A Brief in probability theory

Probability theory is incorporated into machine learning, particularly the subset of artificial intelligence concerned with **predicting outcomes and making decisions**.

In computer science, functions are used to limit the functions outcome to a value between 0 and 1.

It is the core concept as well as a primary prerequisite to understanding the ML models and their applications.

Probability theory is a mathematical framework for quantifying our uncertainty about the world. It allows us (and our software) to reason effectively in situations where being certain is impossible.

Probability theory is at the foundation of many machine learning algorithms. The goal of this post is to cover the vocabulary and mathematics needed before applying probability theory to machine learning applications.

Mathematics of Probability

Probability is all about the possibility of various outcomes. The set of all possible outcomes is called the **sample space**.

The sample space for a coin flip is {heads, tails}. The sample space for the temperature of water is all values between the freezing and boiling point. Only one outcome in the sample space is possible at a time, and the sample space must contain all possible values.

The sample space is often depicted as Ω (capital omega) and a specific outcome as ω (lowercase omega). We represent the probability of an event ω as $P(\omega)$.

The two basic axioms of probability are:

- $0 \leq P(\omega) \leq 1$
- $\sum_{\omega} P(\omega) = 1$

The probability of any event has to be between 0 (impossible) and 1 (certain), and the sum of the probabilities of all events should be 1.

This follows from the fact that the sample space must contain all possible outcomes.

Therefore, we are certain (probability 1) that one of the possible outcomes will occur.

Conditional Probability Formula

Probability of
A and *B*

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

Probability of
A given *B*

Probability of *B*

Bayes' Rule:

The chain rule for two variables in two equivalent ways:

- $P(x, y) = P(x|y) \cdot P(y)$
- $P(x, y) = P(y|x) \cdot P(x)$

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

- If we set both right sides equal to each other and divide by P(*y*), we get Bayes' rule:

$$P(x|y) = \frac{P(y|x) \cdot P(x)}{P(y)}$$

Expectation:

The expected value, or **expectation,** of a function *h*(x) on a random variable x ~ P(x) is the average value of *h*(*x*) weighted by P(x). For a discrete x, we write this as:

$$\mathbb{E}[h(x)] = \sum_x P(x) \cdot h(x)$$

What's the expected value of playing the guessing game at the casino if we assume we have a 1/10 chance of guessing the correct number?

$\mathbb{E}[h(x)]$ = P(winning) · *h*(winning) + P(loosing) · *h*(loosing)= (1/10) · \$8 + (9/10) · (-\$2) = \$0.80 + (-\$1.80) = -\$1. So on average, we'll loose \$1 every time we play!

Variance and Covariance:

We saw variance with respect to a Gaussian distribution when we were talking about continuous random variables. In general, **variance** is a measure of how much random values vary from their mean. Similarly, for functions of random variables, the variance is a measure of the variability of the function's output from its expected value.

$$\mathrm{Var}(h(x)) = \mathbb{E}[(h(x) - \mathbb{E}[h(x)])^2]$$