02 Probability Theory

May 23, 2021

Information: A brief review of basic concepts of Probability related theory used in Machine Learning

Written by: Zihao Xu

Last update date: 05.23.2021

1 Basic concepts

1.1 Introduction

1.1.1 Definition:

- Probability Theory is a mathematical frame work for representing uncertain statements
  - Provides a means of quantifying uncertainty as well as axioms for deriving new uncertainty statements
  - Allows making uncertain statements and reasoning in the presence of uncertainty

1.1.2 Motivation:

- Machine learning must always deal with uncertain quantities and sometimes stochastic (non-deterministic) quantities resulting from:
  - Inherent stochasticity in the system being modeled
  - Incomplete observability
  - Incomplete modeling
- It's usually more practical to use a *simple but uncertain rule* rather than a complex but certain one
  - "Most birds fly" versus "Birds fly, except for ..."
- Uncertainty can be modeled by probability

1.1.3 Interpretation of probability

- **Frequentist probability**
  - Probability theory was originally developed to analyze the *frequencies of events*, which are often *repeatable*
  - When we say that an outcome has a probability p of occurring, it means that if we repeated the experiment infinitely times, then a proportion p of the repetitions would result in that outcome
  - Not seemingly applicable to propositions that are not repeatable
- **Bayesian probability**

  - Use probability to represent a *degree of belief* for an outcome to occur, with 1 indicating absolute certainty that the outcome occurs and 0 indicating absolute certainty that the outcome doesn't occur
  - For properties that we expect common sense reasoning about uncertainty to have, we need to treat Bayesian probabilities as behaving exactly the same as frequentist probabilities.
- **The extension logic to deal with uncertainty**
  - Probability can be seen as the extension of logic to deal with uncertainty. Probability theory provides a set of formal rules for determining the likelihood of a proposition being true given the likelihood of other propositions

1.2 Probability Space

- A *probability space* is defined by the triple $(\Omega, \mathcal{F}, P)$, where:
  - Ω is the *space of possible outcomes* (or outcome space)
  - $\mathcal{F} \subseteq 2^{\Omega}$ is the *space of (measurable) events* (or event space)
  - P is the *probability measure* (or *probability distribution*) that maps an event $E \in \mathcal{F}$ to a real value between 0 and 1 (think of P as a function)
- Given an outcome space Ω, there is some restrictions as to what subset of 2^{Ω} can be considered an event space $\mathcal{F}$:
  - The trivial event Ω and the empty event $\nvdash$ is in $\mathcal{F}$
  - The event space $\mathcal{F}$ is closed under (countable) union
    - if $\alpha, \beta \in \mathcal{F}$, then $\alpha \cup \beta \in \mathcal{F}$
  - The event space $\mathcal{F}$ is closed under complement
    - if $\alpha \in \mathcal{F}$, then $(\Omega \setminus \alpha) \in \mathcal{F}$
- Notice: Event space is not always simply the power set of the outcome space

1.3 Random Variables

1.3.1 Definition

- A *random variable* is a function maps outcomes in the outcome space Ω to real values
  - **Not random nor a variable**
- May be discrete or continuous
  - Discrete: When the image of a random variable is countable, the random variable is called a *discrete random variable*
  - Continuous: When the image of random variable is uncountably infinite (usually an interval), then the random variable is called a *continuous random variable*
  - Image:In mathematics, the image of a function is the set of all output values it may produce.

1.3.2 Notation

- Upper case letters to represent random variables
  - $X, Y, Z, ...$
- Lower case letters to represent the values of random variables
  - $x, y, z, ...$

1.4 Probability Distributions

- A **probability distribution** is a description of how likely a random variable or set of random variables is to take on each of its possible states. The way to describe probability distribution depends on whether the variables are discrete or continuous.

1.4.1 Discrete Variables and Probability Mass Functions

- **Probability Mass Function (PMF)**: A function that gives the probability a discrete random variable is exactly equal to some value
  - Called probability mass function as it divides up a unit mass (the total probability) and places them on different values a random variable can take
  - Sometimes called *discrete density function*
  - The primary means of defining a discrete probability distribution
- **Notations for PMF**:
  - Let X be a discrete random variable, the probability mass function of X is:

$$P(X = x) = \text{Probability that the random variable } X \text{ takes the value } x$$

    - $*$ Can be written in $p_X(x)$ or even directly $p(x)$ when there is no ambiguity
  - Sometimes we define a variable first, then use $\sim$ notation to specify which distribution it follows later: $X \sim P(X)$
- **Properties of PMF**: Given that X is a *discrete random variable*, then:
  - The domain of p must be the set of all possible states of X
  - $\forall x \in X, 0 \leq p(x) \leq 1$. An impossible event has probability 0, and no state can be less probable than that. Likewise, an event that is guaranteed to happen has probability 1, and no state can have a greater chance of occurring
  - $\sum\limits_{x} p(x) = 1$.Usually refer to this property as being *normalized.*
  - Probability of X taking either the value x_1 or the value x_2 (assuming $x_1 \neq x_2$) is:

$$P(X = x_1 \text{ or } X = x_2) \equiv P(X \in \{x_1, x_2\}) = P(X = x_1) + P(X = x_2) = p(x_1) + p(x_2)$$

    - $*$ More generally, the probability that the random variable X takes any value in a set A is given by $P(X \in A) = \sum\limits_{x \in A} p(x)$
  - Let $g(x)$ be a function and $Y = g(X)$ be a new random variable, then the PMF would be:

$$P(y) = P(Y = y) = \sum_{x \in g^{-1}(y)} p(x)$$

1.4.2 Continuous Variables and Probability Density Functions

- **Cumulative Distribution Function(CDF)**:A function that gives the probability that a random variable will take a value less than some value
  - The random variable can be either discrete or continuous
  - Sometimes called *distribution function*
- **Notations for CDF**
  - Let X be a random variable, then its cumulative distribution function is denoted by:

$$F(x) = P(X \leq x)$$

- **Properties of CDF**: Given that X is a random variable, then:

- The domain of F must be the set of all possible states of X
  - $F(x)$ is an increasing function for x
  - $F(-\infty) = \lim_{x \to -\infty} F(x) = 0$
  - $F(+\infty) = \lim_{x \to -\infty} F(x) = 1$
  - $P(a \leq X \leq b) = F(b) - F(a)$
- **Probability Density Function (PDF)**: A function whose value at any given sample in the sample space (the set of possible values taken by the random variable) can be interpreted as providing a *relative likelihood* that the value of the random variable would equal that sample
  - Definition only for continuous random variables
  - While the *absolute likelihood* for a continuous random variable to take on any particular value is 0 (since there is an infinite set of possible values to begin with), the value of the PDF at two different samples can be used to infer how much more likely it is that the random variable would equal one sample compared to the other sample.
- **Notations for PDF**
  - Let X be a continuous random variable, then its probability density function is denoted by:
    $$p(x) \simeq \frac{P(x \leq X \leq x + \Delta x)}{\Delta x}$$
    for some small Δx according to definition
  - Observation:
    $$\begin{aligned} p(x) &= \lim_{\Delta x \to 0} \frac{P(x \leq X \leq x + \Delta x)}{\Delta x} \\ &= \lim_{\Delta x \to 0} \frac{F(x + \Delta x) - F(x)}{\Delta x} \\ &= F'(x) \\ &= \frac{\mathrm{d}F(x)}{\mathrm{d}x} \end{aligned}$$
- **Properties of PDF**: Given that X is a *continuous random variable*, then:
  - The domain of p must be the set of all possible states of X
  - $\forall x \in X, p(x) \geq 0$. Note that $p(x) \leq 1$ is not required
  - $\int p(x)dx = 1$
  - $\int_a^b p(x)dx = F(b) - F(a) = P(a \leq X \leq b)$
  - For any borel subset A of the real numbers:
    $$P(X \in A) = \int_A p(x)dx$$
    . This property holds even for random vectors
  - Let $g(x)$ be a function and $Y = g(X)$ be a new continuous random variable, then the PDF of Y would be:
    $$p(y) = p(x = g^{-1}(y))|\frac{d}{dy}(g^{-1}(y))|$$

1.5 Expectations and Variance

1.5.1 Expectations

- **Definition**: The **expectation**, or **expected value**, of a random variable is denoted by $\mathrm{E}(X)$

- For a discrete random variable X:

$$\mathrm{E}[X] = \underset{x}{\Sigma} x p(x)$$

  where $P(x)$ is the probability mass function and the sum is over all possible values
- For a continuous random variable X:

$$\mathrm{E}[X] = \int_x x p(x) dx$$

  where $p(x)$ is the probability density function and the integral is over all possible values
- Can think of the expectation as the value of the random variable that one should "expect" to get, but notice that the expectation of a random variable is usually not in the possible values of it.

- **Properties**: Let X be a random variable, then:
  - Let $g(x)$ be a function:
    - For a discrete random variable X:

$$\mathrm{E}[g(X)] = \underset{x}{\Sigma} g(x)p(x)$$

    - For a continuous random variable X:

$$\mathrm{E}[g(X)] = \int_x g(x)p(x)dx$$

  - Take any constant c:

$$\mathrm{E}[X + c] = \mathrm{E}[X] + c$$

  - Take any constant λ:

$$\mathrm{E}[\lambda X] = \lambda \mathrm{E}[X]$$

1.5.2 Variance

- **Definition**: The **variance** measures how much the value of a function of a random variable vary as sampling different values from its probability distribution
  - For a discrete/continuous random variable X:

$$\mathrm{V}[X] = \mathrm{E}\left[(X - \mathrm{E}[X])^2\right]$$

  - Can think of the variance as the spread of the random variable around its expectation
  - Square root of the variance is known as the **standard deviation**
- **Properties**:
  - $\mathrm{V}[X] = \mathrm{E}[X^2] - (\mathrm{E}[\mathrm{X}])^2$
  - Take any constant c:

$$\mathrm{V}[X + c] = \mathrm{V}[X]$$

  - Take any constant λ:

$$\mathrm{V}[\lambda X] = \lambda^2 \mathrm{V}[X]$$

1.6 Collections of Random Variables

1.6.1 Joint Distributions, Marginal Distributions and Conditional Distributions

- **Joint Distributions**: Distributions over multiple random variables
  - For discrete random variables X and Y, the **joint probability mass function** of the pair (X, Y) is the function $p(x, y)$ giving the probability that $X = x$ and $Y = y$ is denoted as:
$$p(x, y) = P(X = x, Y = y)$$
    - It's non-negative: $p(x, y) \geq 0$
    - Sum over all the possible values of all random variables is 1: $\underset{x}{\Sigma}\underset{y}{\Sigma} p(x, y) = 1$
    - *Sum Rule*: Marginalizing over the values of one of the random variables gets the PMF of the other: $\underset{y}{\Sigma} p(x, y) = p(x), \underset{x}{\Sigma} p(x, y) = p(y)$. This result is called the **Marginal Distribution**.
  - For continuous random variables X and Y, the **joint probability density function** of the pair (X, Y) is the function $p(x, y)$ giving the probability that $X = x$ and $Y = y$:
$$p(x, y) = \lim_{\Delta x, \Delta y \to 0} \frac{P(x \leq X \leq x + \Delta x, y \leq Y \leq y + \Delta y)}{\Delta x \Delta y}$$
    - It's non-negative: $p(x, y) \geq 0$
    - Sum over all the possible values of all random variables is 1: $\int \int p(x, y) dx dy = 1$
    - *Sum Rule*: Marginalizing over the values of one of the random variables gets the PDF of the other:$\int p(x, y) dy = p(x)$ and $\int p(x, y) dx = p(y)$. The result is called the **Marginal Distribution**
- **Conditional Distributions**: In many cases, the probability of some event given that some other event has happened is wanted. This called a **conditional probability**
  - For random variables X and Y, if $X = x$ is observed and the state of knowledge about Y is need to be updated, then the conditional PDF is:
$$P(Y = y | X = x) = p(y|x) = \frac{P(X = x, Y = y)}{P(X = x)} = \frac{p(x, y)}{p(x)}$$
    - The conditional probability is only defined when $P(X = x) > 0$. The conditional probability conditioned on an event that never happens is meaningless.
  - Chain Rule / Product Rule of Conditional Probabilities
$$P(x^{(1)}, \cdots, x^{(n)}) = P(x^{(1)}) \prod_{i=2}^{n} P(x^{i} | x^{(1)}, \cdots, x^{(i-1)})$$
    - For example:
$$p(a, b, c) = p(a|b, c) p(b, c)$$
$$p(b, c) = p(b|c) p(c)$$
$$p(a, b, c) = p(a|b, c) p(b|c) p(c)$$

1.6.2 Expectations

- Take two random variables X and Y with joint PDF $p(x, y)$, the expectation of their sum is
$$\mathrm{E}[X + Y] = [X] + [Y]$$

1.6.3 Correlated Random Variables

- **Covariance**: A measure of the jointly variability of two random variables
  - Gives some sense of how much two values are linearly related to each other
  - For two jointly distributed random variables X and Y, the covariance is defined as:

$$\mathrm{cov}(X,Y) = \mathrm{E}[(X - \mathrm{E}[X])(Y - \mathrm{E}[Y])]$$

  - Correlation is not causation
  - Two jointly distributed random variables X and Y is called:
    - positively correlated if $\mathrm{cov}(X,Y) > 0$
    - negatively correlated if $\mathrm{cov}(X,Y) < 0$
    - uncorrelated if $\mathrm{cov}(X,Y) \approx 0$
- **Properties of the covariance**:
  - Let X be a random variable:

$$\mathrm{cov}[X,X] = \mathrm{V}[X]$$

  - Let X be a random variable and take any constant λ:

$$\mathrm{cov}[X,\lambda] = 0$$

  - Let X and Y be two random variables:

$$\mathrm{cov}[Y,X] = \mathrm{cov}[X,Y]$$

  - Let X and Y be two random variables and take any constants:

$$\mathrm{cov}[\lambda X, \mu Y] = \lambda\mu \mathrm{cov}[X,Y]$$

  - Let X and Y be two random variables and take any constants:

$$\mathrm{cov}[X+\lambda, Y+\mu] = \mathrm{cov}[X,Y]$$

  - Let X, Y and Z be three random variables:

$$\mathrm{cov}[X, Y+Z] = \mathrm{cov}[X,Y] + \mathrm{cov}[X,Z]$$

  - Let X and Y be two random variable:

$$\mathrm{V}[X+Y] = \mathrm{V}[X] + \mathrm{V}[Y] + 2\mathrm{cov}[X,Y]$$

1.6.4 Independent Random Variables

- **Definition**: Two random variables X and Y are independent conditional on I, denoted as $X \perp Y|I$ if and only if conditioning on one does not tell anything about the other

$$p(x|y,I) = p(x|I)$$

  - When there is no ambiguity, I can be dropped
- **Properties**: Assume X and Y are two independent random variables, then:
  - $p(x,y) = p(x)p(y)$
  - $\mathrm{E}[XY] = \mathrm{E}[X]\mathrm{E}[Y]$
  - $\mathrm{cov}[X,Y] = 0$
    - The reverse is not true. Uncorrelated variables do not have to be independent
  - $\mathrm{V}[X+Y] = \mathrm{V}[X] + \mathrm{V}[Y]$

```
[ ]:
```