times 0.01} = \frac{0.000099}{0.000099 + 0.009999} \approx 0.0098.$$

This matches the earlier calculation and serves as a clear illustration of Bayes' theorem in action.

Bayes' theorem provides a powerful tool for updating probabilities using new evidence. It forms the foundation of many probabilistic models and will appear again when we discuss Bayesian inference, Naive Bayes classifiers, and decision-making under uncertainty.

Example: Applying Bayes' Theorem to Spam Classification

Bayes' theorem is widely used in email spam filtering. To illustrate this, consider a simple dataset of 100 emails, where 20 are spam and 80 are not spam. Without examining any features of the emails, the probability that a randomly selected email is spam is:

$$P(\text{Spam}) = \frac{20}{100} = 0.2, \qquad P(\text{Not Spam}) = 0.8.$$

Suppose we examine the presence of a particular keyword, such as the word "lottery." Among the 20 spam emails, 14 contain this word, and among the 80 non-spam (ham) emails, 10 contain

it. We now want to compute the probability that an email is spam *given* that it contains the word "lottery":

$$P(\text{Spam} \mid \text{Lottery}).$$

Direct Counting Approach

We focus only on the emails that contain the word "lottery." There are:

$$14 \text{ spam emails} + 10 \text{ non-spam emails} = 24 \text{ total}.$$

Among these 24, exactly 14 are spam, so:

$$P(\text{Spam} \mid \text{Lottery}) = \frac{14}{24} = \frac{7}{12} \approx 0.583.$$

Using Bayes' Theorem

We verify this result using Bayes' formula. First compute the following quantities:

$$P(\text{Lottery} \mid \text{Spam}) = \frac{14}{20} = 0.7, \qquad P(\text{Lottery} \mid \text{Not Spam}) = \frac{10}{80} = 0.125.$$

Apply Bayes' theorem:

$$P(\text{Spam} \mid \text{Lottery}) = \frac{P(\text{Spam})\,P(\text{Lottery} \mid \text{Spam})}{P(\text{Spam})\,P(\text{Lottery} \mid \text{Spam}) + P(\text{Not Spam})\,P(\text{Lottery} \mid \text{Not Spam})}.$$

Substitute the values:

$$P(\text{Spam} \mid \text{Lottery}) = \frac{0.2 \times 0.7}{0.2 \times 0.7 + 0.8 \times 0.125} = \frac{0.14}{0.14 + 0.10} \approx 0.583.$$

Both approaches agree. Conditioning on the presence of a specific word substantially increases the estimated probability that an email is spam. This is the basic mechanism behind many simple spam classifiers, where features such as keywords modify the prior probability of spam to produce a more accurate posterior probability.

1.8 Prior, Likelihood, Evidence, and Posterior

Bayes' theorem provides a mathematical framework for updating probabilities when new information becomes available. To understand this update process clearly, it is useful to identify the four fundamental components involved in Bayesian reasoning: the *prior*, the *likelihood*, the *evidence*, and the *posterior*.

Bayes' theorem is:

$$P(A \mid B) = \frac{P(A)\,P(B \mid A)}{P(A)\,P(B \mid A) + P(A')\,P(B \mid A')}.$$

Here is the interpretation of each term:

- **Prior** $P(A)$: The initial belief about event A before receiving any additional information.
- **Likelihood** $P(B \mid A)$: The probability of observing the evidence B assuming A is true.
- **Evidence** $P(B)$: The total probability of observing B, regardless of which underlying event caused it.

- **Posterior** $P(A \mid B)$: The updated probability of event A after observing evidence B.

Illustration: Spam Classification

Consider a dataset of 100 emails, of which 20 are spam and 80 are not spam. Let

$$A = \text{Spam}, \qquad B = \text{Email contains the word “lottery”}.$$

We compute each component of Bayes' theorem as follows.

- **Prior:**
$$P(A) = \frac{20}{100} = 0.2, \qquad P(A') = 0.8.$$
Before examining any features, 20% of emails are spam.
- **Likelihoods:** Among the 20 spam emails, 14 contain the word “lottery”:
$$P(B \mid A) = \frac{14}{20} = 0.7.$$
Among the 80 non-spam emails, 10 contain the word:
$$P(B \mid A') = \frac{10}{80} = 0.125.$$
- **Evidence:**
$$P(B) = P(A)P(B \mid A) + P(A')P(B \mid A') = 0.2 \times 0.7 + 0.8 \times 0.125.$$
- **Posterior:** Applying Bayes' theorem:
$$P(A \mid B) = \frac{0.2 \times 0.7}{0.2 \times 0.7 + 0.8 \times 0.125} \approx 0.583.$$

This means that although the prior probability of spam is only 20%, the probability that an email is spam increases to about 58.3% once we know it contains the word “lottery.”

In general, Bayesian reasoning follows the pattern:

$$\text{Posterior} = \frac{\text{Prior} \times \text{Likelihood}}{\text{Evidence}}.$$

The posterior reflects an updated and more accurate belief because it incorporates useful information contained in the evidence. This mechanism forms the basis of many probabilistic models in machine learning, including Naive Bayes classifiers.

1.9 Naive Bayes Classifier

Bayes' theorem provides a systematic way to update the probability of a class after observing new evidence. In earlier examples, we considered a single feature (e.g., the presence of the word “lottery”) and computed the posterior probability

$$P(\text{Spam} \mid \text{Lottery}).$$

However, real emails contain many informative words. Relying on a single feature usually yields a weak classifier, so we naturally want to model many words simultaneously. For a document

$$X = (w_1, w_2, \ldots, w_n),$$

our goal is to compute:

$$P(\text{Spam} \mid X).$$

Step 1: Start with Bayes' Rule

Bayes' theorem gives:

$$P(\text{Spam} \mid X) = \frac{P(X \mid \text{Spam})\, P(\text{Spam})}{P(X)}.$$

Since $P(X)$ is the same for both classes, we compare only the numerator:

$$P(\text{Spam} \mid X) \propto P(X \mid \text{Spam})\, P(\text{Spam}).$$

The Challenge

To compute $P(X \mid \text{Spam})$ directly, we would need counts for *all* word combinations:

$$(w_1, w_2, \ldots, w_n).$$

This quickly becomes impossible: with thousands of words, most combinations never appear, producing unusable probabilities like $0/0$.

Step 2: The Naive Conditional Independence Assumption

Naive Bayes assumes that the words in a document are *conditionally independent* given the class:

$$P(w_1, w_2, \ldots, w_n \mid \text{Spam}) = \prod_{i=1}^{n} P(w_i \mid \text{Spam}).$$

This assumption is not literally true—words co-occur in meaningful patterns ("free money", "urgent offer")—but the resulting classifier works extremely well in practice.

Step 3: Combine the Steps

Substituting the product into Bayes' rule:

$$P(\text{Spam} \mid w_1, w_2, \ldots, w_n) \propto P(\text{Spam}) \prod_{i=1}^{n} P(w_i \mid \text{Spam}).$$

Interpretation:

- $P(\text{Spam})$: prior probability that a message is spam.
- $P(w_i \mid \text{Spam})$: how strongly word w_i indicates spam.
- The product aggregates evidence across all words.

Step 4: example

Consider the message: *"Congratulations, you won free money!"*

Words such as *free*, *money*, and *won* typically have high likelihood in spam emails:

$$P(\text{free} \mid \text{Spam}), \quad P(\text{money} \mid \text{Spam}), \quad P(\text{won} \mid \text{Spam}),$$

so their combined effect strongly increases the posterior probability of spam.

Step 5: Use Log Probabilities (Avoid Numerical Underflow)

Multiplying many probabilities (each < 1) produces extremely small numbers. To prevent underflow, Naive Bayes uses logarithms:

$$\log P(\text{Spam} \mid X) \propto \log P(\text{Spam}) + \sum_{i=1}^{n} \log P(w_i \mid \text{Spam}).$$

This converts multiplication into addition, making computations stable and efficient.

Step 6: Final Prediction Rule

For two classes (Spam vs. Ham), predict **Spam** if:

$$\log P(\text{Spam}) + \sum_{i=1}^{n} \log P(w_i \mid \text{Spam}) > \log P(\text{Ham}) + \sum_{i=1}^{n} \log P(w_i \mid \text{Ham}).$$

This is the version implemented in most real-world Naive Bayes classifiers.

Numerical Example

Return to a dataset with 100 emails, where 20 are spam and 80 are ham:

$$P(\text{Spam}) = 0.2, \qquad P(\text{Ham}) = 0.8.$$

Suppose the word "lottery" appears 14 times among spam emails and 10 times among ham:

$$P(\text{Lottery} \mid \text{Spam}) = 0.7, \qquad P(\text{Lottery} \mid \text{Ham}) = 0.125.$$

Similarly, assume:

$$P(\text{Winning} \mid \text{Spam}) = 0.75, \qquad P(\text{Winning} \mid \text{Ham}) = 0.1.$$

Applying Naive Bayes:

$$P(\text{Spam} \mid \text{Lottery}, \text{Winning}) = \frac{0.2 \cdot 0.7 \cdot 0.75}{0.2 \cdot 0.7 \cdot 0.75 + 0.8 \cdot 0.125 \cdot 0.1}.$$

Evaluating:

$$P(\text{Spam} \mid \text{Lottery}, \text{Winning}) \approx 0.913.$$

Thus, an email containing both "lottery" and "winning" is classified as spam with over 91% probability.

Naive Bayes remains one of the simplest yet most effective methods for text classification, especially in high-dimensional settings such as spam detection.

1.10 Why Probability Matters in Machine Learning

Up to this point, we have developed several tools from probability theory—conditional probability, Bayes' theorem, and independence. These concepts are not merely theoretical; they form the mathematical backbone of many machine learning models. In fact, a large class of machine learning problems can be stated as follows:

Compute the probability of a target outcome given observed features.

Examples Across Machine Learning

- **Spam Detection.** A classifier estimates the probability that an email is spam based on features such as the words it contains, the sender, or attached files. Mathematically, the model seeks:
$$P(\text{Spam} \mid \text{Features}).$$
- **Sentiment Analysis.** Given a sentence, the model evaluates:
$$P(\text{Positive Sentiment} \mid \text{Words in the Sentence}).$$
This is again a conditional probability problem, where the words act as evidence.
- **Image Recognition.** A model determines whether an image contains a specific object (e.g., a cat). The pixels are treated as features:
$$P(\text{Cat} \mid \text{Pixels}).$$
Modern deep learning architectures compute this probability implicitly, but the underlying formulation remains probabilistic.
- **Medical Decision Support.** Given demographics, symptoms, and medical history, a model estimates:
$$P(\text{Healthy} \mid \text{Patient Data}).$$
Again, this is a conditional probability that incorporates multiple forms of evidence.

Connection to Bayes' Theorem

In earlier examples, such as spam classification, we computed the posterior:

$$P(\text{Spam} \mid \text{Lottery}) = \frac{P(\text{Spam}, \text{Lottery})}{P(\text{Lottery})}.$$

Bayes' theorem allowed us to refine an initial belief (the prior) using new evidence (the features). This same structure appears throughout machine learning models, including Naive Bayes classifiers, logistic regression (via likelihood maximization), and even modern probabilistic deep learning methods.

General Perspective

When viewed abstractly, many machine learning tasks amount to building a model that estimates a conditional probability of the form:

$$P(\text{Target} \mid \text{Features}),$$

and using that probability to make predictions, decisions, or classifications. Whether the features are words, pixels, sensor measurements, or medical symptoms, probability provides a unified framework for reasoning under uncertainty.

2 Probability Distributions

2.1 Random Variables

One of the most fundamental ideas in probability theory—and in machine learning—is the concept of a **random variable**. Unlike the variables you may have encountered in algebra, which take a fixed deterministic value (e.g. $x = 3$), a random variable can take different values depending on the outcome of an underlying random experiment.

Examples of Random Variables

- The temperature recorded at noon each day.
- The number of heads obtained when flipping a fair coin 10 times.
- The time you wait at a bus stop until the next bus arrives.

To build intuition, consider a simple coin flip. Let X denote the number of heads obtained in a single toss. Then X can take only two values:

$$X = \begin{cases} 1, & \text{if the outcome is heads,} \\ 0, & \text{if the outcome is tails.} \end{cases}$$

Since the coin is fair,

$$P(X = 1) = 0.5, \qquad P(X = 0) = 0.5.$$

A Random Variable for Ten Coin Tosses

Now consider flipping the same coin 10 times. The random variable X now represents the total number of heads obtained. The possible values of X range from 0 to 10. For instance:

$$X = 10 \text{ if all tosses are heads,} \qquad X = 9 \text{ if exactly one toss is tails,} \quad \text{and so on.}$$

Because each toss is independent and the probability of heads is 0.5, the probability of a specific sequence (for example HHHHHHHHHH) is:

$$(0.5)^{10}.$$

But computing $P(X = k)$ for each $k = 0, 1, \ldots, 10$ requires grouping together *all* sequences that contain exactly k heads. Before deriving formulas, it is useful to estimate these probabilities through simulation.

Histogram from 500 Experiments

The figure below shows the empirical distribution of X when the 10-toss experiment is repeated 500 times. Blue and orange bars represent two independent simulation runs.

This histogram illustrates several important ideas:

- Outcomes with all heads ($X = 10$) or all tails ($X = 0$) occur very rarely.
- The most frequent outcome is around $X = 5$, which aligns with the intuition that in a fair coin, half of the tosses are heads on average.
- The distribution is roughly symmetric, reflecting the fairness of the coin.

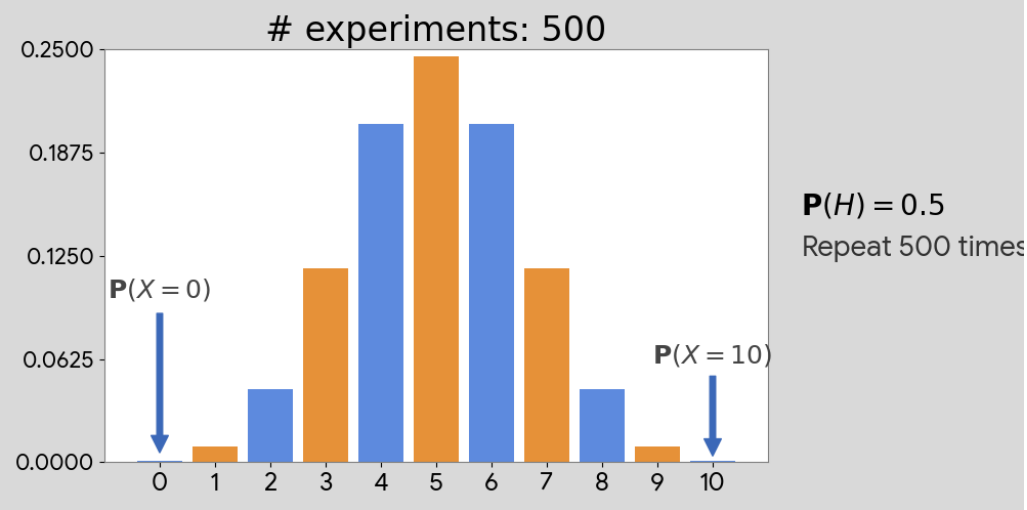

Figure 1: Empirical distribution of the number of heads in 10 fair-coin tosses, repeated 500 times. Extreme outcomes ($X = 0$ and $X = 10$) are rare, while $X = 5$ occurs most frequently.

Why Random Variables Matter

Random variables allow us to model entire experiments compactly. Instead of tracking individual coin-toss outcomes, dice rolls, patient diagnoses, or machine outputs, we focus on the numerical quantity of interest:

$$X = \text{“number of heads”}, \quad X = \text{“number of defective items”}, \quad X = \text{“waiting time”}, \quad \text{etc.}$$

Many real-world quantities in machine learning are naturally represented as random variables:

- The number of users clicking on an advertisement.
- The count of positive predictions in a classifier.
- The time until an event occurs (e.g., system failure).

Discrete vs. Continuous Random Variables

Random variables come in two broad types:

1. **Discrete random variables** take values from a countable set. Examples include:
   - number of heads in n coin tosses;
   - number of times a die shows a 1 in 20 rolls;
   - number of defective products in a shipment.
2. **Continuous random variables** take values from an interval of real numbers. Examples include:
   - waiting time for the next bus;
   - rainfall in millimetres;
   - the height of an athlete’s jump.

The key difference is not just the number of possible values. A discrete random variable may have infinitely many values (e.g., “flip a coin until the first head appears”). The core distinction is:

Discrete: countable values, Continuous: uncountable interval of values.

Deterministic vs. Random Variables

A deterministic variable always has a fixed value or follows a fixed rule, such as:

$$x = 2, \qquad f(x) = x^2.$$

In contrast, a random variable captures uncertainty and variability:

$$X = \text{number of heads in 10 tosses.}$$

This uncertainty is the reason random variables are fundamental in probability, statistics, and machine learning.

2.2 Probability Distributions

In the previous lessons, we focused on computing individual probabilities for specific outcomes. We now take a broader perspective. Suppose we list *all* possible outcomes of an experiment along the horizontal axis, and for each outcome, we assign the probability that it occurs. The collection of these probabilities forms a **probability distribution**.

Example: Three Coin Tosses

Consider tossing three fair coins. Let the random variable

$$X = \text{number of heads in three tosses.}$$

The possible values of X are $0, 1, 2, 3$. Although there are eight possible sequences of heads and tails, each value of X may arise from several different sequences:

- $X = 0$: only one outcome (TTT)
- $X = 1$: three outcomes (HTT, THT, TTH)
- $X = 2$: three outcomes (HHT, HTH, THH)
- $X = 3$: only one outcome (HHH)

Dividing each count by 8, we obtain the probability distribution of X, which explains why obtaining one or two heads is more likely than obtaining zero or three.

Example: Four Coin Tosses

For four fair coin tosses, the random variable X = number of heads takes values $0, 1, 2, 3, 4$. There are:

- 1 outcome with 0 heads,
- 4 outcomes with 1 head,
- 6 outcomes with 2 heads,
- 4 outcomes with 3 heads,
- 1 outcome with 4 heads.

Since there are 16 total outcomes, the probabilities are 1/16, 4/16, 6/16, 4/16, and 1/16 respectively. These can be displayed as a histogram, neatly summarizing the distribution of the random variable.

Example: Five Coin Tosses

Now consider X_3 = number of heads in five fair coin tosses. The possible values are 0 through 5, and the number of favorable outcomes for each is:

$$1,\ 5,\ 10,\ 10,\ 5,\ 1.$$

Since there are 32 total outcomes, the corresponding probabilities are

$$\frac{1}{32},\ \frac{5}{32},\ \frac{10}{32},\ \frac{10}{32},\ \frac{5}{32},\ \frac{1}{32}.$$

This distribution is shown below.

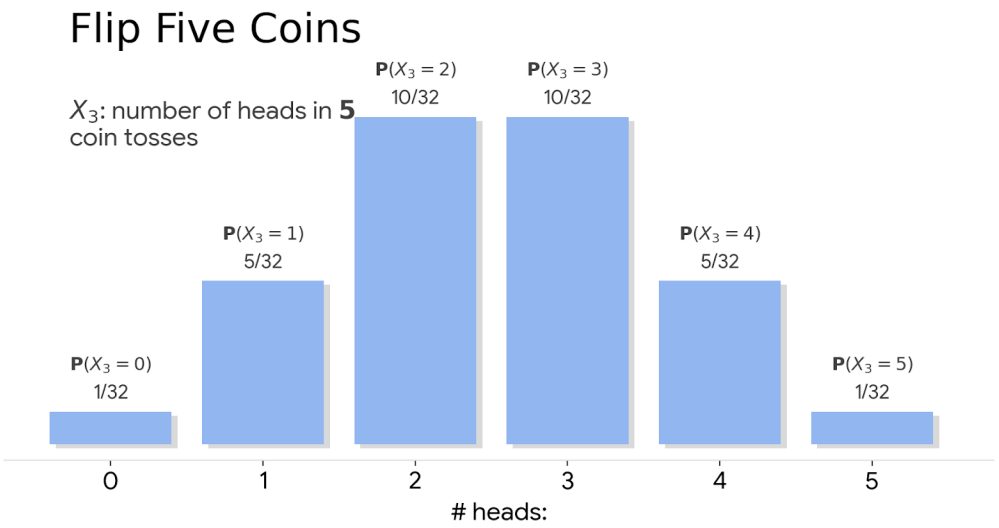

Figure 2: Probability distribution of X_3, the number of heads in five fair coin tosses. Each bar represents $P(X_3 = x)$ for $x = 0, \ldots, 5$.

Probability Mass Function (PMF)

For a discrete random variable X, the probability distribution is captured by its **probability mass function (PMF)**, denoted by $p(x) = P(X = x)$. A PMF must satisfy:

1. $p(x) \geq 0$ for all possible values of x,
2. $\sum_x p(x) = 1$.

The PMFs for the random variables counting heads in 3, 4, and 5 coin tosses all follow a similar pattern. These similarities raise a natural question:

> Is there a single mathematical model that represents all these distributions?

Indeed, such a model exists: the **binomial distribution**, which you will study in the next section.

2.3 The Binomial Distribution

We now turn to one of the most fundamental discrete probability models used throughout statistics and machine learning: the **binomial distribution**. A binomial distribution naturally arises when we repeat the same experiment multiple times under identical conditions—such as tossing a coin n times.

Motivating Example: Counting Heads

Suppose we flip a fair coin 10 times. The number of heads obtained can be any value in

$$0, 1, 2, \ldots, 10,$$

and each value occurs with a particular probability. Plotting these probabilities produces a familiar bell-shaped histogram. This distribution of probabilities is the *binomial distribution*.

Example: Probability of Obtaining Two Heads in Five Tosses

Consider flipping five fair coins. Each individual sequence of heads (H) and tails (T) has probability

$$\left(\frac{1}{2}\right)^5 = \frac{1}{32}.$$

To compute $P(X = 2)$, where X is the number of heads, we must count how many sequences contain exactly two heads. Each such sequence has the same probability, but there are multiple distinct arrangements:

$$\text{H H T T T}, \quad \text{H T H T T}, \quad \ldots$$

In fact, there are 10 such sequences. Therefore,

$$P(X = 2) = \frac{10}{32}.$$

Counting Combinations: The Binomial Coefficient

How do we compute the number of ways to arrange 2 heads and 3 tails? There are 5! ways to arrange five symbols, but we divide out permutations of identical symbols:

$$\binom{5}{2} = \frac{5!}{2!\,3!} = 10.$$

In general,

$$\binom{n}{k} = \frac{n!}{k!\,(n-k)!}$$

counts the number of sequences with k heads in n flips.

A useful symmetry property is:

$$\binom{n}{k} = \binom{n}{n-k},$$

because choosing k heads is equivalent to choosing the $n - k$ tail positions.

Deriving the Binomial PMF

Let X denote the number of heads obtained in n independent coin flips, where each flip has probability p of landing heads. To obtain exactly x heads:

- the probability of any single sequence with x heads is
$$p^x(1-p)^{n-x},$$
- there are $\binom{n}{x}$ such sequences.

Thus the probability mass function (PMF) is:

$$P(X = x) = \binom{n}{x} p^x(1-p)^{n-x}, \quad x = 0, 1, \ldots, n.$$

We say that

$$X \sim \text{Binomial}(n, p).$$

When $p = \frac{1}{2}$, the distribution is symmetric. For other values of p, the distribution skews toward smaller or larger numbers of heads.

Special Case: Five Tosses

For $n = 5$ and $p = \frac{1}{2}$, the PMF becomes:

$$P(X = x) = \binom{5}{x} \left(\frac{1}{2}\right)^5, \quad x = 0, \ldots, 5,$$

which corresponds to the histogram shown in the previous section.

Dice Experiments

The binomial model also applies to experiments that are not literally coin tosses. For example:

Example: What is the probability of obtaining exactly 3 ones when rolling a fair six-sided die five times?

Treat "rolling a 1" as "success" and any other value as "failure." Then:

$$p = P(\text{roll a } 1) = \frac{1}{6}, \qquad 1 - p = \frac{5}{6}.$$

Therefore,

$$P(X = 3) = \binom{5}{3} \left(\frac{1}{6}\right)^3 \left(\frac{5}{6}\right)^2.$$

Another Example: Ten Dice Rolls

If we roll a die 10 times and let X be the number of ones:

$$X \sim \text{Binomial}\left(10, \frac{1}{6}\right).$$

Thus the parameters are:

$$n = 10, \qquad p = \frac{1}{6} \approx 0.1666.$$

This simple model appears frequently in practice because many real-world problems can be reduced to counting the number of "successes" in repeated, independent trials.

2.4 Continuous Probability Distributions

Until now, we have studied **discrete** random variables—variables whose possible outcomes can be placed in a list. For example, if we flip a coin three times, the number of heads can only be

$$0, 1, 2, 3.$$

Similarly, the number of people in a town may be $0, 1, 2, \ldots$, and although the list is large, it is still a list.

However, not all random variables behave this way. Many quantities encountered in real life — such as time, height, distance, or temperature — vary over an *interval* rather than a countable set of distinct values. These are **continuous** random variables.

Discrete vs. Continuous Outcomes

A discrete random variable takes values that can be listed:

$$0, 1, 2, \ldots \quad \text{(countable set)}.$$

A continuous random variable takes values in an interval:

$$[a, b],\ (0, \infty),\ \text{etc.}$$

Such sets contain *uncountably many* numbers. For example, the waiting time for a bus could be:

$$1 \text{ minute},\ 1.01,\ 1.2237,\ \pi,\ \text{etc.}$$

This cannot be listed. Therefore, the probability model must change.

The Key Observation: Exact Values Have Probability Zero

Consider waiting for a call-center representative. What is the probability that your call is answered in *exactly* $1.000000\ldots$ minutes?

$$P(T = 1 \text{ minute}) = 0.$$

Because there are infinitely many possible times in any small interval, the probability of hitting one specific value must be zero. This is a defining characteristic of continuous random variables.

Modeling Probabilities Over Intervals

Since exact values have probability zero, we must instead ask questions such as:

$$P(0 \le T \le 1), \quad P(1 \le T \le 2), \quad \text{etc.}$$

Imagine dividing the possible waiting time (assume it is never more than five minutes) into intervals:

$$[0, 1],\ [1, 2],\ [2, 3],\ [3, 4],\ [4, 5].$$

Assign probabilities to each interval (represented as bars). The heights of these bars sum to 1, giving a discrete approximation of the continuous behavior.

Now, increase the resolution:

- Divide into 30-second intervals. - Then 15-second intervals. - Then smaller and smaller intervals.

As the intervals become narrower, the histogram becomes more detailed. In the limit—dividing into infinitely many infinitely thin intervals—the bars merge into a smooth curve. This curve represents a **continuous probability distribution**.

Area Instead of Sum

For discrete distributions:

$$\sum_x P(X = x) = 1.$$

For continuous distributions, the bars become a curve $f(x)$, and:

$$\int_{-\infty}^{\infty} f(x)\,dx = 1.$$

Here, $f(x)$ is the **probability density function** (PDF), and the area under the curve represents total probability. Probabilities are now computed as:

$$P(a \leq X \leq b) = \int_a^b f(x)\,dx.$$

This shift—from summing probabilities to integrating density—marks the conceptual transition from discrete to continuous probability models.

2.5 Probability Density Functions (PDFs)

When working with discrete random variables, each possible outcome has an associated probability. For example, in a sequence of ten coin tosses, the probability of obtaining exactly three heads is a specific number that we can calculate directly.

For continuous random variables, however, this idea breaks down. Consider the length of a phone call. The probability that a call lasts *exactly*

$$2.000000\ldots \text{ minutes}$$

is

$$P(T = 2) = 0.$$

This is because there are infinitely many possible durations within any interval, making the probability of hitting any single point exactly equal to zero. What *does* make sense is computing probabilities over intervals such as:

$$P(2 \leq T \leq 3), \qquad P(2 \leq T \leq 2.5).$$

These interval probabilities are encoded in a function called the **probability density function (PDF)**.

A Uniform Example

Suppose a call can be answered at any time between 0 and 5 minutes, and all times are equally likely. If we divide this range into five equal intervals, the probability of landing in any one of them is

$$\frac{1}{5} = 0.2.$$

Thus,

$$P(2 \leq T \leq 3) = 0.2.$$

If instead we divide the interval into ten subintervals of width 30 seconds, the probability of landing between 2 and 2.5 minutes is

$$\frac{1}{10} = 0.1.$$

Notice that although the *height* of the bars remains the same in both cases, the *area* changes because the width changes. This demonstrates an important point: for continuous distributions, probabilities correspond to *areas*, not heights.

Returning to a Non-Uniform Example

In more realistic settings, calls are not equally likely to have any duration. Suppose calls are more likely to last between 1–2 minutes or 2–3 minutes and far less likely to reach 4–5 minutes. The probability that a call lasts within an interval such as $[1, 2]$ is the area under the curve over that interval.

If we refine the intervals further—making them smaller and smaller—the histogram becomes more detailed. In the limit, when intervals become infinitesimally small, the histogram transitions into a smooth curve. This is the **probability density function**.

Because a vertical line has zero area, the probability of the call lasting *exactly* 2 minutes is still:

$$P(T = 2) = 0.$$

Definition of the PDF

For a continuous random variable X, the PDF is a function $f_X(x)$ such that:

1. $f_X(x) \geq 0$ for all real x,
2. f_X is defined for all $x \in \mathbb{R}$,
3. The total area under the curve equals 1:

$$\int_{-\infty}^{\infty} f_X(x)\, dx = 1.$$

Probabilities are computed as:

$$P(a \leq X \leq b) = \int_a^b f_X(x)\, dx.$$

PMF vs. PDF: A Summary

- **Discrete random variables:** Take values in a finite or countable set and use a probability mass function (PMF). Each value has a nonzero probability:

$$P(X = x) = p(x).$$

- **Continuous random variables:** Take values in an interval of real numbers and use a probability density function (PDF). Individual points have zero probability:

$$P(X = x) = 0,$$

but probabilities over intervals are meaningful:

$$P(a \leq X \leq b) = \int_a^b f(x)\,dx.$$

Thus, PMFs *assign* probability to each value, while PDFs describe a *density* whose areas correspond to probabilities.

2.6 Cumulative Distribution Function (CDF)

Up to this point, you have learned how probability mass functions (PMFs) describe discrete random variables and how probability density functions (PDFs) describe continuous random variables. However, both of these require you to compute probabilities by either summing individual probabilities (discrete case) or by calculating areas under a curve (continuous case). This can be inconvenient, especially when the probability of interest begins at the lower endpoint of the distribution.

The **cumulative distribution function (CDF)** resolves this issue by directly providing the probability that a random variable takes a value less than or equal to a given number.

Motivation Through the Call Center Example

Consider again the call duration example. If we want the probability that a call lasts between 2 and 3 minutes, we must calculate the area under the PDF between 2 and 3. Instead, it is often easier to compute:

$$P(0 \leq T \leq 3) \quad \text{and} \quad P(0 \leq T \leq 2),$$

and then subtract them. This idea motivates the CDF.

Cumulative Probability in the Discrete Case

Suppose call durations in minutes can fall into the discrete bins $[0, 1), [1, 2), \ldots, [4, 5)$, with known probabilities. The cumulative probability up to 1 minute is simply the area of the first bar. The cumulative probability up to 2 minutes is the sum of the first two bars:

$$P(T \leq 2) = P(0 \leq T < 1) + P(1 \leq T < 2).$$

Continuing this process builds an increasing curve that:

- starts at 0,
- ends at 1,
- and has jumps at the values where probability mass exists.

Cumulative Probability in the Continuous Case

For continuous variables, the same idea applies—but using *areas*:

$$F_X(x) = P(X \leq x) = \int_{-\infty}^{x} f_X(t)\,dt.$$

This produces a smooth, continuous function because individual points have zero probability and therefore cannot create jumps.

The CDF always:

- starts at 0 (after including all probability to the far left),
- approaches 1 as $x \to \infty$,
- never decreases, since probabilities cannot be negative.

Formal Definition

For any random variable X—discrete, continuous, or mixed—the cumulative distribution function is defined as:

$$F_X(x) = P(X \leq x), \qquad x \in \mathbb{R}.$$

Important properties:

1. $0 \leq F_X(x) \leq 1$ for all x,
2. $F_X(x)$ is non-decreasing,
3. $\lim_{x \to -\infty} F_X(x) = 0$,
4. $\lim_{x \to +\infty} F_X(x) = 1$.

For discrete random variables, the CDF has jumps. For continuous random variables, the CDF is smooth.

Relationship Between PDF and CDF

If X is continuous with a PDF $f_X(x)$, then:

$$F_X(x) = \int_{-\infty}^{x} f_X(t)\, dt,$$

and conversely,

$$f_X(x) = \frac{d}{dx} F_X(x).$$

Thus the two functions encode the same information, but in different formats:

- The PDF is used when calculating probabilities over short intervals.
- The CDF is used when we want probabilities from the left endpoint up to a value.

2.7 The Normal Distribution

One of the most important continuous probability distributions in statistics and machine learning is the **normal distribution**. It appears naturally in a wide variety of real-world settings—human height, test scores, measurement errors, and even aggregated noise in machine learning models. Many algorithms, including linear regression and Naive Bayes variants, rely on the mathematical properties of the normal distribution.

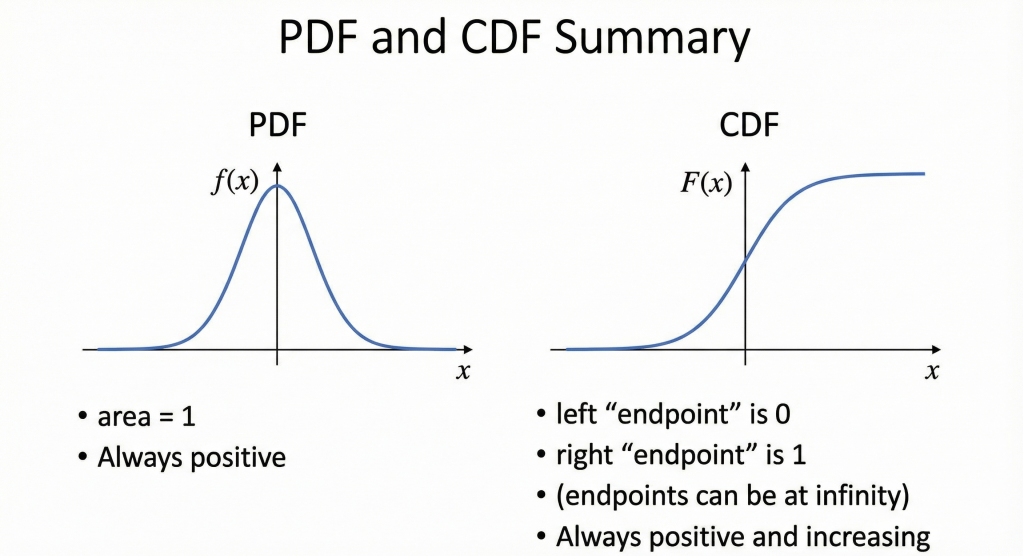

Figure 3: Summary of a probability density function (PDF) and the corresponding cumulative distribution function (CDF).

Definition

A continuous random variable X is said to follow a normal distribution with mean μ and standard deviation $\sigma > 0$ if its probability density function (PDF) is:

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right).$$

We write this as:

$$X \sim \mathcal{N}(\mu, \sigma^2).$$

Important characteristics:

- The distribution is **bell-shaped** and perfectly symmetric around μ.
- The mean μ determines the center of the curve.
- The standard deviation σ controls the spread.
- Larger σ means a wider and flatter curve; smaller σ means a narrower and taller curve.

Standard Normal Distribution

A special case is the **standard normal distribution**, where:

$$\mu = 0, \qquad \sigma = 1.$$

Its random variable is denoted by Z, and:

$$Z \sim \mathcal{N}(0, 1).$$

All normal distributions can be converted ("standardized") into the standard normal via:

$$Z = \frac{X - \mu}{\sigma}.$$

This is extremely useful because probability tables and numerical libraries use the standard normal CDF, denoted $\Phi(z)$.

The Empirical Rule (68–95–99.7 Rule)

A striking property of the normal distribution is that most values fall within a few standard deviations of the mean. The empirical rule states:

$$\begin{aligned} P(\mu - \sigma \leq X \leq \mu + \sigma) &\approx 68\%, \\ P(\mu - 2\sigma \leq X \leq \mu + 2\sigma) &\approx 95\%, \\ P(\mu - 3\sigma \leq X \leq \mu + 3\sigma) &\approx 99.7\%. \end{aligned}$$

This explains why the normal distribution is central in interpreting test scores, quality control, and noise modeling.

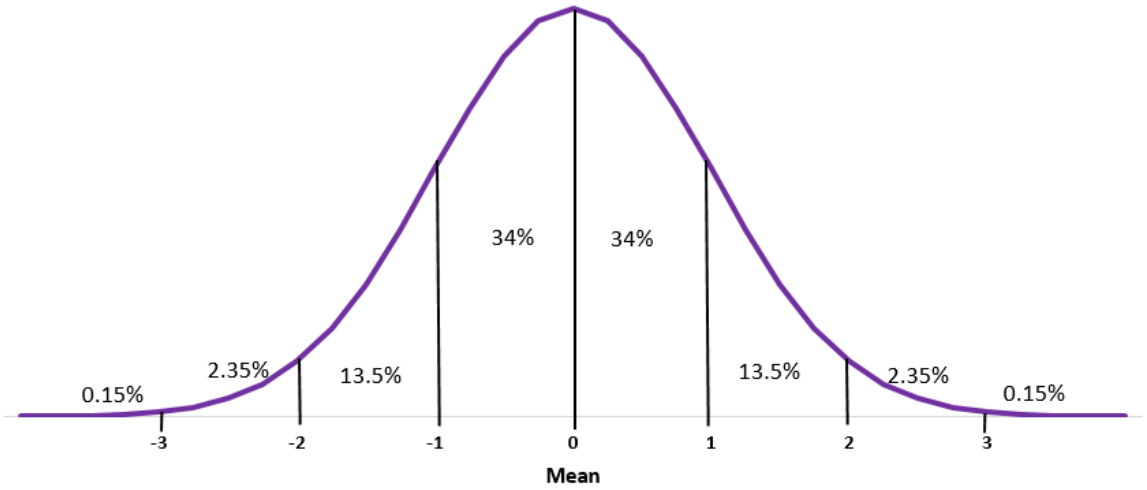

Figure 4: The empirical (68–95–99.7) rule for the normal distribution.

Why the Normal Distribution is Everywhere

The prevalence of the normal distribution is largely explained by the **Central Limit Theorem**. It states that when independent random effects add together—measurement noise, human behavior variability, environmental fluctuations—their sum tends to follow a normal distribution, regardless of the original individual distributions.

This is why:

- exam scores across large populations,
- model prediction errors,
- sensor readings,
- average outcomes of repeated processes

all tend to look approximately normal.

Connection to Machine Learning

In ML, the normal distribution appears in:

- **Gaussian Naive Bayes**, which assumes features follow a normal distribution within each class.
- **Regularization theory**, where Gaussian priors lead to L2 regularization.

- **Optimization noise modeling**, such as stochastic gradient descent.
- **Initialization of neural networks**, where weights are often drawn from normal distributions.

Its mathematical simplicity makes the Gaussian distribution a cornerstone in probabilistic modeling.

3 Describing Probability Distributions

A probability distribution tells us which outcomes are possible and how likely each outcome is. To understand the behaviour of a random variable, we need numerical summaries that describe its long–run average outcome and the amount of variation around that average. Two of the most fundamental quantities for this purpose are the **expected value** and the **variance**.

3.1 Expected Value

The expected value (or *mean*) of a random variable represents the value we would observe *on average* if an experiment were repeated many times. It is a probability–weighted average of all possible outcomes.

Intuition

The expected value is not the value you obtain in a single trial. Instead, it is the long–run average across a very large number of trials.

For example, the expected value of a fair die roll is 3.5. Even though 3.5 is impossible on any single roll, it accurately reflects the average value over many repetitions.

Definition (Discrete Random Variables)

Suppose a discrete random variable X takes values $x_1, x_2, \ldots, x_n$ with probabilities $p(x_i)$. The expected value is

$$\mathbb{E}[X] = \sum_{i=1}^{n} x_i \, p(x_i).$$

This formula simply multiplies each possible value by its probability and adds the results.

Example: Rolling a Fair Die

A fair die has six equally likely outcomes. The expected value is

$$\mathbb{E}[X] = \frac{1+2+3+4+5+6}{6} = 3.5.$$

Example: Number of Heads in 5 Tosses

Let X be the number of heads in 5 fair coin flips. Then $X \sim \text{Binomial}(5, 0.5)$, whose mean is

$$\mathbb{E}[X] = np = 5 \times 0.5 = 2.5.$$

Although one cannot obtain 2.5 heads, this is the long–run average over many such experiments.

Definition (Continuous Random Variables)

If a continuous random variable has a probability density function $f(x)$, then the expected value is

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} x\, f(x)\, dx.$$

The interpretation remains the same: a weighted average where the density provides the weights.

Interpretation

The expected value represents the *balance point* of a distribution. In skewed distributions, extreme values can pull the mean toward the tail, so the mean may differ substantially from the median.

Summary of Expected Value

- The expected value is the long–run average outcome of a random variable.
- Discrete case: $\mathbb{E}[X] = \sum x_i p(x_i)$.
- Continuous case: $\mathbb{E}[X] = \int x f(x)\, dx$.
- It represents the distribution's center of mass.

3.2 Variance

While the expected value describes the center of a distribution, it tells us nothing about how much the outcomes vary. Variance measures how far the values of a random variable spread out from the mean.

Motivation

Consider two random variables:

- Win or lose \$1 with equal probability.
- Win or lose \$100 with equal probability.

Both have expected value 0. Yet the second clearly exhibits much greater variability. Variance quantifies this difference in spread.

Definition

The variance of a random variable X is defined as

$$\boxed{\mathrm{Var}(X) = \mathbb{E}\Big[(X - \mathbb{E}[X])^2\Big].}$$

This is the expected squared deviation from the mean.

Why Square the Deviations?

If we simply averaged the deviations $X - \mathbb{E}[X]$, positive and negative values would always cancel. Squaring ensures that all deviations contribute positively and gives more weight to large deviations.

Alternative Formula

Variance is often easier to compute using the identity

$$\boxed{\mathrm{Var}(X) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2.}$$

This follows by expanding $(X - \mu)^2$ and applying linearity of expectation.

Example: Variance of a Fair Die

First compute

$$\mathbb{E}[X] = 3.5, \qquad \mathbb{E}[X^2] = \frac{1^2 + 2^2 + 3^2 + 4^2 + 5^2 + 6^2}{6} = \frac{91}{6}.$$

Then

$$\mathrm{Var}(X) = \frac{91}{6} - (3.5)^2 = 2.9167.$$

Example: Variance of a Binomial Variable

For $X \sim \mathrm{Binomial}(n, p)$,

$$\mathrm{Var}(X) = np(1 - p).$$

For 5 coin tosses:

$$\mathrm{Var}(X) = 5 \times 0.5 \times 0.5 = 1.25.$$

Summary of Variance

- Variance measures how spread out a distribution is.
- Defined as $\mathrm{Var}(X) = \mathbb{E}[(X - \mathbb{E}[X])^2]$.
- Equivalent formula: $\mathrm{Var}(X) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2$.
- Scaling by a multiplies variance by a^2; shifting by b has no effect.