Probability Theory in Artificial Intelligence

Ipshita Bonhi Upoma

Department of Computer Science and Engineering
School of Data and Sciences

Probability Theory in AI

Why is Probability Theory Important in AI?

- Helps manage uncertainty in decision-making and predictions.
- Essential for models like:
  - Naive Bayes (classification)
  - Hidden Markov Models (sequence modeling)
  - Variational Autoencoders (generative models)
- Enables data-driven, informed decisions, even with noisy or incomplete data.

Basic Concepts

- **Experiment:** A process yielding outcomes (e.g., coin toss, dice roll).
- **Sample Space S:** Set of all possible outcomes, e.g.,

$$S = \{\text{Heads}, \text{Tails}\}$$

- **Event E:** Subset of outcomes, e.g., $E = \{\text{Tails}\}$.
- **Probability $P(E)$:** Likelihood of E, ranging from 0 (impossible) to 1 (certain).

Classical Probability

Classical Theory of Probability:

$$P(E) = \frac{\text{Number of desired possible outcomes}}{\text{Number of all equally possible outcomes}}.$$

Assumption: All outcomes have equal possibility.

Examples of Classical Probability

Example 1: Tossing a Coin

- Sample Space: $S = \{\text{Heads}, \text{Tails}\}$
- Event: $E = \{\text{Tails}\}$
- Probability: $P(E) = \frac{1}{2}$

Example 2: Rolling a Die

- Sample Space: $S = \{1, 2, 3, 4, 5, 6\}$
- Event: $E = \{3, 4, 5, 6\}$
- Probability: $P(E) = \frac{4}{6} = \frac{2}{3}$

Statistical Probability: Rolling a Die

How to Calculate Statistical Probability?

- Conduct the experiment multiple times.
- Collect data and tabulate the results.

Example: Calculating Probability of Rolling a 3 or More

- Experiment: Roll a die 600 times.
- Goal: Compute the probability of rolling a value of 3 or greater.

Results of Rolling a Die Experiment

Outcome (Value)	1	2	3	4	5	6
No. of Times Appeared	95	105	110	94	97	99

Table: Results of 600 Rolls of a Die

Probability Calculation:

$$P(E) = \frac{110 + 94 + 97 + 99}{600} = \frac{400}{600} = \frac{2}{3}.$$

Thus, the probability of rolling a 3 or greater is $\frac{2}{3}$.

Complement Rule

Definition: The probability that an event E does not occur is:

$$P(E^c) = 1 - P(E)$$

Addition Rule

Definition: For two events A and B, the probability that either A or B occurs is:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

Multiplication Rule

Definition: If two events A and B are independent, the probability that both events occur is:

$$P(A \cap B) = P(A) \times P(B)$$

Conditional Probability

Definition: The probability of event A occurring given that event B has occurred is:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}, \quad \text{provided } P(B) > 0.$$

Example of Conditional Probability

Example: A teacher knows that:

- 40 students passed a math test.
- 30 of these students attended an extra review session.

Find: The probability that a student attended the review session given they passed the test.

$$P(\text{Review}|\text{Pass}) = \frac{P(\text{Review and Pass})}{P(\text{Pass})}$$

$$P(\text{Pass}) = \frac{40}{100} = 0.4, \quad P(\text{Review and Pass}) = \frac{30}{100} = 0.3$$

$$P(\text{Review}|\text{Pass}) = \frac{0.3}{0.4} = 0.75$$

Product Rule

Definition: From conditional probability:

$$P(A \cap B) = P(A|B) \cdot P(B)$$

Chain Rule

Definition: The joint probability of multiple events $x_1, x_2, \ldots, x_n$ can be expressed as:

$$
\begin{aligned}
P(x_1, x_2, \ldots, x_n) &= \prod_{i=1}^{n} P(x_i | x_1, \ldots, x_{i-1}) && (1) \\
&= P(x_1)P(x_1 \mid x_2)P(x_1 \mid x_2, x_3) \ldots && (2)
\end{aligned}
$$

This rule helps compute complex joint probabilities step-by-step using conditional probabilities.

Bayes' Rule: Derivation

Step 1: Conditional Probabilities

$$P(A|B) = \frac{P(A \cap B)}{P(B)}, \quad P(B|A) = \frac{P(A \cap B)}{P(A)}.$$

Step 2: Rewrite Joint Probability

$$P(A \cap B) = P(A|B) \cdot P(B) = P(B|A) \cdot P(A).$$

Bayes' Rule:

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

Bayes' Rule

Formula:

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

- **Prior:** $P(A)$ — Initial probability of A.
- **Likelihood:** $P(B|A)$ — Probability of B given A.
- **Evidence:** $P(B)$ — Normalizing factor.

Significance of Bayes' Rule

Applications:

- **Weather Prediction:** Compute $P(\text{Rain}|\text{Cloudy})$.
- **Medical Diagnosis:** Find the probability of a disease given test results.
- Decision Tree algorithm.
- **NLP Applications:** Sentiment analysis and spam detection.

Bayes' Rule forms the foundation of Bayesian inference and is widely used in AI and statistics.

Example: Weather Prediction

Problem: Compute the probability of rain (A) given the sky is cloudy (B).

- $P(A) = 0.3$ (prior probability of rain).
- $P(B|A) = 0.8$ (likelihood: cloudy given rain).
- $P(B) = 0.6$ (evidence: probability of cloudy).

Solution:

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)} = \frac{0.8 \cdot 0.3}{0.6} = 0.4$$

Thus, the probability of rain given that it is cloudy is 0.4 or 40%.

Example: Sentiment Analysis

Problem: Find the probability that a tweet is positive (A) given that it contains the word "good" (B).

- $P(A) = 0.7$ (prior probability of positivity).
- $P(B|A) = 0.6$ (likelihood of "good" in positive tweets).
- $P(B) = 0.5$ (probability of "good" in any tweet).

Solution:

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)} = \frac{0.6 \cdot 0.7}{0.5} = 0.84$$

Thus, the probability that the tweet is positive is 84%.

Discrete Random Variables

Definition: A **discrete random variable** is a variable that can take on a countable number of distinct values. These values can be finite or infinite.

Examples:

- The number of heads in a series of coin tosses.
- The number of students passing an exam.

Notation for Discrete Random Variables

- **Random Variable:** Usually denoted by a capital letter, such as X, Y, or Z.
- **Possible Values:** Denoted by lowercase letters, such as $x_1, x_2, \ldots$.
- **Probability Mass Function (PMF):** The probability that the random variable X takes a specific value x_i is represented as:

$$P(X = x_i) \quad \text{or} \quad p(x_i).$$

- **Cumulative Distribution Function (CDF):** The cumulative probability up to a value x_i is represented as:

$$F(x_i) = P(X \leq x_i).$$

Probability Mass Function (PMF)

Definition: The PMF gives the probability distribution of a discrete random variable:

$$P(X = x_i) = \text{Probability of } X \text{ taking value } x_i.$$

Example: For a fair 6-sided die:

$$P(X = x) = \frac{1}{6}, \quad x \in \{1, 2, 3, 4, 5, 6\}.$$

Cumulative Distribution Function (CDF)

Definition: The CDF gives the cumulative probability:

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

Example: For a 6-sided die:

$$F(3) = P(X \leq 3) = P(X = 1) + P(X = 2) + P(X = 3) = \frac{3}{6}.$$

Vector Notation for Random Variables

- A set of n random variables can be represented as a **vector**:

$$\mathbf{X} = (X_1, X_2, \ldots, X_n)$$

  where X_i are individual random variables.

- The possible values of a random variable can also be represented as a **vector**:

$$\mathbf{x} = (x_1, x_2, \ldots, x_m)$$

  where x_i represents a specific value the random variable can take.

Example: Vector Notation for PMF

Example: Let X represent the outcome of rolling a fair six-sided die.

Vector of Possible Values:

$$\mathbf{x} = (1, 2, 3, 4, 5, 6)$$

Vector of Probabilities (PMF):

$$\mathbf{p} = \left(\frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}\right)$$

Cumulative Distribution Function (CDF):

$$\begin{aligned}\mathbf{F} &= (F(1), F(2), F(3), F(4), F(5), F(6)) \\ &= \left(\frac{1}{6}, \frac{1}{3}, \frac{1}{2}, \frac{2}{3}, \frac{5}{6}, 1\right)\end{aligned}$$

Discrete Random Variables with Categorical Outcomes

Example: Weather Conditions

- Let X represent the weather on a given day.

$$X = \{\text{Sunny}, \text{Cloudy}, \text{Rainy}\}$$

- **PMF:**

$$P(X = \text{Sunny}) = 0.5, \quad P(X = \text{Cloudy}) = 0.3, \quad P(X = \text{Rainy}) = 0.2$$

- **CDF:** The cumulative probabilities are:

$$F(\text{Sunny}) = 0.5, \quad F(\text{Cloudy}) = 0.8, \quad F(\text{Rainy}) = 1.0$$

Data to Probability

Example: TV Show Survey A survey is conducted among 500 people. The results are summarized below:

Category	Male	Female	Total
GOT	80	120	200
TBBT	100	25	125
Others	50	125	175
Total	230	270	500

Table: Survey Data on TV Show Preferences

Probability Distribution

Distribution:

- $P(\text{GOT}) = 0.4, \quad P(\text{TBBT}) = 0.25, \quad P(\text{Others}) = 0.35.$
- $P(\text{Male}) = 0.46, \quad P(\text{Female}) = 0.54.$

Probability distribution table from data

Category	Male	Female	Total
GOT	.16	.24	.40
TBBT	.2	.05	.25
Others	.1	.25	.35
Total	.46	.54	1

Table: Probability distribution table

Marginal Probability

Definition: Marginal probability, also known as simple probability, represents the probability of a single event occurring:

$$P(A) = \sum_{B} P(A \cap B)$$

Example 1: From the TV survey data:

$$P(\text{TBBT}) = 0.2 + 0.05 = 0.25$$

Example 2: Probability of being female:

$$P(\text{Female}) = 0.24 + 0.05 + 0.25 = 0.54$$

Marginal Probability Distribution

Category	Male	Female	Total
GOT	.16	.24	.40
TBBT	.2	.05	.25
Others	.1	.25	.35
Total	.46	.54	1

Table: Probability distribution table

Joint Probability

Definition: Joint probability is the probability of two or more events occurring simultaneously:

$$P(A \cap B)$$

Example: Probability of a person being female and liking TBBT:

$$P(\text{Female} \cap \text{TBBT}) = 0.05$$

Joint Probability Distribution

Definition: Joint Probability Distribution (JPD) lists all joint probabilities of events:

$$\sum_{i,j} P(A_i \cap B_j) = 1$$

Category	**Male**	**Female**	**Total**
GOT	0.16	0.24	0.40
TBBT	0.20	0.05	0.25
Others	0.10	0.25	0.35
Total	0.46	0.54	1.00

Table: Joint Probability Distribution

Conditional Probability Using JPD

Example: Probability of liking GOT given that the person is male:

$$P(\text{GOT}|\text{Male}) = \frac{P(\text{GOT} \cap \text{Male})}{P(\text{Male})} = \frac{0.16}{0.46} \approx 0.347$$

Conditional Independence

Definition: Two events A and B are conditionally independent given C if:

$$P(A \cap B|C) = P(A|C) \cdot P(B|C)$$

Absolute Independence:

- $P(A \cap B) = P(A) \cdot P(B)$
- Equivalently: $P(A) = P(A|B)$ and $P(B) = P(B|A)$

Example: Independence Check

Problem: Are male viewers and liking GOT independent?
Solution:

$$P(\text{Male} \cap \text{GOT}) = 0.16, \quad P(\text{Male}) = 0.46, \quad P(\text{GOT}) = 0.40$$

$$P(\text{Male}) \cdot P(\text{GOT}) = 0.46 \cdot 0.40 = 0.184$$

Conclusion: Since $P(\text{Male} \cap \text{GOT}) \neq P(\text{Male}) \cdot P(\text{GOT})$, they are **not independent**.

Example: Student Preparation

$p(\text{smart} \wedge \text{study} \wedge \text{prep})$	smart		¬smart	
	study	¬study	study	¬study
prepared	0.432	0.160	0.084	0.008
¬prepared	0.048	0.160	0.036	0.072

Table: Probability distribution for Student personality and preparation

- Is smart conditionally independent of prepared, given study?
- Is study conditionally independent of prepared, given smart?
- Is smart conditionally independent of prepared, given study?
- Is study conditionally independent of prepared, given smart?

Example: Is Smartness Independent of Preparation?

Check: Conditional Independence of Smartness and Preparation given Study.

$$P(\text{Smart} \cap \text{Prepared} | \text{Study}) = P(\text{Smart} | \text{Study}) \cdot P(\text{Prepared} | \text{Study})$$

Solution:

$$\begin{aligned} P(\text{Smart} \cap \text{Prepared} | \text{Study}) &= \frac{P(\text{Smart} \cap \text{Prepared} \cap \text{Study})}{P(\text{Study})} \\ &= \frac{0.432}{0.432 + 0.048 + 0.084 + 0.036} \end{aligned}$$

Example: Is Smartness Independent of Preparation?

Solution:

$$
\begin{aligned}
P(\text{Smart}|\text{Study}) &= \frac{P(\text{Smart} \cap \text{Study})}{P(\text{Study})} \\
&= \frac{0.432 + 0.048}{0.432 + 0.048 + 0.084 + 0.036} \\
P(\text{Prepared} \mid \text{Study}) &= \frac{P(\text{Prepared} \cap \text{Study})}{P(\text{Study})} \\
&= \frac{0.432 + 0.084}{0.432 + 0.048 + 0.084 + 0.036}
\end{aligned}
$$

Conditionally independent if

$$P(\text{Smart} \cap \text{Prepared}|\text{Study}) = P(\text{Smart}|\text{Study}) \cdot P(\text{Prepared}|\text{Study})$$

Example: Additional Checks

Questions to Explore:

- Is Smartness conditionally independent of Preparation, given Study?
- Is Study conditionally independent of Preparation, given Smartness?

Steps: 1. Calculate probabilities using the joint distribution table. 2. Check if the equality conditions for independence are satisfied:

$$P(A \cap B|C) = P(A|C) \cdot P(B|C)$$

Law of Total Probability

Definition: The law of total probability allows us to compute the probability of an event by summing over all possible conditions:

$$P(Y) = \sum_i P(Y|Z_i) \cdot P(Z_i)$$

Example:

$$P(A) = P(A \cap B) + P(A \cap \neg B)$$

$$P(A) = P(A|B) \cdot P(B) + P(A|\neg B) \cdot P(\neg B)$$

Example: Umbrella Problem

Problem: Compute $P(Y = 1)$, the probability of carrying an umbrella.
Given:

- Weather conditions (Z): Sunny, Cloudy, Rainy.
- $P(Y = 1|Z = 1) = 0.1, \quad P(Y = 1|Z = 2) = 0.3, \quad P(Y = 1|Z = 3) = 0.9.$
- $P(Z = 1) = 0.4, \quad P(Z = 2) = 0.3, \quad P(Z = 3) = 0.3.$

Solution: Umbrella Problem

Using the law of total probability:

$$P(Y = 1) = \sum_i P(Y = 1|Z = i) \cdot P(Z = i)$$

$$P(Y = 1) = (0.1 \cdot 0.4) + (0.3 \cdot 0.3) + (0.9 \cdot 0.3)$$

$$P(Y = 1) = 0.04 + 0.09 + 0.27 = 0.40$$

Thus, the probability of carrying an umbrella is $P(Y = 1) = 0.40$ or 40%.

Questions?

Thank you for your attention!