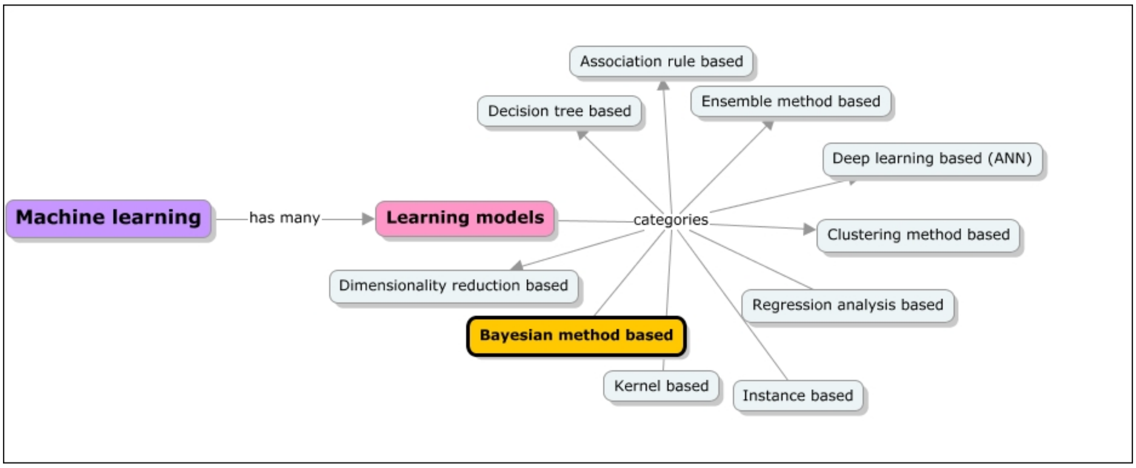
Chapter 9. Bayesian learning

In this chapter, we will go back to covering an important, statistical-based method of learning called the Bayesian method learning, and in particular, the Naïve Bayes algorithm among others. The statistical models generally have an explicit probability model, which reveals the probability of an instance belonging to a particular class rather than just classification while solving a classification problem. Before taking a deep dive into the Bayesian learning, you will learn some important concepts under statistics such as probability distribution and the Bayes theorem which is the heart of Bayesian learning.

Bayesian learning is a supervised learning technique where the goal is to build a model of the distribution of class labels that have a concrete definition of the target attribute. Naïve Bayes is based on applying Bayes' theorem with the *naïve* assumption of independence between each and every pair of features.

You will learn the basics and advanced concepts of this technique and get hands-on implementation guidance in using Apache Mahout, R, Julia, Apache Spark, and Python to implement the means - clustering algorithm.

The following figure depicts different learning models covered in this book, and the technique highlighted will be dealt with in detail in this chapter:



The topics listed here are covered in depth in this chapter:

- An overview of Bayesian statistics and core principles or concepts of probability, distribution, and other relevant statistical measures
- Bayes' theorem and its mechanics
- Deep dive into the Naïve Bayes algorithm and variations of Naïve Bayes classifiers like multinomial and Bernoulli classifiers
- A detailed explanation of some real-world problems or use cases that the Bayesian learning technique can address
- Sample implementation using Apache Mahout, R, Apache Spark, Julia, and Python (scikit-learn) libraries and modules

Bayesian learning

Under supervised learning techniques, the learning models that are categorized under statistical methods are instance-based learning methods and the Bayesian learning method. Before we understand the Bayesian learning method, we will first cover an overview of concepts of probabilistic modeling and Bayesian statistics that are relevant in the context of Machine learning. The core concepts of statistics are very deep, and what will be covered in the next few sections is primarily focused on equipping you with a basic understanding of the dynamic and diverse field of probabilistic Machine learning, which is sufficient to interpret the functioning of the Bayesian learning methods.

Statistician's thinking

The objective of statisticians is to answer questions asked by people from various domains using data. The typical engineering methods use some subjective/objective methods that do not require data to answer the questions. But, statisticians always look at the data to answer questions. They also incorporate variability (the probability that measurements taken on the exact quantity at two different times will slightly differ) in all their models.

Let's take an example: *was M.F. Hussain a good painter?* One method of answering this question measures the paintings based on some accepted norms (by the person or community) of the quality of paintings. The answer in such a case may be based on creative expression, color usage, form, and shape. *I believe M.F. Hussain is a good painter.* In this case, this response can be fairly subjective (which means that the response you get from one person can be very different from the response you get from another). The statistician's method of answering this is very different. They first collect the data from a sample of people who are considered experts in assessing the quality of paintings (university professors of art, other artists, art collectors, and more). Then, after analyzing the data, they will come up with a conclusion such as: "75% of the university professors of arts, 83% of the professional artists, and 96% of the art collectors from the data of 3000 participants of the survey (with equal number of participants from each category) opined that Mr. M.F. Hussain is a good painter". Hence, it can be stated that he is considered a good painter by most. Very obviously, this is a very objective measure.

Important terms and definitions

The following are the essential parameters and concepts that are used to assess and understand the data. They are explained as definitions in some cases and with examples and formulae in others. They are classified as "vocabulary" and "statistical quantities". You will come across some of these terms in the next sections of this chapter:

Term	Definition
Population	This is the universe of data. Typically, statisticians want to make a prediction about a group of objects (Indians, galaxies, countries, and more). All the members of the group are called the population.
Sample	Most of the times, it is not feasible to work on the entire population. So, statisticians collect a representative sample from the population and do all their calculations on them. The subset of the population that is chosen for the analysis is called a **sample**. It is always cheaper to compile the sample com-

Term	Definition
	pared to the population or census. There are several techniques to collect samples: • **Stratified sampling**: This is defined as the process of dividing the members of the population into homogeneous subgroups before sampling. Each subgroup should be mutually exclusive, and every element of the population should be assigned to a subgroup. • **Cluster sampling**: This method of sampling ensure n unique clusters where each cluster has elements with no repetition.
Sample size	This is an obvious dilemma that every statistician has been through. How big should be the size of the sample? The bigger the sample, the higher will be the accuracy. However, the cost of collection and analysis also rise accordingly. So, the challenge is to find an optimum sample size where the results are accurate, and the costs are lower.
Sampling Bias	Bias is a systematic error that impacts the outcome in some way. Sampling bias is a consistent error that arises due to the sample selection.
Variable	It is one of the measurements of the sample or population. If we are taking all the members of a class, then their age, academic background, gender, height, and so on, become the variables. Some variables are independent. This means they do not depend on any other variable. Some are dependent.
Randomness	An event is called random if its outcome is uncertain before it happens. An example of a random

Term	Definition
	event is the value of the price of gold tomorrow afternoon at 1 P.M.
Mean	It is equal to the sum of all the values in the sample divided by the total number of observations in the sample.
Median	Median is a midpoint value between the lowest and highest value of a data set. This is also called the second quartile (designated Q2) = cuts data set in half = 50th percentile. If there is no exact midpoint (that is, the observations in the sample are even), then the median is the average of the two points in the middle.
Mode	This is the most frequently occurring value of the variable. A data can be unimodal (single mode), or multimodal (frequent multiple values). If the data obeys normal distribution (about which you will learn later), the mode is obtained using the empirical formula: *mean – mode = 3 x (mean - median)*
Standard deviation	It is an average measure of how much each measurement in the sample deviates from the mean. Standard deviation is also called the standard deviation of the mean. $\sigma = \sqrt{\frac{1}{n-1}\sum_{i=1}^{i=n}\left(x_i - \overline{x}\right)^2}$

Probability

Before we start understanding the probability, let's first look at why we need to consider uncertainty in the first place. Any real-life action is always associated with the uncertainty of the result or the outcome. Let's take some examples; will I be able to catch the train on time today? Will the sales of our top-selling product continue to be in the top position this quarter? If I toss a coin, will I get a heads or tails? Will I be able to go to the airport in *t* minutes?

There can be many sources of uncertainty:

- Uncertainty due to lack of knowledge, as a result of insufficient data, incomplete analysis, and inaccurate measurements
- Otherwise, uncertainty can also be due to complexity, as a result of incomplete processing conditions

In the real world, we need to use probabilities and uncertainties to summarize our lack of knowledge and ability to predict an outcome.

Let's elaborate on the last previous example.

Can I go to the airport in 25 minutes? There could be many problems, such as incomplete observations on the road conditions, noisy sensors (traffic reports), or uncertainty in action, say a flat tire or complexity in modeling the traffic. To predict the outcome, there should definitely be some assumptions made, and we need to deal with uncertainty in a principled way; this is called **probability**. In short, probability is a study of randomness and uncertainty.

In probability, an experiment is something that can be repeated and has uncertainty in the result. A single outcome of an experiment is referred to as a single event, and an event is a collection of outcomes. A **sample space** probability is a list of all the possible outcomes of an experiment.

The probability of the event *E* is represented as *P(E)* and is defined as the likelihood of this event occurring.

NOTE

> The Probability of an Event P(E) = the number of ways an event can happen / the number of possible outcomes

For example, for a coin that is tossed, there are two possibilities: heads or tails.

The probability of heads is *P(H) = ½ = 0.5*

When a dice is thrown, there are six possibilities, which are 1, 2, 3, 4, 5, and 6.

The probability of 1 is *P(1) = 1/6 = 0.16667*

The probability of rolling any event, *E*, *P(E)*, must be between *0* and *1* (inclusive).

$0 \leq P(E) \leq 1$

The value of *0* for probability indicates that an event is impossible, and the value of *1* indicates the certainty of the event. If there are *n* events, then the summation of the probability of each event is *1*. This can be represented as:

If *S = {e1, e2, ….en}* then *P(e1) +P(e2)+…P(en) = 1*

There are many ways to determine the probability:

- **Classical method**: This is the method that we used to define probability in the previous section. This method requires equally likely outcomes. So, if an experiment has equally likely *n* events and there are *m* possibilities, the event *E* can then occur.
  P(E) = the number of ways the event E can occur / the number of possible outcomes = m/n.
  For example, a bag of chocolates contains five brown covered chocolates, six yellow covered chocolates, two red covered chocolates, eight orange covered chocolates, two blue covered chocolates, and seven green covered chocolates. Suppose that a candy is randomly selected. What is the probability of a candy being brown?
  P (B) = 5/30

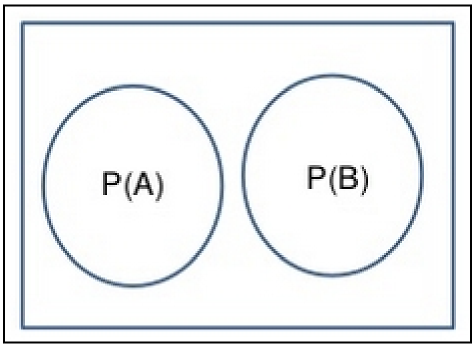
- **Empirical method**: The empirical method of probability computation is also called relative frequency, as this formula requires the number of times an experiment is repeated. This method defines the probability of the event *E*, which is the number of times an event is observed over the total number of times the experiment is repeated. The basis on which the probability is computed in this case is either observations or experiences.
  P(E) = Frequency of E / the number of trials of the experiment.
  For example, we want to compute the probability of a grad student to pick medicine as their major. We pick, let's say, a sample of 200 students and 55 of them pick medicine as majors, then:
  P(someone picking medicine) = 55/200 = 0.275
- **Subjective method**: This method of probability uses some fair and computed, or educated assumptions. It usually describes an individual's perception of the likelihood of an event to occur. This means the individual's degree of belief in the event is considered, and thus can be biased. For example, there is a 40% probability that the physics professor would not turn up to take the class.

Types of events

Events can be mutually exclusive, independent, or dependent in nature.

Mutually exclusive or disjoint events

Mutually exclusive events are the events that cannot happen at the same time. In short, the probability of the two events occurring at the same time is *0*. *P(1)* and *P(5)*. When a dice is rolled, there are mutually exclusive events. A Venn diagram representation of mutually exclusive events is depicted here:



For mutually exclusive events A and B the Addition rule is:

P(A or B) = P(A) + P(B)

For mutually exclusive events A and B the Multiplication rule is:

P(A and B) = P(A) x P(B)

Independent events

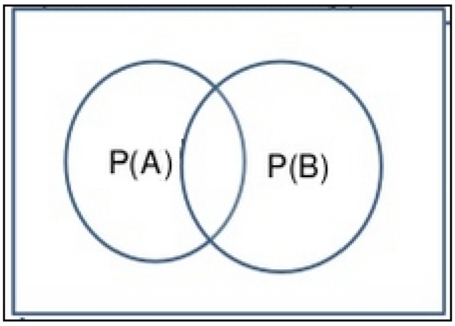
If the outcome of one event does not impact the outcome of another event, the two events are called independent events. For example, event A is that it rained on Sunday, and event B is the car having a flat tire. These two events are not related and the probability of one does not impact the other. An independent event can be mutually exclusive but not vice versa.

Multiplication rule in the case of independent events A and B is:

P(A and B) = P(A) x P(B)

Dependent events

Dependent events are the events where the occurrence of one event can influence the occurrence of another event. For example, a student who takes English as their first major can take political science as the second major. The Venn representation of dependent events is depicted here:



Addition rule for dependent event A and B is:

P(A or B) = P(A) + P(B) – P(A and B)

Multiplication rule for dependent event A and B is:

P(A and B) = P(A) x P(B)

Types of probability

In this section, we will take a look at the different types of probabilities, which are listed as follows:

- **Prior and posterior probability**: Prior probability is the probability that an event E occurs without any prior information or knowledge of any assumptions in the problem context.
  Let's take an example. If your friend was travelling by air and you were asked if they have a man or a woman as their neighbor, as the basis formula of probability works, there is a 0.5 (50%) probability that it can be a man and a 0.5 (50%) probability that it can be a woman. These values can change when more information is provided, and the probability that is measured then is called the posterior probability.
- **Conditional probability**: Conditional probability is defined as the probability that an event occurs, given another event already occurred. *P(B|A)* is interpreted as the probability of event B, given event A.
  For example, let's compute the probability that a person will be hit by a car while walking on the road. Let *H* be a discrete random variable describing the probability of a person being hit by a car, taking the hit as 1 and not as 0.
  Let *L* be a discrete random variable describing the probability of the cross traffic's traffic light state at a given moment, taking one from *{red, yellow, green}*:
  P(L=red) = 0.7,
  P(L=yellow) = 0.1,
  P(L=green) = 0.2.
  P(H=1|L=R) = 0.99,
  P(H|L=Y) = 0.9 and
  P(H|L=G) = 0.2.
  Using the conditional probability formulae, we get the following:
  *P(H=1 and L=R) = P(L=R)*P(H|L=R) = 0.693;*
  *P(H=1 and L=Y) = 0.1*0.9 = 0.09*
  Similarly, if the probability of getting hit while red is on is 0.99, the probability of not getting hit is 0.01. So, *P(H=0|L=R) = 0.01*. From these, we can compute the probability of *H=0* and *L=R*.

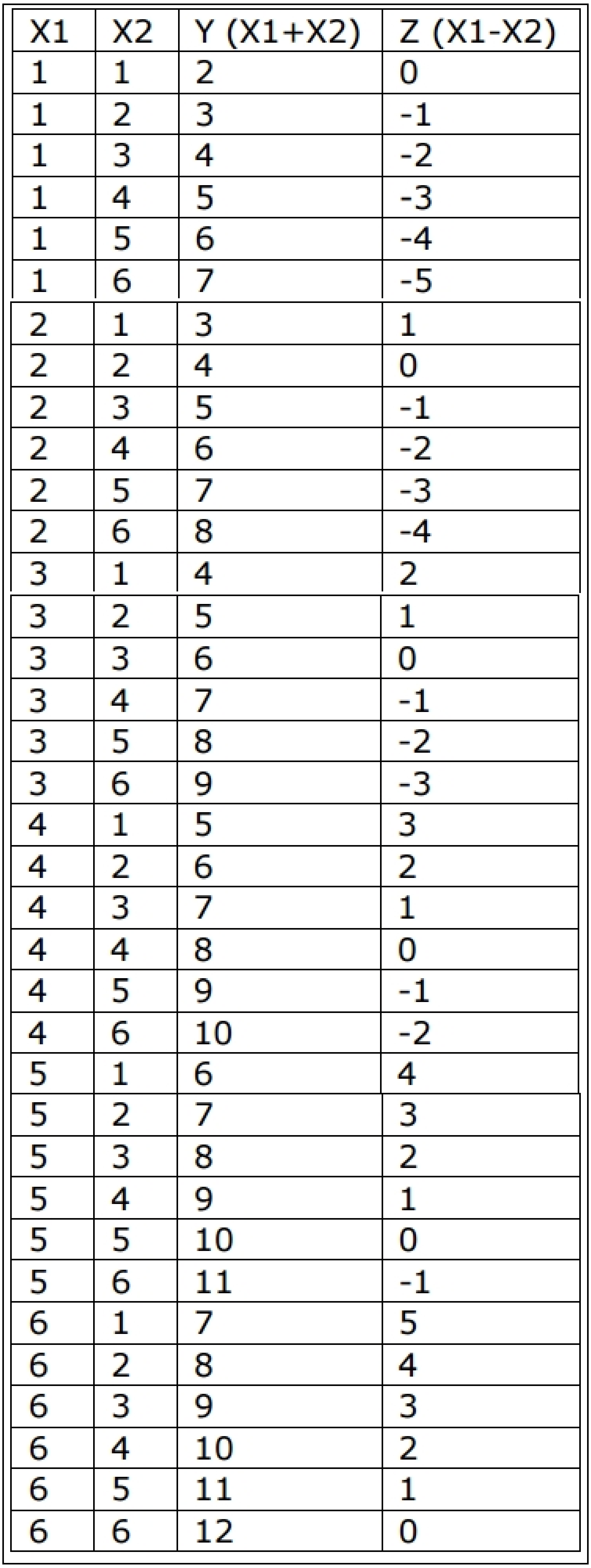
- **Joint probability**: Joint probability is the probability of two or more things happening together. In a two variable case, *f(x,y|θ)* is the joint probability distribution, where *f* is the probability of *x* and *y* together as a pair, given the distribution parameters—*θ*. For discrete random variables, the joint probability mass function is:
  P(X and Y) = P(X).P(Y|X) =P(Y).P(X|Y)
  You already saw this while studying the conditional probability. Since these are probabilities, we have the following:

  $$\sum_{x}\sum_{y} P\left(X = x \, and \, Y = y\right) = 1$$

- **Marginal probability**: Marginal probability is represented by *f(x|θ)* where *f* is the probability density of *x* for all the possible values of *y*, given the distribution parameters—*θ*. The marginal probability in a random distribution is determined from the joint distribution of *x* and *y* by summing over all the values of *y*. In a continuous distribution, it is determined by integrating over all the values of *y*. This is called **integrating out** the variable *y*. For discrete random variables, the marginal probability mass function can be written as *P(X = x)*. This is as follows:

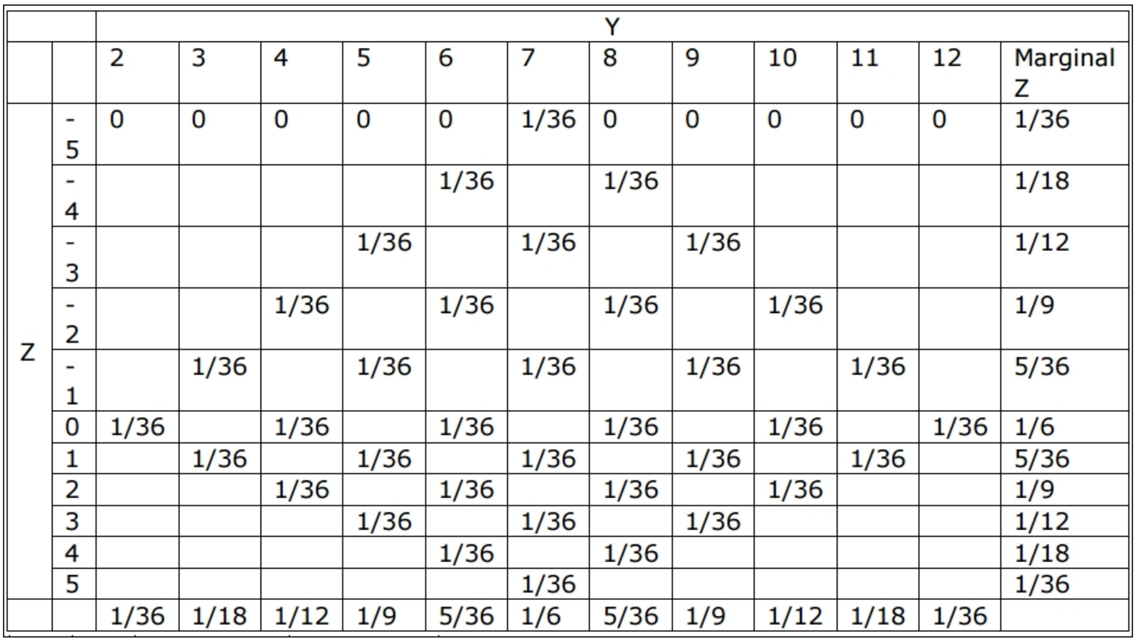
  $$P\left(X = x\right) = \sum_{y} P\left(X = x, Y = y\right) = \sum_{y} P\left(X = x \mid Y = y\right) P\left(Y = y\right)$$

  From the above equation, *P(X = x,Y = y)* is the joint distribution of *X* and *Y*, and *P(X = x|Y = y)* is the conditional distribution of *X*, given *Y*. The variable *Y* is marginalized out. These bivariate marginal and joint probabilities for discrete random variables are often displayed as two-way tables (as illustrated next). We will show the computations in a worked out problem in the next section.
  For example, suppose two dices are rolled, and the sequence of scores *(X1, X2)* is recorded. Let *Y=X1+X2* and *Z=X1−X2* denote the sum and difference of the scores respectively. Find the probability density function of *(Y, Z)*. Find the probability density function of *Y*. Find the probability density function of *Z*. Are *Y* and *Z* independent?
  Assuming that *X1* and *X2* are independent, they can take 36 possibilities, as shown in the table here:



X1	X2	Y (X1+X2)	Z (X1-X2)
1	1	2	0
1	2	3	-1
1	3	4	-2
1	4	5	-3
1	5	6	-4
1	6	7	-5
2	1	3	1
2	2	4	0
2	3	5	-1
2	4	6	-2
2	5	7	-3
2	6	8	-4
3	1	4	2
3	2	5	1
3	3	6	0
3	4	7	-1
3	5	8	-2
3	6	9	-3
4	1	5	3
4	2	6	2
4	3	7	1
4	4	8	0
4	5	9	-1
4	6	10	-2
5	1	6	4
5	2	7	3
5	3	8	2
5	4	9	1
5	5	10	0
5	6	11	-1
6	1	7	5
6	2	8	4
6	3	9	3
6	4	10	2
6	5	11	1
6	6	12	0

Let's now construct the joint, marginal, and conditional table. In this, we will have values of *Z* as rows and *Y* as columns. *Y* varies from 2 to 12 and *Z* varies from -5 to 5. We can fill all the conditional distributions just by counting. For example, take *Z=-1*; we see that this happens when *Y=3, 5, 7, 9, 11*. We also note that the probability of each one of them (say, the conditional probability that *Z=-1*, given *Y=3*) is *1/36*. We can fill the table like this for all the values:

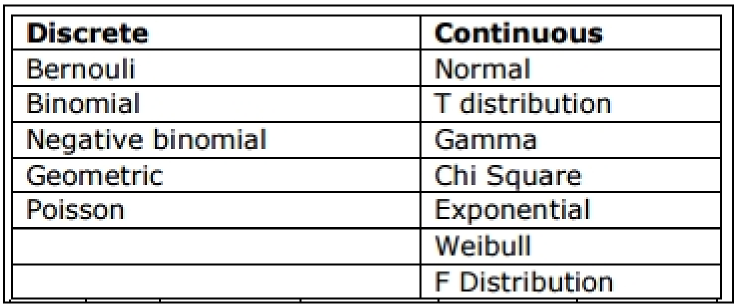


		Y											
		2	3	4	5	6	7	8	9	10	11	12	Marginal Z
Z	-5	0	0	0	0	0	1/36	0	0	0	0	0	1/36
	-4					1/36		1/36					1/18
	-3				1/36		1/36		1/36				1/12
	-2			1/36		1/36		1/36		1/36			1/9
	-1		1/36		1/36		1/36		1/36		1/36		5/36
	0	1/36		1/36		1/36		1/36		1/36		1/36	1/6
	1		1/36		1/36		1/36		1/36		1/36		5/36
	2			1/36		1/36		1/36		1/36			1/9
	3				1/36		1/36		1/36				1/12
	4					1/36		1/36					1/18
	5						1/36						1/36
		1/36	1/18	1/12	1/9	5/36	1/6	5/36	1/9	1/12	1/18	1/36	

So, the bottom row is the marginal distribution of *Y*. The right-most column is the marginal distribution of *Z*. The total table is the joint distribution. Clearly, they are dependent.

Distribution

Distributions are either discrete or continuous probability distributions, depending on whether they define probabilities associated with discrete variables or continuous variables:



Discrete	Continuous
Bernouli	Normal
Binomial	T distribution
Negative binomial	Gamma
Geometric	Chi Square
Poisson	Exponential
	Weibull
	F Distribution

We will cover a few of the previously mentioned distributions here.

In this section, our major emphasis is on modeling and describing a given property of the data. To understand how crucial this skill is, let's look at a few examples:

- A bank wants to look at the amount of cash withdrawn per transaction in an ATM machine over a period of time to determine the limits of transaction
- A retailer wants to understand the number of broken toys that he is getting in every shipment
- A manufacturer wants to understand how the diameter of a probe is varying between various manufacturing cycles
- A pharmaceutical company wants to understand how the blood pressures of millions of patients are impacted by its new drug

In all these cases, we need to come up with some precise quantitative description of how the observed quantity is behaving. This section is all about this. Anyway, intuitively, what do you think are the qualities that you would like to measure to gain an understanding?

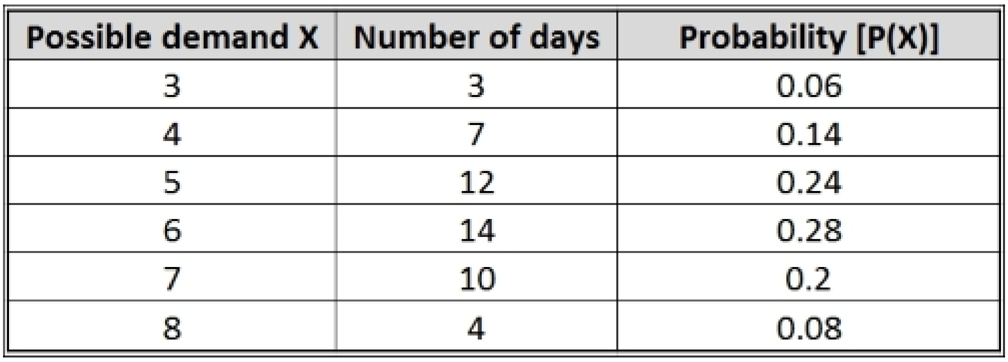
- What are all the values that a given variable is taking?
- What is the probability of taking a given value and what values have the highest probability?
- What is the mean/median, and how much is the variance?
- Given a value, can we tell how many observations fall into it and how many fall away from it?
- Can we give a range of values where we can tell 90% of the data lies?

Actually, if we can answer these questions, and more importantly if we develop a technique to describe such quantities, we are more or less unstoppable as far as this property is considered!

There are two prime observations to be made here. First, a property when distributed the way it is has all the qualities it takes to be a random variable (knowing one value of the quantity does not help us know the next value). Then, if we know the probability mass function or the distribution function of this random variable, we can compute all the previous matter. This is why it is so important to understand mathematics. In general, we follow (for that matter, almost anybody interested in analyzing the data that follows) a systematic process in describing a quantity:

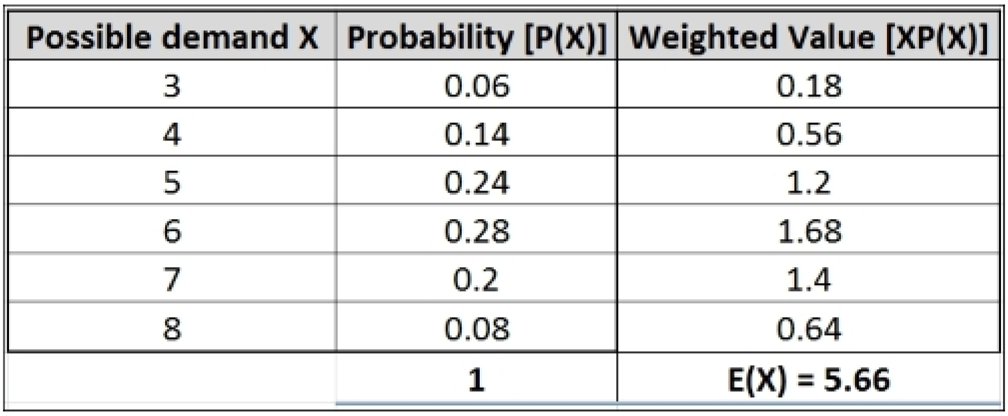
1. We will first understand the random variable.
2. Next, we will compute the probability mass (or distribution) function.
3. Then, we will predict the all-important parameters (mean and variance).
4. Then, we will check with experimental data to see how good our approximations are.

For example, the number of vans that have been requested for rental at a car rental agency during a 50-day period is identified in the following table. The observed frequencies have been converted into probabilities for this 50-day period in the last column of the table:



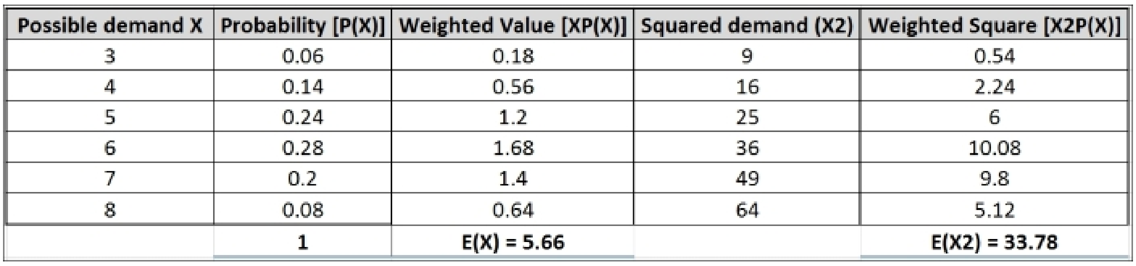
Possible demand X	Number of days	Probability [P(X)]
3	3	0.06
4	7	0.14
5	12	0.24
6	14	0.28
7	10	0.2
8	4	0.08

The expected value is 5.66 vans, as shown here:



Possible demand X	Probability [P(X)]	Weighted Value [XP(X)]
3	0.06	0.18
4	0.14	0.56
5	0.24	1.2
6	0.28	1.68
7	0.2	1.4
8	0.08	0.64
	1	**E(X) = 5.66**

Similarly, variance computation is given next:



Possible demand X	Probability [P(X)]	Weighted Value [XP(X)]	Squared demand (X2)	Weighted Square [X2P(X)]
3	0.06	0.18	9	0.54
4	0.14	0.56	16	2.24
5	0.24	1.2	25	6
6	0.28	1.68	36	10.08
7	0.2	1.4	49	9.8
8	0.08	0.64	64	5.12
	1	**E(X) = 5.66**		**E(X2) = 33.78**

The standard deviation is a square root of variance and is equal to 1.32 vans. Let's systematically analyze various distributions.

Bernoulli distribution

This is the simplest distribution that one can think of. Many a times, a property takes only discrete values; like a coin toss, a roll of the dice, the gender of people, and so on. Even if they are not exactly discrete, we can transform them by binning in some cases. For example, when we look at the net worth of individuals, we can redivide them as rich and poor (**discrete quantity**) based on the exact wealth they have (**continuous quantity**). Let's say that the probability of the property taking a given value is *p* (of course, the probability of it not taking is *(1-p)*). If we collect the large sample sufficiently, then how does the dataset look? Well, there will be some positives (where the variable took the value) and negatives (where the variable does not take the value). Assume that we denote positive with 1 and negative with 0.

Then, we have the following:

The mean = weighted average of probabilities = 1*p +0*(1-p) = p

Binomial distribution

This is an extension of the Bernoulli idea. Let's take a specific example. You are working in a population bureau and have the data of all the families in a state. Let's say you want to identify the probability of having two male children in families that have exactly two children. As you can see, a family can have two children in only four different ways: MM, MF, FM, and FF. As we consider having a male child as the event of interest, then the probability that there are only male children is *0.25 (1/4)*. The probability of there being one male child is *0.5 (0.25+0.25) (1/4+1/4),* and no male child is *0.25 (1/4)*.

So, if you look at 100 families, what is the probability that 20 families have exactly two male children? We will come to the solution later. Let's extend the argument to find the probability of having all the males in families with three children: The total possibilities are FFF, FFM, FMF, FMM, MFM, MMF, MFF, and MMM (eight total possibilities). The probability for all three to be male is *1/8*. The probability for two of the three being male is *3/8*. The probability for one of the three to be male is *3/8*. The probability for none to be male is *1/8*. Note that the total probability of all the events is always equal to 1.

Poisson probability distribution

Now, let's try to extend the Binomial theorem to infinite trials, but with a catch. The examples that we have taken (coin toss and more) have an interesting property. The probability of the event occurring in a trial does not change even if we increase the number of trials. However, there are a great number of examples, whereas the number of trials (or its equivalent) increases, the corresponding probability of the event decreases. So, we need to reduce the time interval to zero, or the number of observations to infinity to ensure that we see only a single success or failure in any trial. In this limiting case, the probability that we see *r* successes in *n* observations can be computed as follows:

$$\lim_{n\to\infty} c_r^n \left(\frac{\lambda}{n}\right)^r \left(1-\frac{\lambda}{n}\right)^{n-r}$$

$$\lim_{n\to\infty} \frac{n!}{r!(n-r)!}\frac{\lambda^r}{n^r}\left(1-\frac{\lambda}{n}\right)^{n-r}$$

The probability distribution of a Poisson random variable *X* is as given below. This considers representing the number of successes occurring in a given time interval:

$$P(X=r)=\frac{\lambda^r e^{-\lambda}}{r!}$$

Here, *r* is the $r^{\mathbf{th}}$ trial and *λ* = a mean number of successes in the given time interval or the region of space.

Exponential distribution

Let's now look at the Poisson example and ask ourselves a different question. What is the probability that the inspector does not see the first car until *t* hours? In this case, it may not be relevant, but when we work on the failure of a component, it makes sense to understand what time the probability of not seeing the failure is high. So, let's say the sighting of the car (or first failure) follows the Poisson process. Then, let's define *L*, a random variable that is the probability that the inspector will not see the first car until time *t* as the time before the first sighting of the car. From the Poisson distribution, the probability that she will not see the first car in 1 hour is as follows:

$$P(X=0)=\frac{e^{-\lambda}\lambda^{0}}{0!}=e^{-\lambda}$$

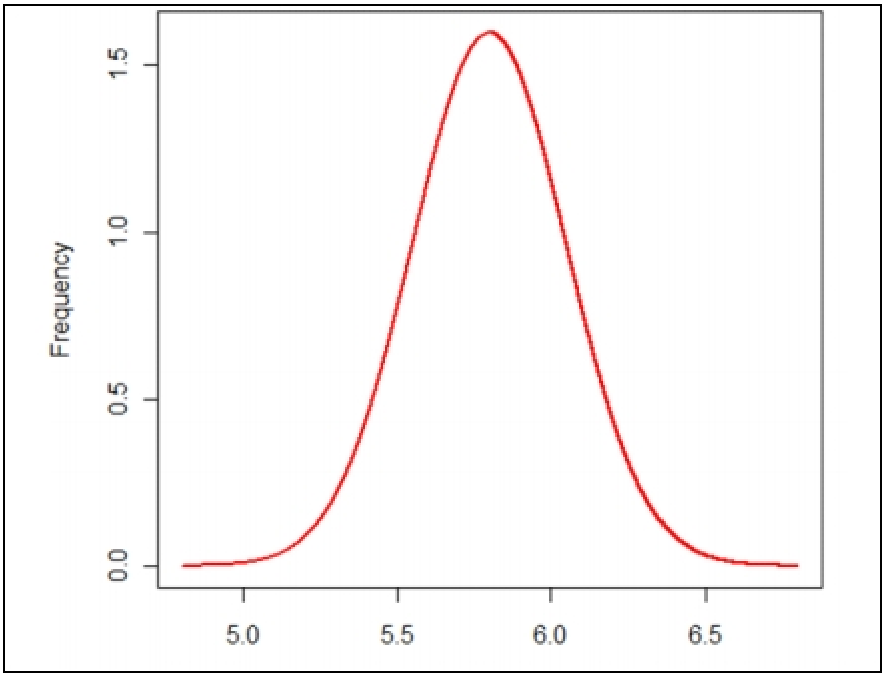
The probability that that she will not see a car in the second hour also is the same, and the probability that she will not see the car in *t* hours is $e^{-\lambda t}$ *($e^{-\lambda}$ * $e^{-\lambda}$ *…times)*. The probability that she will see the car in the first *t* hours then is $1\text{-}e^{-\lambda t}$.

The applications of exponential distribution are as follows:

- Time to the first failure in a Poisson process
- Distance of the dispersion of seeds from the parent plant
- The expected lifetime of an organism, ignoring the aging process (where the end occurs due to accidents, infections, and more)

Normal distribution

Normal distribution is a very widely used class of continuous distribution. It is also often called the bell curve because the graph of its probability density resembles a bell. Most of the real-life data such as weights, heights, and more (particularly when there are large collections) can be well approximated by a normal distribution.



Once we know the values of the heights, the number of samples that have this value can be mathematically described as follows:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}$$

Here, σ is the standard deviation and μ is the mean. To describe a normal distribution, we just need to know two concepts (average and SD).

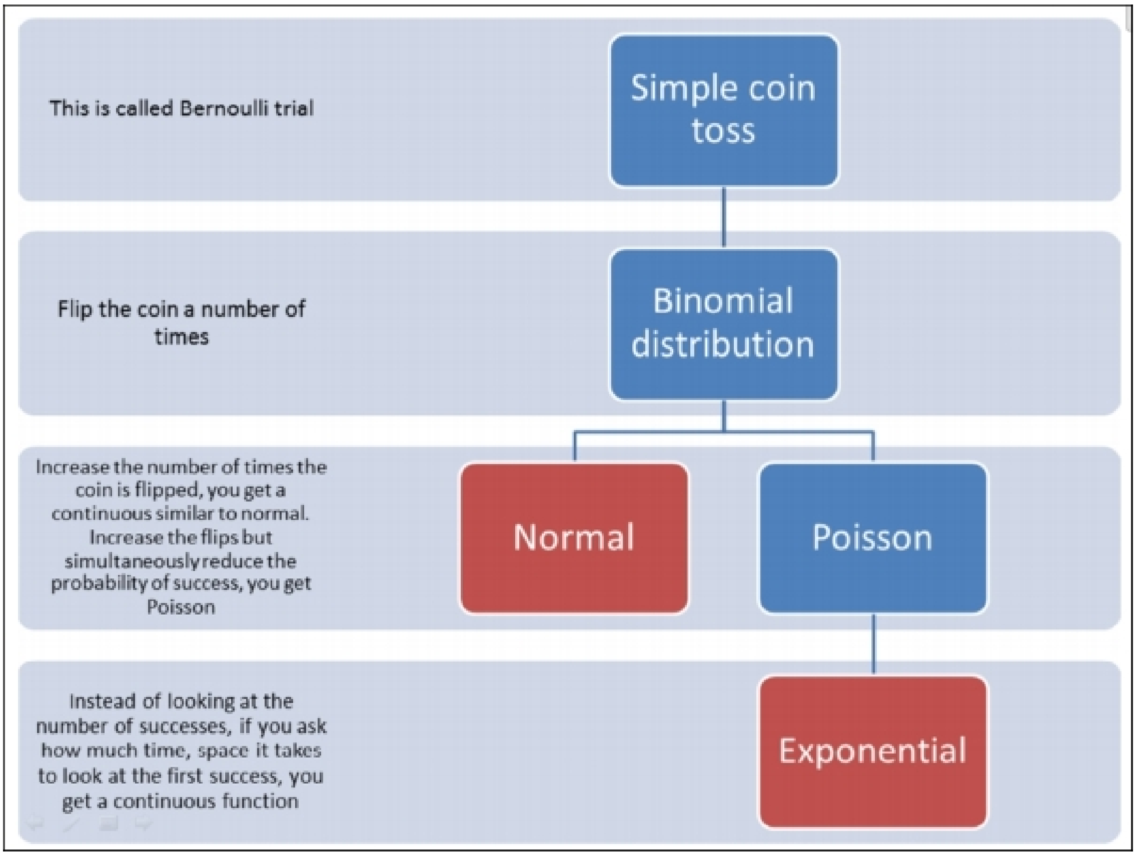
Every normal curve adheres to the following *rule*:

- About 68% of the area under the curve falls within one standard deviation of the mean
- About 95% of the area under the curve falls within two standard deviations of the mean
- About 99.7% of the area under the curve falls within three standard deviations of the mean

Collectively, these points are known as the **empirical rule** or the **68-95-99.7 rule**.

Relationship between the distributions

While we know that more or less everything converges to a normal distribution, it is best to understand where each one fits. The following chart helps in this:



This is called Bernoulli trial

Flip the coin a number of times

Increase the number of times the coin is flipped, you get a continuous similar to normal. Increase the flips but simultaneously reduce the probability of success, you get Poisson

Instead of looking at the number of successes, if you ask how much time, space it takes to look at the first success, you get a continuous function

Bayes' theorem

Before we go into the Bayes' theorem, we mentioned at the beginning of this chapter what is at the Bayesian learning is the Bayes theorem.

Let's start with an example. Assume that there are two bowls of nuts; the first bowl contains 30 cashew nuts and 10 pistachios and the second bowl contains 20 of each. Let's choose one bowl randomly and pick a nut with eyes closed. The nut is cashew. Now, what is the probability that the bowl chosen is the first bowl? This is a conditional probability.

So, *p(Bowl 1|cashew)* or the probability that it is bowl 1, given the nut is cashew, is not an easy and obvious one to crack.

If the question was to put the other way, *p(cashew|bowl1)* or the probability that the nut is cashew, given bowl 1 is easy, *p(cashew|Bowl 1) = ¾*.

As we know, *p(cashew|Bowl 1)* is not the same as *p(Bowl 1|cashew)*, but we can use one value to get another value, and this is what Bayes' theorem is all about.

The first step of defining the Bayes' theorem conjunction is commutative; following are the steps:

p (A and B) =p (B and A),

Further, the probability of A and B is the probability of A and the probability of B, given A:

p (A and B) = p (A) p (B|A), similarly

p (B and A) = p (B) p (A|B)

so,

p (A) p (B|A) = p (B) p (A|B) and

$$p(A \mid B) = \frac{p(A)\,p(B \mid A)}{p(B)}$$

And that's Bayes' theorem!

It might not be very obvious, but it is a very powerful definition.

Now, let's apply this to solve the previous *nut* problem to find *p(bowl1 cashew)*, and we can derive it if we can get *p(cashew|bowl 1)*:

p (bowl1 cashew) = (p(bowl1) p(cashew|bowl1)) / p (cashew)

p (bowl1) = ½

p (cashew|bowl1) = ¾

p (cashew) = total cashews / total nuts (between bowl1 and bowl2) = 50/80 = 5/8

Putting it together, we have the following:

p (bowl1 cashew) = ((1/2) (3/4))/(5/8)= 3/5 = 0.6

The additional aspect that needs to be considered now is how to feature in the changes that come over time as the new data comes in. This way, the probability of a hypothesis can be measured in the context of the data at a given point in time. This is called the diachronic interpretation of the Bayes' theorem.

Following is the restating Bayes' theorem with the hypothesis (*H*) for the given data (*D*):

$$p(H \mid D) = \frac{p(H)\,p(D \mid H)}{p(D)}$$

p (H) is the probability of the hypothesis *H* before seeing the data *D*.

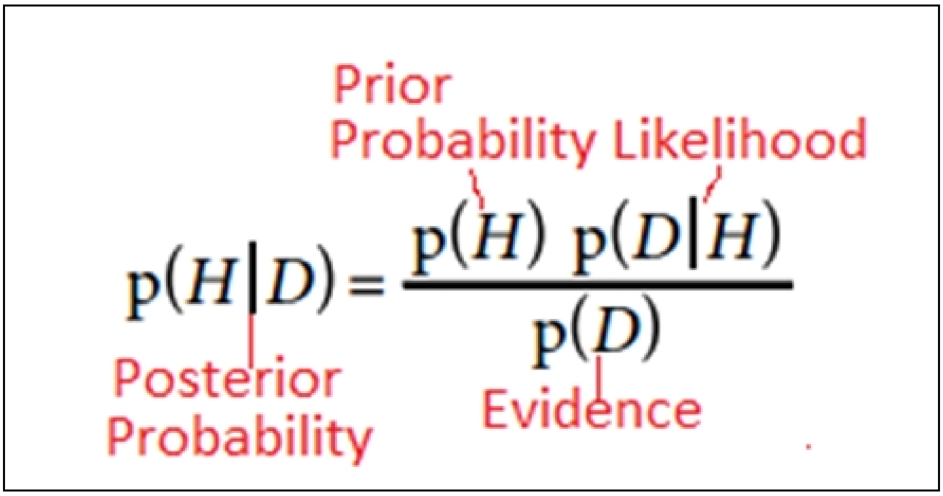
p (D) is the probability of data *D* under any hypothesis, which is usually constant.

p (H|D) is the probability of the hypothesis *H* after seeing the data *D*.

p (D|H) is the probability of data *D* given the hypothesis *H*.

NOTE

p (H) is called prior probability; *p (H|D)* is posterior probability; *p (D|H)* is the likelihood; and *p (D)* is the evidence:



Naïve Bayes classifier

In this section, we will look at the Naïve Bayes classifiers and how they are used to solve the classification problems. The Naïve Bayes classifier technique is based on the Bayes' theorem and assumes the predictors to be independent, which means knowing the value of one attribute does influence the value of any other attribute. The independence assumption is what makes Naïve Bayes *naïve*.

Naïve Bayes classifiers are easy to build, do not involve any iterative process, and work very well with large datasets. Despite its simplicity, Naïve Bayes is known to have often outperformed other classification methods.

We need to compute the probability of an assumption given a class.

That is, $P(x_1, x_2, ...x_{n|y})$. Obviously, there are multiple pieces of evidence represented by $x_1, x_2, ...x_n$.

Hence, we start with an assumption that $x_1, x_2, ...x_n$ are conditionally independent, given *y*. Another simple way of defining this is that we need to predict an outcome given multiple evidence as against a single evidence. To simplify, we uncouple these multiple pieces of evidence:

P(Outcome|Multiple Evidence) = [P(Evidence1|Outcome) x P(Evidence2|outcome) x ... x P(EvidenceN|outcome)] x P(Outcome) / P(Multiple Evidence)

This is also written as follows:

P(Outcome|Evidence) = P(Likelihood of Evidence) x Prior probability of outcome / P(Evidence)

In order to apply Naïve Bayes to predict an outcome, the previously mentioned formula will need to be run for every outcome. Just run this formula for each possible outcome, and in the case of a classification problem, the outcome will be a class. We will look at the famous fruit problem to help you understand this easily.

Given any three important characteristics of a fruit, we will need to predict what fruit it is. To simplify the case, let's take three attributes—long, sweet, and yellow; and three classes of fruit—banana, orange, and others.

Let there be 1,000 data points in the training set, and this is how the available information looks like:

Type	Long	Not long	Sweet	Not sweet	Yellow	Not yellow	Total
Banana	400	100	350	150	450	50	500
Orange	0	300	150	150	300	0	300
Others	100	100	150	50	50	150	200
Total	500	500	650	350	800	200	1000

Some derived values/prior probabilities from the previous table are as follows:

Probability of Class

p (Banana)= 0.5 (500/1000)

p (Orange)= 0.3

p (Others) = 0.2

Probability of Evidence

p (Long)= 0.5

p (Sweet)= 0.65

p (Yellow) = 0.8

Probability of Likelihood

p (Long|Banana) = 0.8

p (Long/Orange) = 0 P(Yellow/Other Fruit) =50/200 = 0.25

p (Not Yellow|Other Fruit)= 0.75

Now, given a fruit, let's classify it based on attributes. First, we run probability for each of the three outcomes, take the highest probability, and then classify it:

p (Banana|/Long, Sweet and Yellow) = p (Long|Banana) x p (Sweet|Banana) x p (Yellow|Banana) x p (banana) /p (Long) xp (Sweet) x. p (Yellow)

p (Banana||Long, Sweet and Yellow) =0.8 x 0.7 x 0.9 x 0.5 / p (evidence)

p (Banana||Long, Sweet and Yellow) =0.252/ p (evidence)

p (Orange||Long, Sweet and Yellow) = 0

p (Other Fruit/Long, Sweet and Yellow) = p (Long/Other fruit) x p (Sweet/Other fruit) x p (Yellow/Other fruit) x p (Other Fruit)

= (100/200 x 150/200 x 50/150 x 200/1000) / p (evidence)

= 0.01875/ p (evidence)

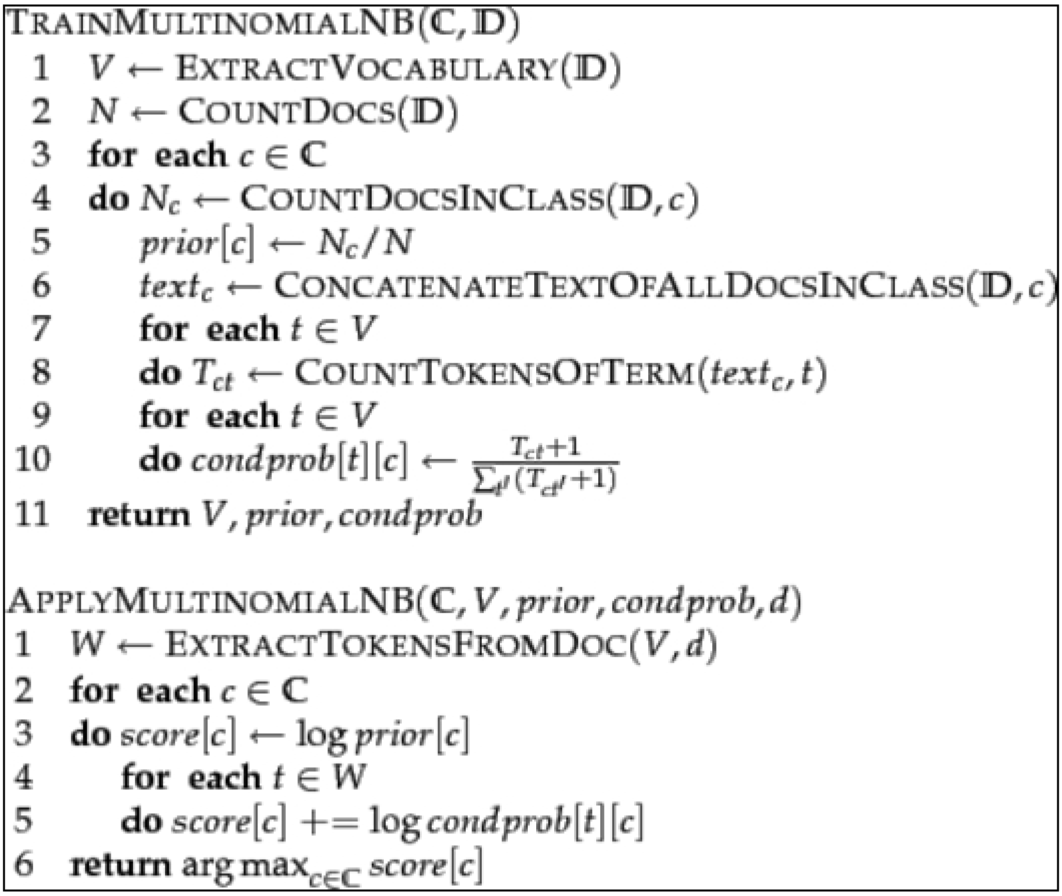
With the largest margin of *0.252 >> 0.01875*, we can now classify this Sweet/Long/Yellow fruit as likely to be a *Banana*.

As Naïve Bayes assumes a gaussian distribution for each of the features, it is also called the Gaussian Naïve Bayes classifier.

Naïve Bayes is particularly good when there is missing data. In the next sections, let's look at different types of Naïve Bayes classifiers.

Multinomial Naïve Bayes classifier

As we have seen in the previous section, Naïve Bayes assumes independence of the model against the distribution for a feature. In the case of a multinomial Naïve Bayes, the $p(x_i|y)$ is a multinomial distribution; in short, a multinomial distribution is assumed for each of the features. The case that fits this variant is that of a document where we need to compute the word count. A simple algorithm of multinomial Naïve Bayes is given here:

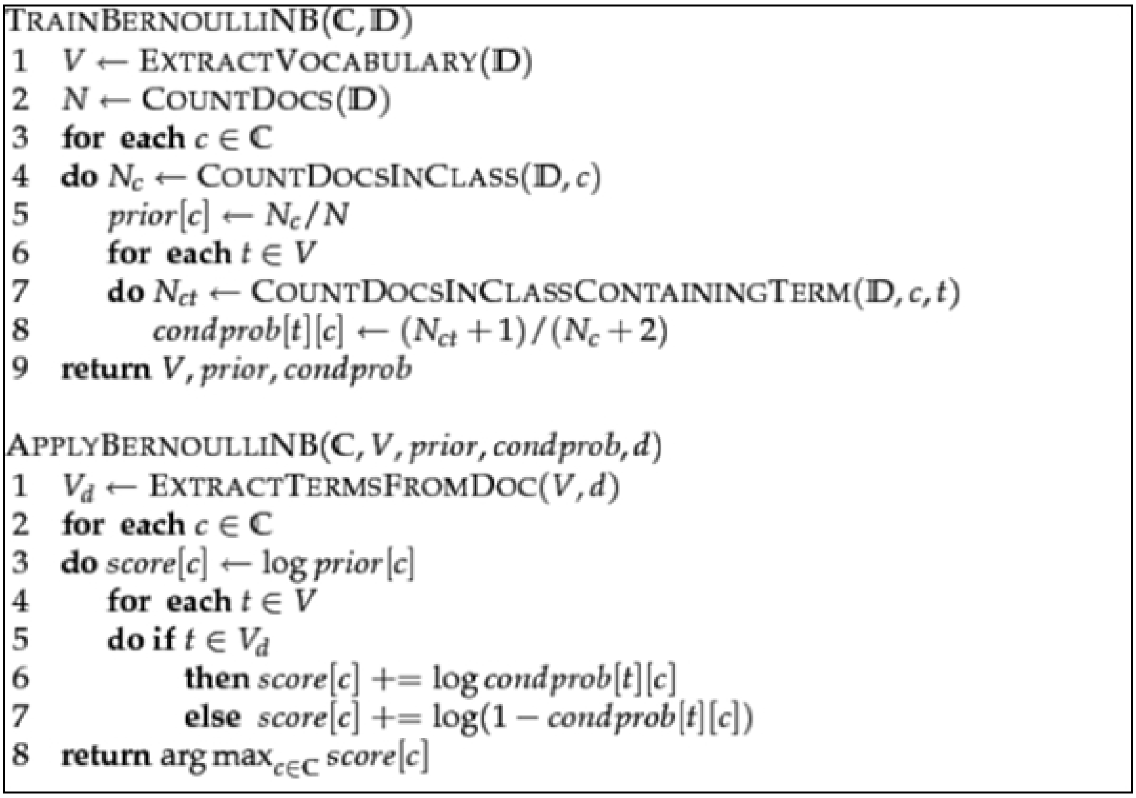


```
TRAINMULTINOMIALNB($\mathbb{C}$, $\mathbb{D}$)
  $V \leftarrow$ EXTRACTVOCABULARY($\mathbb{D}$)
  $N \leftarrow$ COUNTDOCS($\mathbb{D}$)
  for each $c \in \mathbb{C}$
  do $N_c \leftarrow$ COUNTDOCSINCLASS($\mathbb{D}, c$)
      $prior[c] \leftarrow N_c / N$
      $text_c \leftarrow$ CONCATENATETEXTOFALLDOCSINCLASS($\mathbb{D}, c$)
      for each $t \in V$
      do $T_{ct} \leftarrow$ COUNTTOKENSOFTERM($text_c, t$)
      for each $t \in V$
      do $condprob[t][c] \leftarrow \frac{T_{ct}+1}{\sum_{t'}(T_{ct'}+1)}$
  return $V, prior, condprob$

APPLYMULTINOMIALNB($\mathbb{C}, V, prior, condprob, d$)
  $W \leftarrow$ EXTRACTTOKENSFROMDOC($V, d$)
  for each $c \in \mathbb{C}$
  do $score[c] \leftarrow \log prior[c]$
      for each $t \in W$
      do $score[c] \mathrel{+}= \log condprob[t][c]$
  return $\arg\max_{c \in \mathbb{C}} score[c]$
```

The Bernoulli Naïve Bayes classifier

The Bernoulli Naïve Bayes classifier attaches a Boolean indicator to a word as one if it belongs to a document under examination and zero if it does not. The focus of this variation is that it considers the count of occurrence or non-occurrence of a word in a specific document under consideration. The non-occurrence of a word is an important value as it is used in the computation of the conditional probabilities of the occurrence of a word. The Bernoulli Naïve Bayes algorithm is detailed here:



```
TrainBernoulliNB($\mathbb{C}, \mathbb{D}$)
  $V \leftarrow$ ExtractVocabulary($\mathbb{D}$)
  $N \leftarrow$ CountDocs($\mathbb{D}$)
  for each $c \in \mathbb{C}$
  do $N_c \leftarrow$ CountDocsInClass($\mathbb{D}, c$)
      $prior[c] \leftarrow N_c / N$
      for each $t \in V$
      do $N_{ct} \leftarrow$ CountDocsInClassContainingTerm($\mathbb{D}, c, t$)
          $condprob[t][c] \leftarrow (N_{ct} + 1)/(N_c + 2)$
  return $V, prior, condprob$

ApplyBernoulliNB($\mathbb{C}, V, prior, condprob, d$)
  $V_d \leftarrow$ ExtractTermsFromDoc($V, d$)
  for each $c \in \mathbb{C}$
  do $score[c] \leftarrow \log prior[c]$
      for each $t \in V$
      do if $t \in V_d$
            then $score[c]$ += $\log condprob[t][c]$
            else $score[c]$ += $\log(1 - condprob[t][c])$
  return $\arg\max_{c \in \mathbb{C}} score[c]$
```

	Multinomial Naïve Bayes	Bernoulli Naïve Bayes
Model Variable	Here, a token is generated and checked for occurrence in a position	Here, a document is generated and checked for occurrence in a document
Document	$d = \langle t_1, \ldots, t_k, \ldots, t_{n_d} \rangle, t_k \in V$	$d = \langle e_1, \ldots, e_i, \ldots, e_M \rangle,$ $e_i \in \{0,1\}$
Estimation of the parameter	$\hat{P}(X = t \mid c)$	$\hat{P}(U_i = e \mid c)$
Rule		

	Multinomial Naïve Bayes	Bernoulli Naïve Bayes
	$\hat{P}(c)\prod_{1\le k\le n_d}\hat{P}(X=t_k \mid c)$	$\hat{P}(c)\prod_{t_i\in V}\hat{P}(U_i=e_i \mid c)$
Occurrences	This considers multiple occurrences	This considers single occurrences
Size of the document	Large documents are handled	Good with smaller documents
Features	This supports handling more features	This is good with lesser features
Estimation of a term	$\hat{P}(X=the \mid c)\approx 0.05$	$\hat{P}(U_{the}=1 \mid c)\approx 1.0$