Understanding the Fundamentals of Probability and Random Variables in Machine Learning

Rishabh Iyer

March 19, 2024

1 Introduction

Probability theory is a cornerstone of machine learning, providing the framework for understanding uncertainty, making predictions, and modeling complex systems. This article introduces the fundamental concepts of probability, explores the definition and properties of random variables, and discusses the concept of entropy, setting the stage for their application in decision trees and probabilistic models.

2 Basics of Probability

Probability offers a way to quantify uncertainty. It involves the study of events, the outcomes of experiments, and the likelihood of these outcomes.

2.1 Probability Axioms

The axioms of probability form the foundation of probability theory, ensuring that probabilities are consistently assigned to events. These axioms are:

1. The probability of any event A, denoted $P(A)$, is a non-negative real number.
2. The probability of the sample space S, representing all possible outcomes, is 1: $P(S) = 1$.
3. For any sequence of mutually exclusive events $A_1, A_2, \ldots$, the probability of their union is the sum of their probabilities: $P(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$.

Below, we consider three examples of probabilistic processes:

Dice Roll: Consider a fair six-sided dice with the sample space $S = \{1, 2, 3, 4, 5, 6\}$. The probability of rolling a 3, $P(3)$, is $\frac{1}{6}$, satisfying the non-negativity axiom. The probability of the sample space, $P(S)$, is 1, and for mutually exclusive events, like rolling a 2 or 3, $P(2 \text{ or } 3) = P(2) + P(3) = \frac{1}{6} + \frac{1}{6} = \frac{1}{3}$.

Coin Flip: With a fair coin, the sample space is $S = \{H, T\}$. The probability of getting heads, $P(H)$, is $\frac{1}{2}$, illustrating the non-negativity axiom. The sum of probabilities of all outcomes, $P(H) + P(T)$, equals 1, fulfilling the probability of the sample space axiom.

Pair of Dice: Rolling a pair of fair six-sided dice increases the sample space to $S = \{(i, j) | i, j \in \{1, 2, 3, 4, 5, 6\}\}$, where i and j represent the outcomes of the first and second dice, respectively. The probability of rolling a sum of 7, for example, is $P(\text{Sum} = 7) = \frac{6}{36} = \frac{1}{6}$, as there are 6 outcomes that produce a sum of 7: $((1, 6), (2, 5), (3, 4), (4, 3), (5, 2), (6, 1))$.

2.2 Conditional Probability, Bayes' Rule, and Independence

Conditional probability, $P(A|B)$, represents the probability of event A given that event B has occurred. Bayes' Rule provides a way to update our beliefs based on new evidence:

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

Two events A and B are independent if the occurrence of one does not affect the probability of the occurrence of the other, i.e., $P(A|B) = P(A)$. The joint probability distribution of two random variables describes the probability of simultaneous occurrences of their outcomes.
Below, we consider three examples of conditional probabilities, Bayes rule, and Independence:

Dice Roll: The probability of rolling an even number given that the roll is greater than 4 is $P(\text{Even}| > 4) = P(6| > 4) = \frac{1}{2}$, since only 5 and 6 are greater than 4, and only 6 is even.

Coin Flip: If we flip two coins, the probability of the second coin being heads given that the first coin is heads is $P(H_2|H_1) = P(H_2) = \frac{1}{2}$, demonstrating independence (see the next subsection) since the outcome of the first flip does not affect the second.

Pair of Dice: The probability of the sum being 7 given that the first dice shows a 4 is $P(\text{Sum} = 7|\text{First dice} = 4) = P(3) = \frac{1}{6}$, since only one outcome $(4, 3)$ leads to a sum of 7 when the first dice is 4. The outcome of the first dice is independent of the outcome of the second dice. For example, $P(\text{First dice} = 1 \text{ and Second dice} = 2) = P(\text{First dice} = 1) \times P(\text{Second dice} = 2) = \frac{1}{6} \times \frac{1}{6} = \frac{1}{36}$.

3 Random Variables

A random variable is a function that assigns a real number to each outcome in the sample space of a random experiment. Random variables can be discrete or continuous, depending on their value set.

Expectation and Variance: The expectation (or expected value) of a random variable provides a measure of its central tendency, while the variance measures the spread of its values.

Below, are examples of Random variables from the dice roll and coin flip.

Dice Roll: Define a random variable X representing the outcome of a dice roll. So $X = 1$ corresponds to the dice roll of 1 and so on. The expected value $E[X]$ is 3.5, calculated as the average of all possible outcomes.

Coin Flip: Let Y be a random variable representing the number of heads in a single flip, with $Y = 1$ for heads and $Y = 0$ for tails. The expected value $E[Y]$ is $\frac{1}{2}$, reflecting the fair chance of heads or tails.

Pair of Dice: Define a random variable Z as the sum of the outcomes of rolling a pair of dice. The expected value $E[Z]$ is calculated by considering all possible sums and their probabilities. For example, the probability of $Z = 7$ is $\frac{1}{6}$, and so on for other sums. The expected value is:

$$E[Z] = \sum_{z=2}^{12} z \cdot P(Z = z) = \frac{1}{36}(2 \cdot 1 + 3 \cdot 2 + \ldots + 12 \cdot 1) = 7$$

4 Entropy

Entropy measures the uncertainty or randomness of a random variable. For a discrete random variable X with possible values $\{x_1, \ldots, x_n\}$ and probability mass function $P(X)$, the entropy $H(X)$ is defined as:

$$H(X) = -\sum_{i=1}^{n} P(x_i) \log P(x_i)$$

Entropy is a foundational concept in information theory and plays a crucial role in decision trees and other machine learning models, where it can be used to quantify the information gain associated with a particular split.

Dice Roll: The entropy of the dice roll, with each outcome equally likely, is calculated as $H(X) = -\sum_{i=1}^{6} \frac{1}{6} \log_2 \frac{1}{6} \approx 2.585$ bits, indicating the uncertainty of the dice roll outcome. This is the highest entropy. The lowest entropy is when one of the outcomes (say 1) has probability 1 and everything else is probability 0. In that case, the entropy is 0.

Coin Flip: The entropy of a fair coin flip is $H(Y) = -\left(\frac{1}{2} \log_2 \frac{1}{2} + \frac{1}{2} \log_2 \frac{1}{2}\right) = 1$ bit, representing the maximum uncertainty in this binary event. Again, this is the highest entropy. The lowest entropy is when one of the events has probability 1 (say heads) and other is 0 (tails), in which case, the entropy is 0.

Pair of Dice: The entropy of the sum Z of a pair of dice, considering the varying probabilities of each sum (e.g., $P(Z = 7) = \frac{1}{6}$, while $P(Z = 2) = P(Z = 12) = \frac{1}{36}$), is calculated as:

$$H(Z) = -\sum_{z=2}^{12} P(Z = z) \log_2 P(Z = z)$$

This entropy value quantifies the uncertainty in the sum of the outcomes of rolling a pair of dice, reflecting the distribution of sums rather than individual outcomes.

5 Conclusion

The principles of probability, the properties of random variables, and the concept of entropy are fundamental to many aspects of machine learning. Understanding these concepts is essential for developing and analyzing models that can learn from data and make predictions about the world.