Probability

AcadView

May 17, 2018

1 Overview

Probability theory is the study of uncertainty. Through this class, we will be relying on concepts from probability theory for deriving machine learning algorithms. These notes attempt to cover the basics of probability theory at a level appropriate for the course. In these notes, we provide a basic treatment of probability that does not address the finer details.

2 Elements of probability

In order to define a probability on a set we need a few basic elements,

- **Sample space** Ω: The set of all the outcomes of a random experiment. Here, each outcome $\omega \in \Omega$ can be thought of as a complete description of the state of the real world at the end of the experiment.

- **Set of events (or event space)** F**:** A set whose elements $A \in F$ (called events) are subsets of Ω

- **Probability measure:** A function $P : F \rightarrow R$ that satisfies the following properties,

  - $P(A) \geq 0$, for all $A \in F$
  - $P(\Omega) = 1$
  - If A_1,A_2,... are disjoint events (i.e.,$A_i \cap A_j = \emptyset$ whenever i 6= j), then

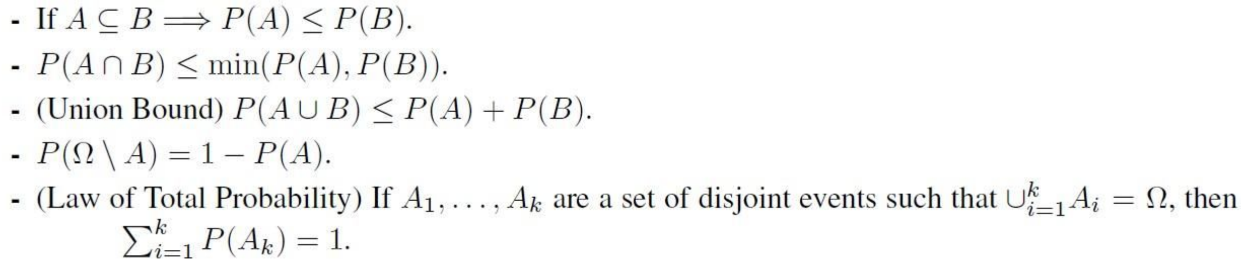
$$P(\cup_i A_i) = {}^{\mathrm{P}}{}_i P(A_i)$$

These three properties are called the Axioms of Probability.

Example: Consider the event of tossing a six-sided die. The sample space is Ω = {1,2,3,4,5,6} We can define different event spaces on this sample space. For example, the simplest event space is the trivial event space *F* = { ϕ,Ω}.

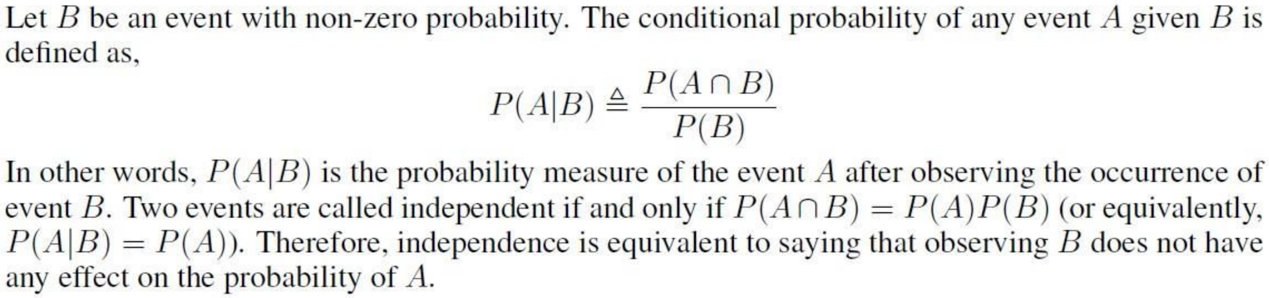
Another event space is the set of all subsets of Ω. For thefirst event space, the unique probability measure satisfying the requirements above is given by *P*(ϕ) = 0;*P*(Ω) = 1. For the second event space, one valid probability measure is to assign the probability of each set in the event space to be $\frac{i}{6}$ where *i* is the number of elements of that set; for example, $P(\{1,2,3,4\}) = \frac{4}{6}$ and $P(\{1,2,3\}) = \frac{3}{6}$.

Properties:



- If $A \subseteq B \Longrightarrow P(A) \leq P(B)$.
- $P(A \cap B) \leq \min(P(A), P(B))$.
- (Union Bound) $P(A \cup B) \leq P(A) + P(B)$.
- $P(\Omega \setminus A) = 1 - P(A)$.
- (Law of Total Probability) If $A_1, \ldots, A_k$ are a set of disjoint events such that $\cup_{i=1}^{k} A_i = \Omega$, then $\sum_{i=1}^{k} P(A_k) = 1$.

2.1 Conditional probability and independence



Let B be an event with non-zero probability. The conditional probability of any event A given B is defined as,

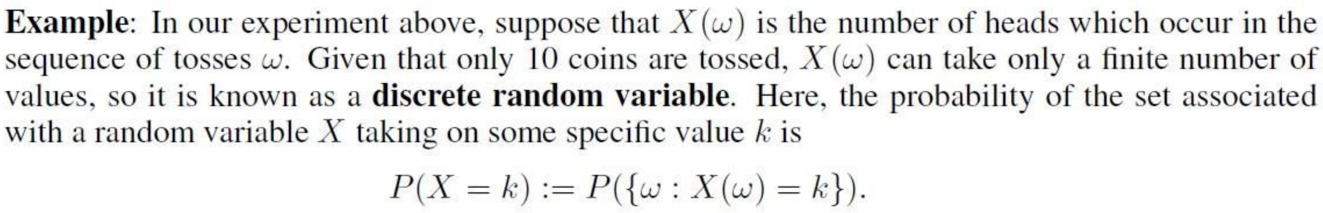
$$P(A|B) \triangleq \frac{P(A \cap B)}{P(B)}$$

In other words, $P(A|B)$ is the probability measure of the event A after observing the occurrence of event B. Two events are called independent if and only if $P(A \cap B) = P(A)P(B)$ (or equivalently, $P(A|B) = P(A)$). Therefore, independence is equivalent to saying that observing B does not have any effect on the probability of A.

3 Random variables

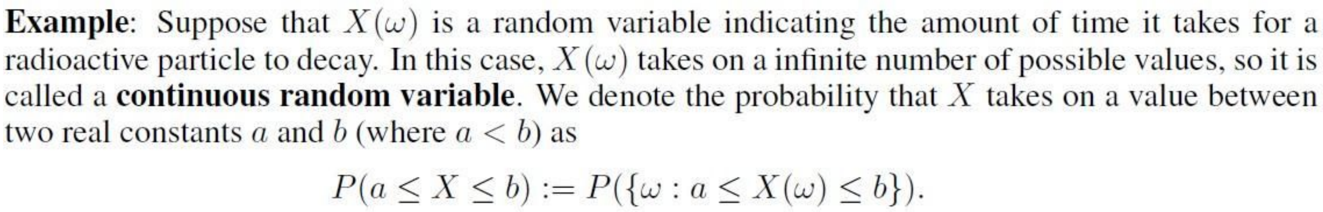
Consider an experiment in which we flip 10 coins, and we want to know the number of coins that come up heads. Here, the elements of the sample space Ω are 10-length sequences of heads and tails. For example, we might have w_0 = {*H*,*H*,*T*,*H*,*T*,*H*,*H*,*T*,*T*,*T*} ∈ Ω. However, in practice,we usually do not care about the probability of obtaining any particular sequence of heads and tails.

Instead we usually care about real-valued functions of outcomes, such as the number of heads that appear among our 10 tosses, or the length of the longest run of tails. These functions, under some technical conditions, are known as random variables.



Example: In our experiment above, suppose that $X(\omega)$ is the number of heads which occur in the sequence of tosses ω. Given that only 10 coins are tossed, $X(\omega)$ can take only a finite number of values, so it is known as a **discrete random variable**. Here, the probability of the set associated with a random variable X taking on some specific value k is

$$P(X = k) := P(\{\omega : X(\omega) = k\}).$$

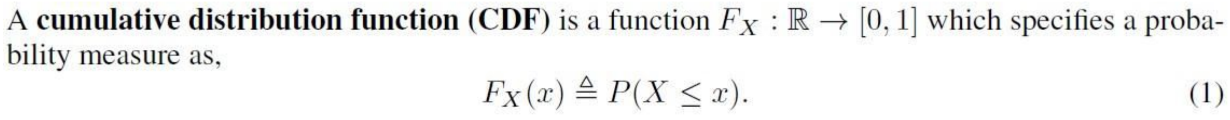


Example: Suppose that $X(\omega)$ is a random variable indicating the amount of time it takes for a radioactive particle to decay. In this case, $X(\omega)$ takes on a infinite number of possible values, so it is called a **continuous random variable**. We denote the probability that X takes on a value between two real constants a and b (where $a < b$) as

$$P(a \leq X \leq b) := P(\{\omega : a \leq X(\omega) \leq b\}).$$

3.1 Cumulative distribution functions

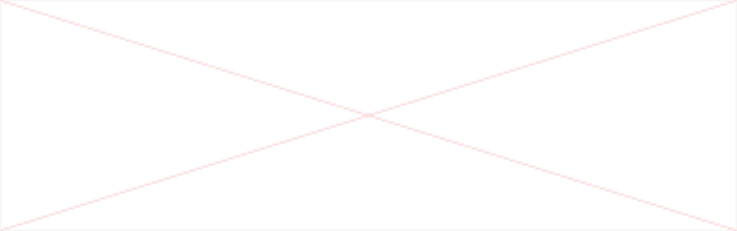
In order to specify the probability measures used when dealing with random variables, it is often convenient to specify alternative functions (CDFs, PDFs, and PMFs) from which the probability measure governing an experiment immediately follows. In this section and the next two sections, we describe each of these types of functions in turn.

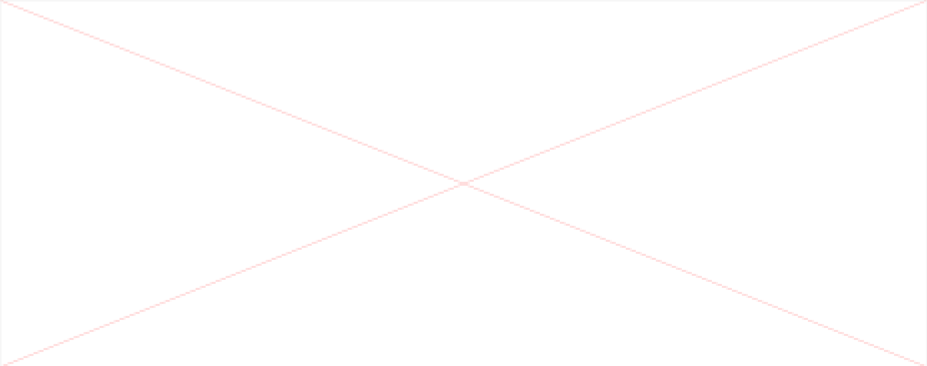


A **cumulative distribution function (CDF)** is a function $F_X : \mathbb{R} \to [0, 1]$ which specifies a probability measure as,

$$F_X(x) \triangleq P(X \leq x). \qquad (1)$$

Properties

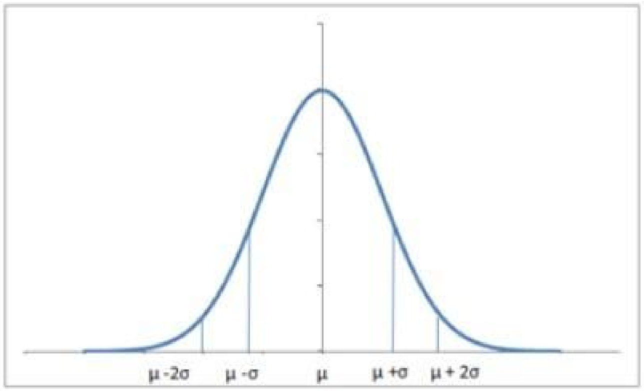




3.2 Probability density functions

3.3 Normal distribution

The normal distribution informally called as a bell curve looks like this:

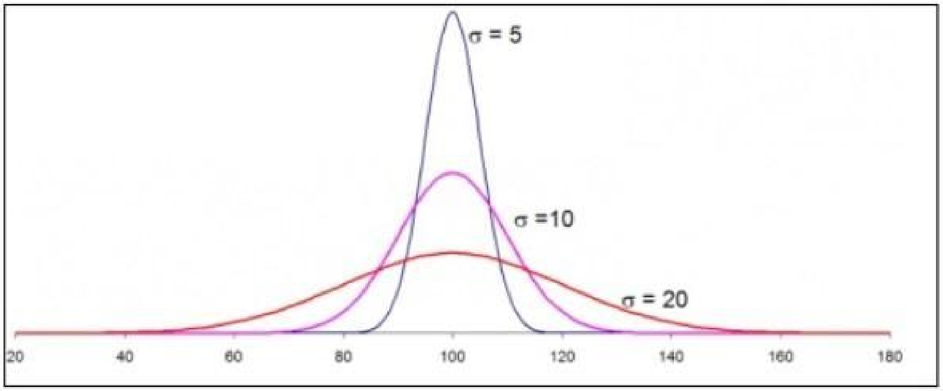


The equation of the normal distribution happens to be:

$$f(x|\mu, \sigma^2) = \frac{1}{\sqrt{(2\sigma^2\pi)}}\left(e^{-\frac{(x-\mu)^2}{2\sigma^2}}\right)$$

Here μ is the mean of the data while σ is the standard deviation of the data.

The normal distribution is perfectly symmetrical about the mean. The probabilities move similarly in both directions around the mean. The total area under the curve is 1, since summing up all the possible probabilities would give 1.
The distribution might vary a bit depending upon how spread the data is. If the data has a very high range and standard deviation, the normally distributed curve would be spread out and flatter, since a large number of values would be sufficiently away from the mean.

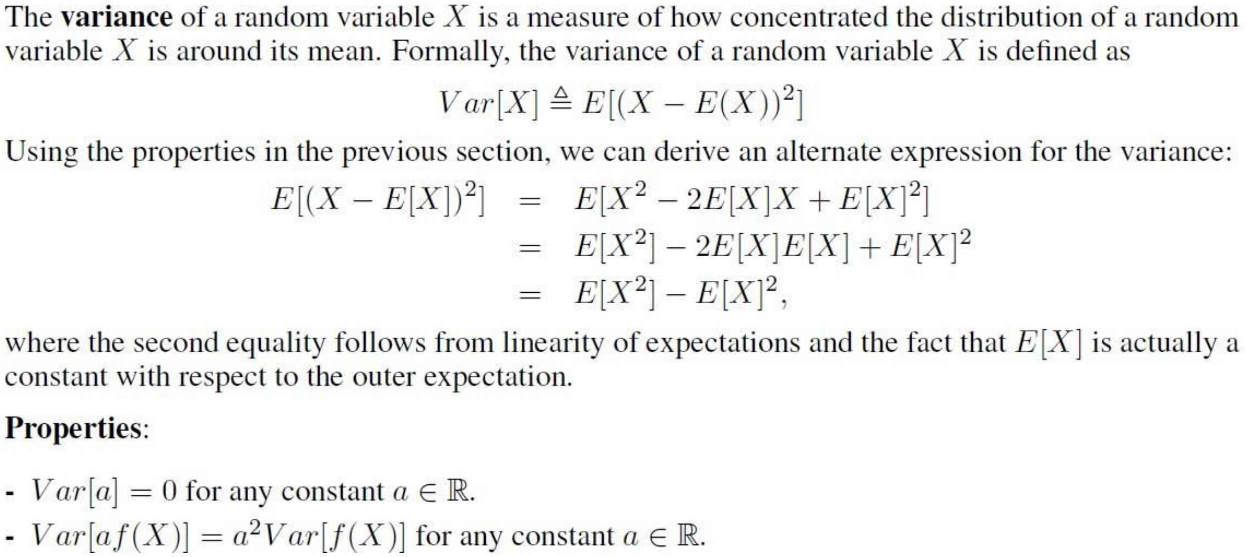


Also, if a lot of values are away from the mean, the probability for data being around the mean also drops. Similarly, if the standard deviation is low, which means most of the values are near around the mean, there is high probability of the sample mean being around the mean and the distribution is a lot skinnier. The higher the standard deviation, the thicker and flatter the curve.

- Area under a probability density function gives the probability for the random variable to be in that range.
- If I have a population data and I take random samples of equal size from the data, the sample means are approximately normally distributed

- There is large probability for the means to be around the actual mean of the data, than to be farther away
- Normal distributions for higher standard deviations are flatter as compared to those for lower standard deviations

3.4 Variance



The **variance** of a random variable X is a measure of how concentrated the distribution of a random variable X is around its mean. Formally, the variance of a random variable X is defined as

$$Var[X] \triangleq E[(X - E(X))^2]$$

Using the properties in the previous section, we can derive an alternate expression for the variance:

$$\begin{aligned} E[(X - E[X])^2] &= E[X^2 - 2E[X]X + E[X]^2] \\ &= E[X^2] - 2E[X]E[X] + E[X]^2 \\ &= E[X^2] - E[X]^2, \end{aligned}$$

where the second equality follows from linearity of expectations and the fact that $E[X]$ is actually a constant with respect to the outer expectation.

Properties:

- $Var[a] = 0$ for any constant $a \in \mathbb{R}$.
- $Var[af(X)] = a^2 Var[f(X)]$ for any constant $a \in \mathbb{R}$.

3.5 Some common random variables

Discrete random variables

- $X \sim Bernoulli(p)$ (where $0 \leq p \leq 1$): one if a coin with heads probability p comes up heads, zero otherwise.

$$p(x) = \begin{cases} p & \text{if } p = 1 \\ 1 - p & \text{if } p = 0 \end{cases}$$

- $X \sim Binomial(n, p)$ (where $0 \leq p \leq 1$): the number of heads in n independent flips of a coin with heads probability p.

$$p(x) = \binom{n}{x} p^x (1 - p)^{n-x}$$

- $X \sim Geometric(p)$ (where $p > 0$): the number of flips of a coin with heads probability p until the first heads.

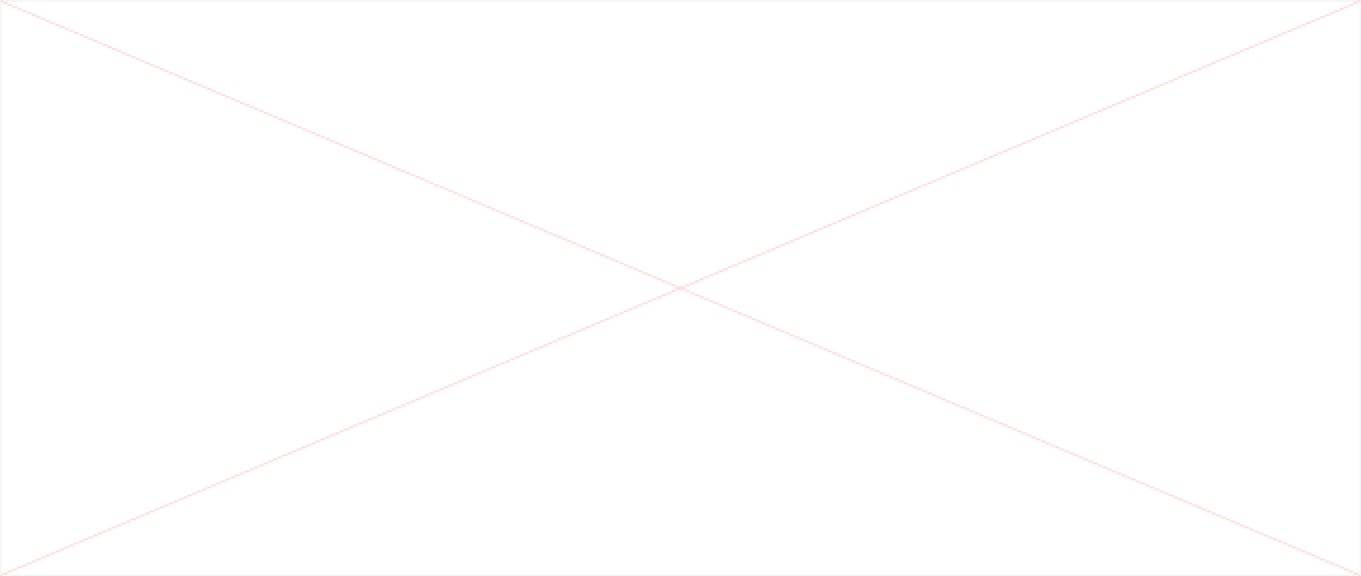
$$p(x) = p(1 - p)^{x-1}$$

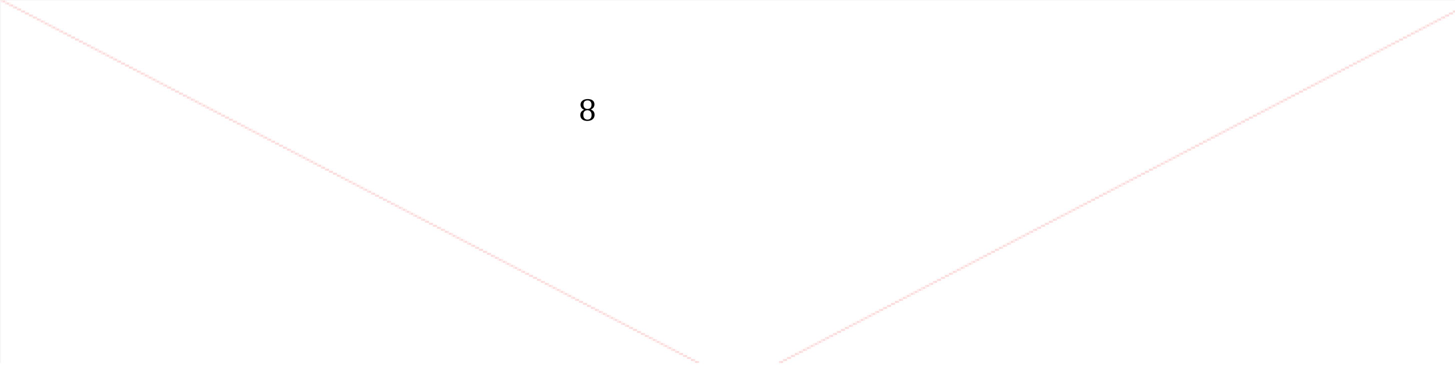
- $X \sim Poisson(\lambda)$ (where $\lambda > 0$): a probability distribution over the nonnegative integers used for modeling the frequency of rare events.

$$p(x) = e^{-\lambda} \frac{\lambda^x}{x!}$$

4 Conditional Probability and Bayes' theorem

Conditional Probability is the study of the probability of two things happening together. The way to do this is by applying Bayes' theorem which provides a simple way for calculating conditional probabilities.





Speaking mathematically, the probability of the model given the data is probability of the data given the model times the ratio of the independent probability of the model and the independent probability of the data.

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

In most cases, you can't just plug numbers into an equation; You have to figure out what your ”tests” and ”events” are first. For two events, A and B, Bayes' theorem allows you to figure out p(A|B) (the probability that event A happened, given that test B was positive) from p(B|A) (the probability that test B happened, given that event A happened). It can be a little tricky to wrap your head around as technically you're working backwards; you may have to switch your tests and events around, which can get confusing. An example should clarify what I mean by ”switch the tests and events around.”

Bayes' theorem is simple but has profound implications. The degree of belief in a machine learning model can also be thought of as probabilities and machine learning can be thought of as learning models of data. Thus, we can consider multiple models, find out the probabilities they have given the data and then consider the model which has the higher probability. In practice, this may not be that simple, but at least it will be easier to understand and not fish on routes with zero probabilities. **Example:**

You might be interested in finding out a patients probability of having liver disease if they are an alcoholic. ”Being an alcoholic” is the test (kind of like a litmus test) for liver disease.

A could mean the event ”Patient has liver disease.” Past data tells you that 10% ofpatients entering your clinic have liver disease. P(A) = 0.10.

B could mean the litmus test that ”Patient is an alcoholic.” Five percent of the clinicspatients are alcoholics. P(B) = 0.05.

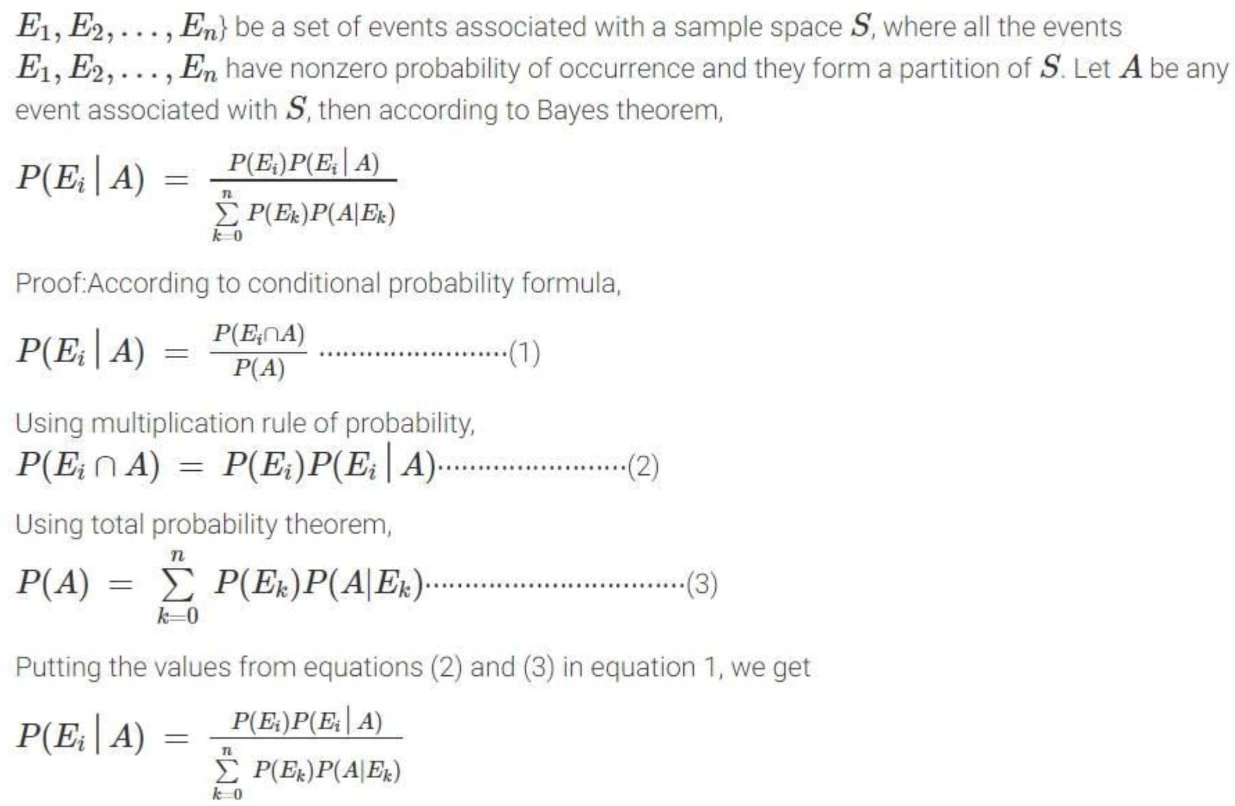
You might also know that among those patients diagnosed with liver disease, 7% are alcoholics. This is your B|A: the probability that a patient is alcoholic, given that they have liver disease, is 7%.

Bayes theorem tells you:

P(A|B) = (0.07 * 0.1)/0.05 = 0.14

In other words, if the patient is an alcoholic, their chances of having liver disease is 0.14 (14%). This is a large increase from the 10% suggested by past data. But its still unlikely that any particular patient has liver disease.

Derivation of Bayes Formula



$E_1, E_2, \ldots, E_n\}$ be a set of events associated with a sample space S, where all the events $E_1, E_2, \ldots, E_n$ have nonzero probability of occurrence and they form a partition of S. Let A be any event associated with S, then according to Bayes theorem,

$$P(E_i \mid A) = \frac{P(E_i)P(E_i \mid A)}{\sum_{k=0}^{n} P(E_k)P(A|E_k)}$$

Proof:According to conditional probability formula,

$$P(E_i \mid A) = \frac{P(E_i \cap A)}{P(A)} \cdots\cdots\cdots\cdots\cdots(1)$$

Using multiplication rule of probability,

$$P(E_i \cap A) = P(E_i)P(E_i \mid A) \cdots\cdots\cdots\cdots\cdots(2)$$

Using total probability theorem,

$$P(A) = \sum_{k=0}^{n} P(E_k)P(A|E_k) \cdots\cdots\cdots\cdots\cdots(3)$$

Putting the values from equations (2) and (3) in equation 1, we get

$$P(E_i \mid A) = \frac{P(E_i)P(E_i \mid A)}{\sum_{k=0}^{n} P(E_k)P(A|E_k)}$$