A BRIEF ANALYSIS ABOUT RANDOM VARIABLES AND PROBABILITY DISTRIBUTIONS

NAGENDRAPPA.G.
ASSISTANT PROFESSOR
DEPARTMENT OF MATHEMATICS
GOVERNMENT FIRST GRADE COLLEGE KORATAGERE TUMKUR DIST KARNATAKA

ABSTRACT

In machine learning, we often deal with uncertainty and stochastic quantities, due to one of the reasons being incomplete observability therefore, we most likely work with sampled data. Probability theory is a branch of mathematics concerned with the study of random phenomena and is often considered one of the fundamental pillars of machine learning. It is however a huge field to cover and very easy to get lost in, especially when being self-taught. Loosely speaking, the random variable is a variable whose value depends on the outcome of a random event. We can also describe it as a function that maps from the sample space to a measurable space (e.g. a real number). The probability mass function (PMF) describes the probability distribution over a discrete random variable. In other terms, it is a function that returns the probability of a random variable being exactly equal to a specific value.

KEYWORDS;- *RANDOM VARIABLES, PROBABILITY DISTRIBUTIONS, NON-NEGATIVITY, PROBABILITY MASS FUNCTION, BINOMIAL DISTRIBUTION..*

INTRODUCTION

Probability theory is a branch of mathematics concerned with the study of random phenomena and is often considered one of the fundamental pillars of machine learning. It is however a huge field to cover and very easy to get lost in, especially when being self-taught.

In the following sections, we are going to cover some fundamental aspects especially relevant to machine learning the random variable and the probability distribution.

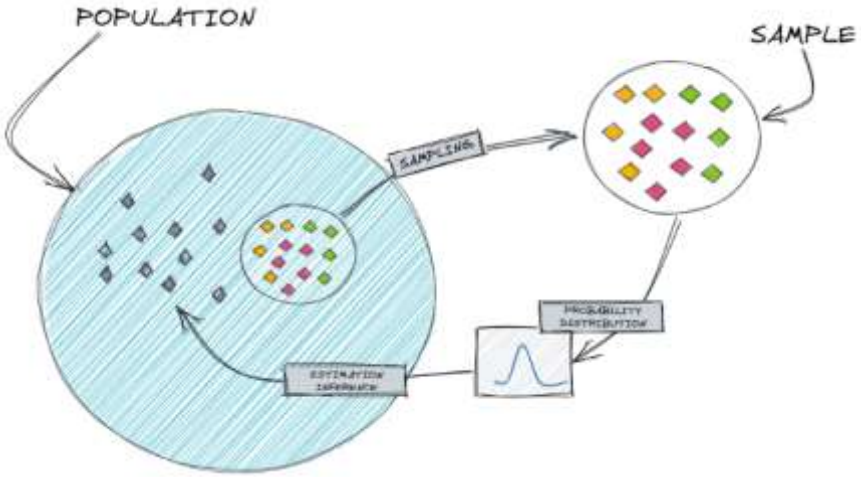
But before diving headfirst into the depth of probability theory, let's try to answer the question of why those concepts are important to understand and why we should even care in the first place.

Why Probability?

In machine learning, we often deal with uncertainty and stochastic quantities, due to one of the reasons being incomplete observability therefore, we most likely work with sampled data.

Now, suppose we want to draw reliable conclusions about the behavior of a random variable, despite the fact that we only have limited data and we simply do not know the entire population.

Hence, we need some kind of way to generalize from the sampled data to the population, or in other words — we need to estimate the true data-generating process.



Estimating the data-generating process [Image by Author]

Understanding the probability distribution, allows us to compute the probability of a certain outcome by also accounting for the variability in the results. Thus, it enables us to generalize from the sample to the population, estimate the data-generating function and predict the behavior of a random variable more accurately.

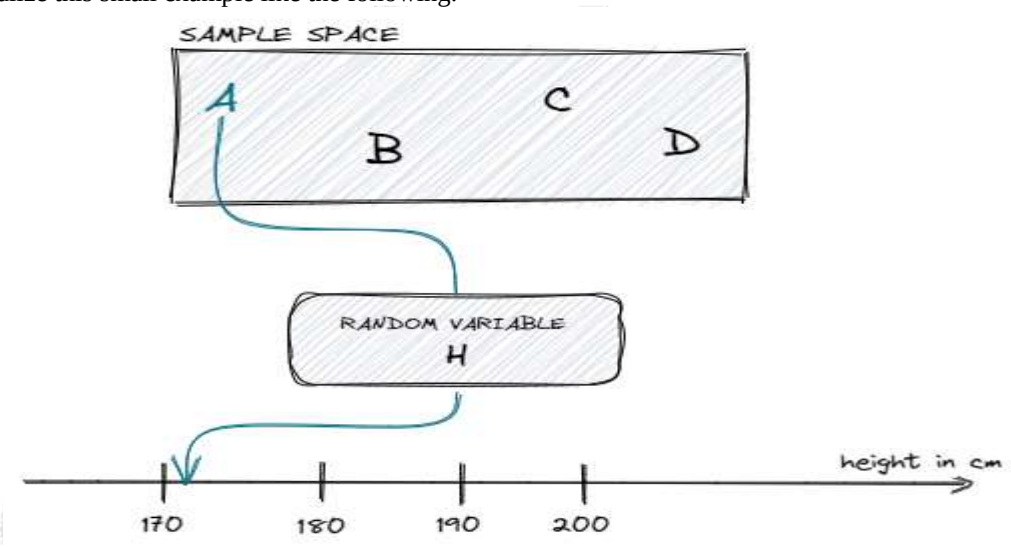
Introducing the Random Variable

Loosely speaking, the random variable is a variable whose value depends on the outcome of a random event. We can also describe it as a function that maps from the sample space to a measurable space (e.g. a real number).

Let's assume, we have a sample space containing 4 students {A, B, C, D}. If we now randomly pick student A and measure the height in centimeters, we can think of the random variable (H)as the function with the input of studentand the output of heightas a real number.

$$H(student) = height$$

We can visualize this small example like the following:



An Example of a random variable [Image by Author]

Depending on the outcome which student is randomly picked — our random variable (H) can take on different states or different values in terms of height in centimeters.

A random variable can be either discrete or continuous.

If our random variable can take only a finite or countably infinite number of distinct values, then it is discrete. Examples of a discrete random variable include the number of students in a class, test questions answered correctly, the number of children in a family, etc.

Our random variable, however, is continuous if between any two values of our variable are an infinite number of other valid values. We can think of quantities such as pressure, height, mass, and distance as examples of continuous random variables.

When we couple our random variable with a probability distribution we can answer the following question: How likely is it for our random variable to take a specific state? Which is basically the same as asking for the probability.

Now, we are left with one question that remains— what is a probability distribution?

Probability Distribution

The description of how likely a random variable takes one of its possible states can be given by a probability distribution. Thus, the probability distribution is a mathematical function that gives the probabilities of different outcomes for an experiment.

More generally it can be described as the function which maps an input space A related to the sample space to a real number, namely the probability.

$$P : A \to \mathbb{R}$$

For the above function to characterize a probability distribution, it must follow all of the Kolmogorov axioms:

1. **Non-negativity**
2. **No probability exceeds 1**
3. **Additivity of any countable disjoint (mutually exclusive) events**

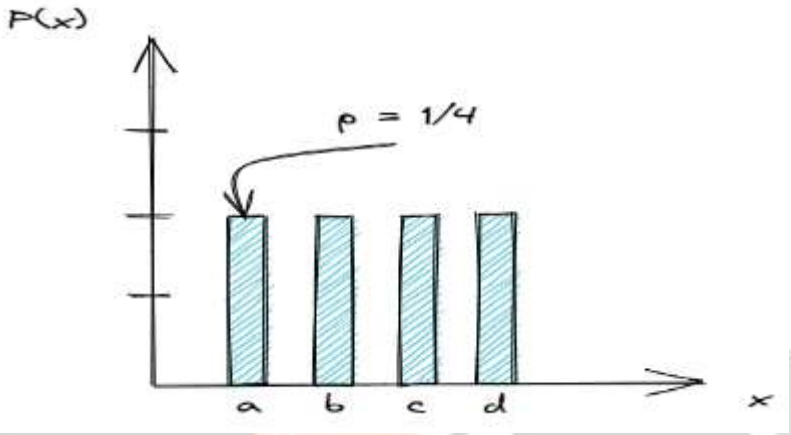
The way we describe a probability distribution depends on whether the random variable is discrete or continuous, which will result in a probability mass or density function respectively.

Probability Mass Function

The probability mass function (PMF) describes the probability distribution over a discrete random variable. In other terms, it is a function that returns the probability of a random variable being exactly equal to a specific value.

The returned probability lies in the range [0, 1] and the sum of all probabilities for every state equals one.

Let's imagine a plot where the x-axis describes the states and the y-axis shows the probability of a certain state. Thinking this way allows us to envision the probability or the PMF as a barplot sitting on top of a state.



An example (uniform) PMF [Image by Author]

In the following, we will learn about three common discrete probability distributions: The Bernoulli, binomial and geometric distribution.

Bernoulli distribution

Named after the Swiss mathematician Jacob Bernoulli, the Bernoulli distribution is a discrete probability distribution of a single binary random variable, which either takes the value 1 or 0.

Loosely speaking, we can think of the Bernoulli distribution as a model giving the set of possible outcomes for a single experiment, that can be answered with a simple yes-no question.

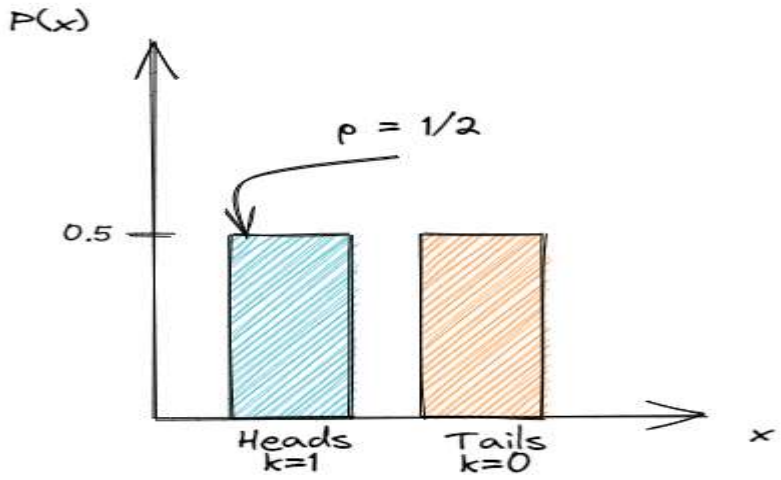
More formally the function can be stated as the following equation

$$f(k;p) = \begin{cases} q = 1 - p & \text{if k=0} \\ p & \text{if k=1} \end{cases}$$

$$f(k;p) = p^k(1-p)^{1-k} \quad \text{for k} \in \{0,1\}$$

Which basically evaluates to p if k=1 or to (1-p) if k=0. Thus, the Bernoulli distribution is parametrized by just a single parameter p.

Suppose, we toss a fair coin once. The probability of obtaining heads is P(Heads) = 0.5. Visualizing the PMF we get the following plot:



An example of a Bernoulli Trial [Image by Author]

Note: The Bernoulli distribution takes either the value 1 or 0, which makes it particularly useful as an indicator or dummy variable.

Since the Bernoulli Distribution models only a single trial, it can also be viewed as a special case of the binomial distribution.

Binomial Distribution

The binomial distribution describes the discrete probability distribution of the number of successes in a sequence of n independent trials, each with a binary outcome. The success or failure is given by the probability p or (1-p) respectively.

Thus, the binomial distribution is parametrized by the parameters

$$n \in \mathbb{N}, \quad p \in [0,1]$$

More formally the binomial distribution can be expressed with the following equation:

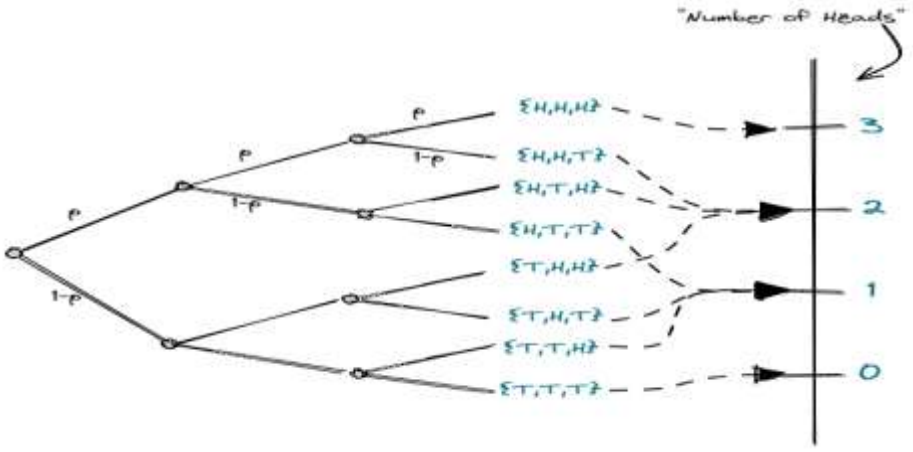
$$f(k;n,p) = \binom{n}{k} p^k (1-p)^{n-k}$$

The success of k is given by the probability p to the power of k, whereas the probability of failure is defined by (1-p) to the power of n minus k, which is basically the number of trials minus the one trial where we get k.

Since the event of success k can occur anywhere in n trials, we have **"n choose k"** ways to distribute the success.

Let's pick up our coin-tossing example from before and build on it.

Now, we are going to flip the fair coin three times, while being interested in the random variable describing the number of heads obtained.



Number of heads in three coin flips [Image by Author]

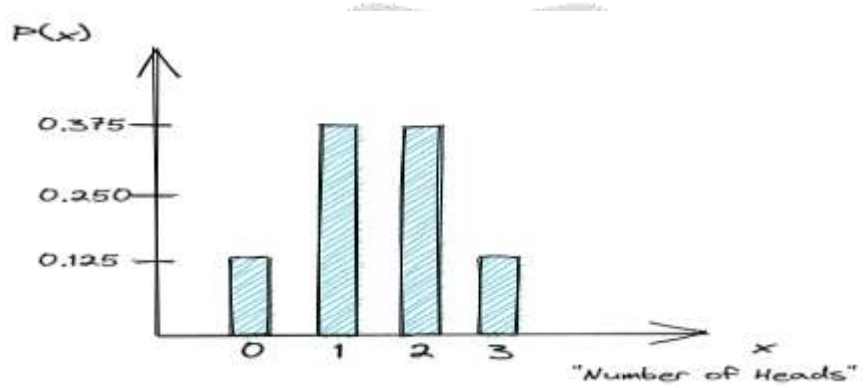
If we want to compute the probability of the coin coming up as heads two times, we can simply use the equation from before and pluck in the values

$$P(2) = \binom{3}{2} p^2 (1-p)^{3-2}$$

$$P(2) = 3(0.5)^2 (0.5)^1$$

$$P(2) = 0.375$$

Which results in a probability P(2) = 0.375. If we proceed in the same way for the remaining probabilities, we get the following distribution:



The binomial distribution of three coin flips [Image by Author]

Geometric Distribution

Suppose, we are interested in the number of times we have to flip a coin until it comes up heads for the first time.

The geometric distribution gives the probability of the first success occurrence, requiring n independent trials, with a success probability of p.

More formally it can be stated as

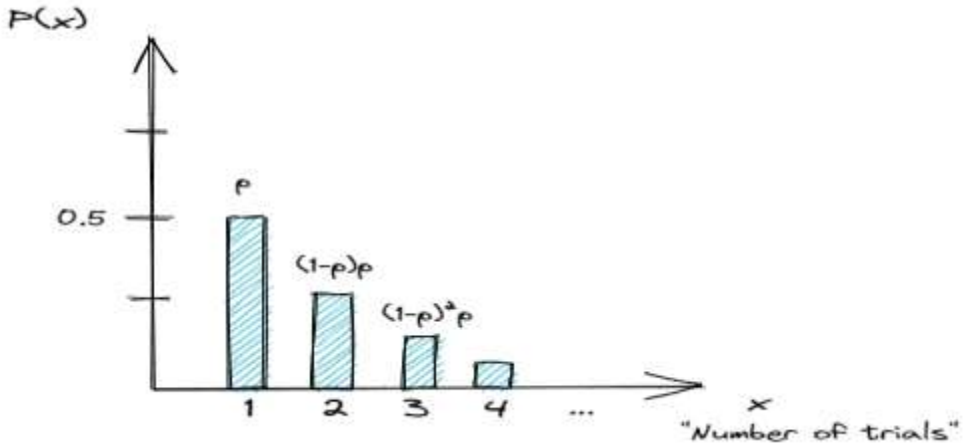
$$P(n) = (1-p)^{n-1} p$$

Which computes the probability of the number of trials needed up to and including the success event.

The following assumptions need to be true, in order to calculate the geometric distribution:

1. **Independence**
2. **For each trial, there are only two possible outcomes**
3. **The probability of success is the same for every trial**

Let's visualize the geometric distribution by answering the question for the probability of the number of trials needed for the coin to come up heads for the first time.



The geometric distribution until first heads [Image by Author]

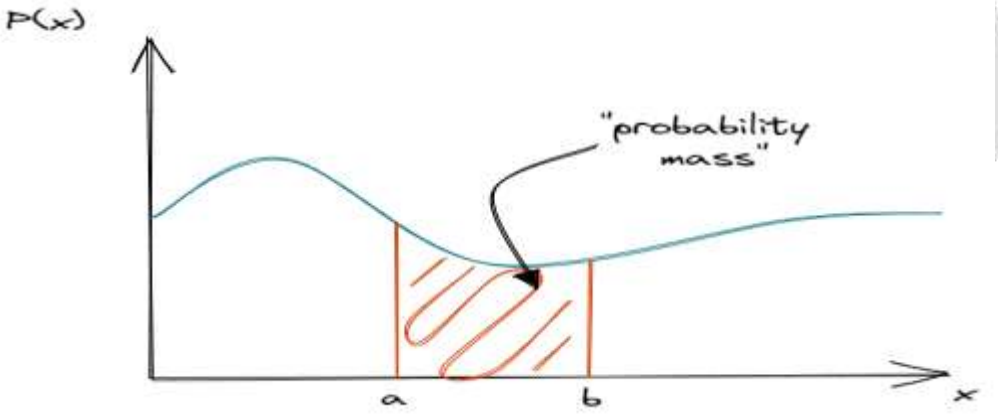
Probability Density Function

In the earlier sections, we learned that a random variable can either be discrete or continuous. If it is discrete, we can describe the probability distribution with a probability mass function.

Now, we are dealing with continuous variables — hence, we need to describe the probability distribution with a probability density function (PDF).

The PDF, contrary to the PMF, does not give the probability of a random variable taking a specific state directly. Instead, it describes the probability of landing inside an infinitesimal region. In other terms, the PDF describes the probability of a random variable lying between a particular range of values.

In order to find the actual probability mass, we need to integrate, which yields the area under the density function but above the x-axis.



An example probability density function [Image by Author]

The probability density function must be non-negative and its integral needs to be 1.

$$(1) \qquad p(x) \geq 0$$

$$(2) \qquad \int p(x)\delta x = 1$$

One of the most common continuous probability distributions is the gaussian or normal distribution.

Gaussian Distribution

The Gaussian distribution is often considered a sensible choice to represent a real-valued random variable, whose distribution is unknown.

This is mainly due to the central limit theorem, which, loosely speaking, states that the average of many independent random variables with finite mean and variance is itself a random variable which is normally distributed as the number of observations increases.

This is especially useful since it allows us to model complicated systems as Gaussian distributed, even if the individual parts follow a more complicated structure or distribution.

Another reason it is a common choice for modeling a distribution over a continuous variable is the fact that it inserts the least amount of prior knowledge.

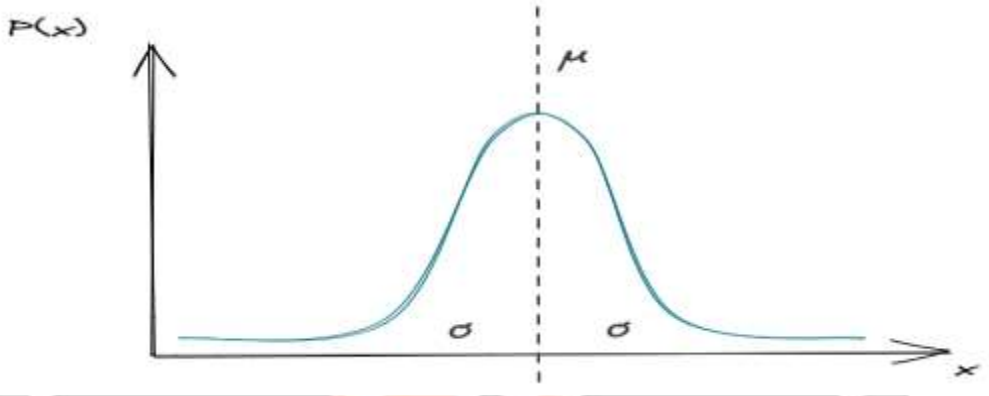
More formally, the Gaussian distribution can be stated as

$$\mathcal{N}(x;\mu,\sigma^2) = \sqrt{\frac{1}{2\pi\sigma^2}}\exp(-\frac{1}{2\sigma^2}(x-\mu)^2)$$

where the parameter μ is the mean and σ² describes the variance.

In simple terms, the mean will be responsible for defining the central peak of the bell-shaped distribution, whereas the variance or the standard deviation defines its width.

We can visualize the normal distribution as the following:



An example of a normal distribution [Image by Author]

Conclusion

In this article, we talked about random variables, probability distributions, how they are related, and how we can interpret them. We also distinguished discrete and continuous random variables by introducing some of the most common probability mass and density functions.

Although it is possible to apply learning algorithms without knowing the basics of probability distributions and still get decent results a deeper understanding of the subject will enable us to make better choices, assumptions, and predictions about the true behavior of a random variable.

References

- Deep Learning (Ian J. Goodfellow, Yoshua Bengio and Aaron Courville), Chapter 3, MIT Press, 2016.
- Castañeda; V. Arunachalam & S. Dharmaraja (2012). Introduction to Probability and Stochastic Processes with Applications. Wiley. p. 67. ISBN 9781118344941.
- Billingsley, Patrick (1995). Probability and Measure (3rd ed.). Wiley. p. 187. ISBN 9781466575592.
- Bertsekas, Dimitri P. (2002). Introduction to Probability. Tsitsiklis, John N., Τσιτσικλής, Γιάννης Ν. Belmont, Mass.: Athena Scientific. ISBN 188652940X. OCLC 51441829.
- Steigerwald, Douglas G. "Economics 245A – Introduction to Measure Theory" (PDF). University of California, Santa Barbara. Retrieved April 26, 2013.