CHAPTER 1

PROBABILITY THEORY

1.1 Probability Theory in AI

Probability theory is essential for AI systems to make decisions, predictions, and inferences under uncertainty. It underpins machine learning algorithms, from classification (e.g., Naive Bayes) to generative models (e.g., Hidden Markov Models and Variational Autoencoders).
In Machine Learning, probability theory is used in handling uncertainty, making predictions, learning from data, and updating models with observations of new data. It allows systems to make informed, data-driven decisions, even when the data is noisy or incomplete.
Models like Bayesian networks, Gaussian mixtures, and Markov models use probability theory to represent complex relationships and uncertainty in data. Techniques like Bayesian inference enable models to update predictions as new data is observed, refining the model over time.

1.2 Basic Concepts

Experiment: A process that results in one outcome from a set of possible outcomes. For example, tossing a coin or rolling a die.

Sample Space (S): The set of all possible outcomes of an experiment. For a fair coin, the sample space is

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

Event (E): A subset of the sample space. An event may consist of one or more outcomes. For example, getting heads in a coin toss is an event.

Probability of an Event (P(E)): The measure of how likely an event is to occur. Probability values range from 0 (impossible event) to 1 (certain event).

Classical Theory of Probability:

$$P(E) = \frac{\text{Number of desired possible outcomes}}{\text{Number of all equally possible outcomes}}. \tag{1.1}$$

Assumption: All outcomes have equal possibility.

Example 1: Tossing a Coin

Experiment: Tossing a fair coin.

Sample Space (S): The set of all possible outcomes.

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

Event (E): Getting "Tails" on the toss.

$$E = \{\text{Tails}\}.$$

Probability of an Event $P(E)$: The probability of getting “Tails” on the toss (assuming a fair coin).

$$P(E) = \frac{1}{2}.$$

Example 2: Rolling a Die

Experiment: Rolling a fair six-sided die.

Sample Space (S):

$$S = \{1, 2, 3, 4, 5, 6\}.$$

Event (E): Rolling a 3 or greater.

$$E = \{3, 4, 5, 6\}.$$

Probability of an Event $P(E)$: The probability of rolling a 3 or greater.

$$P(E) = \frac{4}{6} = \frac{2}{3}.$$

Statistical Theory of Probability: Run the experiment a large number of times. Create a table and collect data accordingly.

Example 1: Calculating the statistical the probability of getting a 3 or more if we throw a die.

Throw the die 600 times and then calculate the average number of times a 3 or greater value appears. To calculate the probability of rolling a 3 or greater using the statistical approach, we can simulate rolling a die 600 times. The results are shown in the table below:

Value	1	2	3	4	5	6
No. of times appeared	95	105	110	94	97	99

Table 1.1: Number of times each value appeared in 600 rolls of a die

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

Atomic Theory: For any event A, its probability will be

$$0 \leq P(A) \leq 1. \tag{1.2}$$

The sum of the probabilities of all possible events is 1:

$$\sum P(E_i) = 1. \tag{1.3}$$

1.3 Basic Probability Rules

Complement Rule: The probability that event E does not occur is:

$$P(E^c) = 1 - P(E) \tag{1.4}$$

where E^c denotes the complement of event E (i.e., all outcomes where E does not happen).

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

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) \tag{1.5}$$

where $A \cap B$ is the event where both A and B occur.

Multiplication Rule: If two events A and B are independent (i.e., the outcome of one does not affect the other), the probability that both events occur is:

$$P(A \cap B) = P(A) \times P(B) \tag{1.6}$$

Conditional Probability: The probability of event A occurring given that event B has occurred is called *conditional probability* and is denoted by $P(A|B)$. It is calculated as:

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \quad \text{(provided } P(B) > 0\text{)} \tag{1.7}$$

This concept helps in situations where we want to know how likely one event is, given that we have additional information about another event. Here, $P(B)$ acts as a normalizing constant.

Example: Suppose you are a teacher at a school and you have information about students passing or failing a math test. Out of 100 students, 40 students passed the test. Additionally, you know that 30 students who passed the test also attended an extra review session.
You want to find the probability that a student attended the extra review session given that they passed the test.

Solution: We want to calculate the **Conditional probability**

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

From the given information we calculate,

$$P(\text{Pass}) = \frac{40}{100} = 0.4$$

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

Thus,

$$P(\text{Review}|\text{Pass}) = \frac{.3}{.4} = .75$$

This means if a student has passed, there is a 75% chance that they have attended the review session.

Product Rule: From the conditional probability (1.7) we can get,

$$P(A \cap B) = P(A|B)P(B) \tag{1.8}$$

Chain Rule: We can further use the conditional probability to find the joint probablity of multiple events, $x_1, x_2, \ldots x_n$.

$$\begin{aligned} P(x_1, x_2, \ldots x_n) &= P(x_1)p(x_1|x_2)P(x_1|x_2, x_3)\ldots \qquad (1.9) \\ &= \prod_{i}^{n} P(x_i)P(x_1|x_1, \ldots, x_{i-1}) \qquad (1.10) \end{aligned}$$

1.4 Bayes' Rule: Derivation and Examples

1.4.1 Derivation of Bayes' Rule

Bayes' Rule allows us to update our beliefs based on new evidence. It is derived from the definition of conditional probability. Recall that the conditional probability of event A given event B is:

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \tag{1.11}$$

Similarly, the conditional probability of event B given event A is:

$$P(B|A) = \frac{P(A \cap B)}{P(A)} \tag{1.12}$$

From these two definitions, we can express $P(A \cap B)$ as:

$$P(A \cap B) = P(A|B) \cdot P(B) = P(B|A) \cdot P(A) \tag{1.13}$$

Rearranging this to solve for $P(A|B)$, we get:

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)} \tag{1.14}$$

This is the famous Bayes' Rule.

1.4.2 Significance of Bayes' Rule

Bayes' Rule is a powerful tool for updating the probability of an event based on new evidence. It is widely used in various fields, including weather prediction, medical diagnosis, and NLP tasks such as sentiment analysis.

Weather Example Using Bayes' Rule

Consider a weather prediction scenario. We want to compute the probability that it will rain tomorrow, (A) given that the sky is cloudy, (B).

From observed data, we are given the following information:

- $P(A) = 0.30$ (the prior probability that it will rain),
- $P(B|A) = 0.80$ (the likelihood: the probability that the sky will be cloudy given that it rains),
- $P(B) = 0.60$ (the probability that the sky is cloudy, regardless of whether it rains or not).

We want to calculate $P(A|B)$, the probability that it will rain given that the sky is cloudy.

Bayes' Rule tells us:

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)} = \frac{0.80 \cdot 0.30}{0.60} = \frac{0.24}{0.60} = 0.40 \tag{1.15}$$

Thus, the probability that it will rain tomorrow given that the sky is cloudy is 0.40, or 40%.

Example of Sentiment Analysis using Bayes' Rule

Bayes' Rule can also be applied in Natural Language Processing (NLP), for instance, in sentiment analysis, where we want to classify a piece of text as either positive or negative.

For example, we want to classify a tweet as either positive or negative based on the presence of the word "good".

Let's define the following events: - A = Positive sentiment - B = The word "good" appears in the tweet

We want to calculate $P(A|B)$, the probability that a tweet has a positive sentiment, given that the word "good" appears.

From observed data, we know:

- $P(A) = 0.70$ (prior probability that the tweet is positive),
- $P(B|A) = 0.60$ (likelihood: the probability that the word "good" appears in a positive tweet),
- $P(B) = 0.50$ (probability that the word "good" appears in any tweet, regardless of sentiment).

Using Bayes' Rule, we calculate:

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)} = \frac{0.60 \cdot 0.70}{0.50} = \frac{0.42}{0.50} = 0.84 \tag{1.16}$$

Thus, the probability that the tweet has a positive sentiment, given that the word "good" appears, is 0.84 or 84%.

1.5 Discrete Random Variables

A discrete random variable is a variable that can take on a countable number of distinct values. These values may be finite or infinite but are always countable. Discrete random variables typically arise in situations where the outcomes are distinct and can be listed, such as the number of heads in a series of coin tosses or the number of students passing an exam.

1.5.1 Notation for Discrete Random Variables

Random Variable: A discrete random variable is usually denoted by a capital letter, such as X, Y, or Z.

Possible Values: The possible values that a discrete random variable can take are denoted by lowercase letters, such as $x_1, x_2, x_3, \ldots$.

Probability Mass Function (PMF): The probability that the random variable X takes a specific value x_i is represented as $P(X = x_i)$ or $p(x_i)$. The Probability Mass Function (PMF) gives the probability distribution of a discrete random variable.

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

$$F(x_i) = P(X \leq x_i) \tag{1.17}$$

which sums the probabilities for all values less than or equal to x_i.

1.5.2 Properties of Discrete Random Variables

- **Non-Negativity**: For any x, $P(X = x) \geq 0$.
- **Normalization**: The sum of the probabilities for all possible values must equal to 1:

$$\sum_x P(X = x) = 1 \tag{1.18}$$

- **Countability**: The possible values $x_1, x_2, x_3, \ldots$ must be countable.

1.5.3 Basic Vector Notation

Vector of Random Variables: A set of n random variables can be represented as a vector. For example, a vector $\mathbf{X}$ of random variables is often written as:

$$\mathbf{X} = (X_1, X_2, \ldots, X_n) \tag{1.19}$$

where X_i are the individual random variables.

Vector of Possible Values: A random variable X can take a set of discrete values, which can also be represented as a vector of values. For example:

$$\mathbf{x} = (x_1, x_2, \ldots, x_m) \tag{1.20}$$

where x_i represents a specific outcome or value of the random variable.

Probability Mass Function (PMF): The PMF can be represented as a vector of probabilities. If X is a discrete random variable taking values $x_1, x_2, \ldots, x_m$, the PMF is a vector of probabilities:

$$\mathbf{p} = (p(x_1), p(x_2), \ldots, p(x_m)) \tag{1.21}$$

where $p(x_i) = P(X = x_i)$ is the probability that X takes the value x_i.

Example: Let X be a discrete random variable representing the outcome of rolling a fair 6-sided die. The possible outcomes for X are the integers from 1 to 6. The vector notation representing the values the discrete random variable can take is:

$$\mathbf{x} = (1, 2, 3, 4, 5, 6) \tag{1.22}$$

Since the die is fair, the probability of each outcome is equal. Thus, the probability of each outcome is:

$$P(X = x) = \frac{1}{6} \quad \text{for each} \quad x \in \{1, 2, 3, 4, 5, 6\} \tag{1.23}$$

We can represent the PMF as a vector of probabilities:

$$\mathbf{P} = \left(\frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}\right) \tag{1.24}$$

Cumulative Distribution Function (CDF): The CDF gives the cumulative probability up to each value of X. The CDF vector is denoted as:

$$\mathbf{F} = (F(1), F(2), F(3), F(4), F(5), F(6)) \tag{1.25}$$

1.5.4 Discrete Random Variable with Categorical Outcomes

A discrete random variable can take categorical outcomes, where the outcomes are not numeric. For example, consider a random variable X representing the weather, where the possible outcomes are $\{\text{Sunny}, \text{Cloudy}, \text{Rainy}\}$.

Example: Let X be the random variable representing the weather on a given day. The possible outcomes are:

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

The Probability Mass Function (PMF) for X is given by:

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

Cumulative Distribution Function (CDF): The CDF for X is calculated as the cumulative probability up to each outcome. Since the outcomes are ordered as "Sunny", "Cloudy", and "Rainy", the CDF is:

$$F(\text{Sunny}) = P(X \leq \text{Sunny}) = P(X = \text{Sunny}) = 0.5$$

$$F(\text{Cloudy}) = P(X \leq \text{Cloudy}) = P(X = \text{Sunny}) + P(X = \text{Cloudy}) = 0.5 + 0.3 = 0.8$$

$$F(\text{Rainy}) = P(X \leq \text{Rainy}) = P(X = \text{Sunny}) + P(X = \text{Cloudy}) + P(X = \text{Rainy}) = 0.5 + 0.3 + 0.2 = 1$$

Thus, the CDF for X is:

$$F(X) = \begin{cases} 0.5 & \text{if } X = \text{Sunny} \\ 0.8 & \text{if } X = \text{Cloudy} \\ 1 & \text{if } X = \text{Rainy} \end{cases} \tag{1.26}$$

1.6 Data to Probability

Example: A survey is conducted on 500 people both male and female about which tv shows they liked. From the 500 people we got the following response.

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

Table 1.2: Survey Data for TV Shows Viewership by Gender

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

Table 1.3: Probability distribution from Survey Data for TV Show Viewership by Gender

1.6.1 Marginal Probability

Marginal probability, also known as **Simple Probability** represents the probability of a particular event.

The marginal probability is obtained by summing the joint probabilities across the rows or columns.

Example: For the joint probability distribution given in 1.3 find the probability of a viewer watching TBBT.

Solution: Sum over all the values of row TBBT.

$$P(\text{TBBT}) = .2 + .05 = .25.$$

This is the value on the cell for TBBT on the Total column.

1.6.2 Marginal Probability Distribution

All the probability that occurs in the margin for a particular variable/event can be summed up to create Marginal Probability Distribution. For example, the "Total" column and "Total" row in 1.3.

1.6.3 Joint Probability

Joint Probability is the probability of two or more events occurring at the same time. For two events, this is $P(A \cap B)$. We can easily calculate it from the probability distribution.

Example: What is the joint probability of a person being Female and liking TBBT? **Solution:** We can find it directly from the probability distribution table.

$$P(T^F) = 0.05$$

1.6.4 Joint Probability Distribution

All the probability that occur jointly can be summed up to create Joint Probability Distribution (JPD). Sum of JPD is 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 1.4: Joint Probability Distribution Table is highlighted

1.7 Understanding Conditional Probability using Probability Distribution Table

We can calculate conditional probability using the joint distribution table.

Example: What is the probability of a person liking GOT given that person is male? **Ans:** Using values directly from table 1.3

$$\begin{aligned} P(G|M) &= \frac{P(G \cap M)}{P(M)} \\ &= \frac{0.16}{0.46} \\ &= 0.347 \end{aligned}$$

1.7.1 Conditional Independence

No effect of one variable on another variable. $P(A|B) = P(A)$
$P(A \cap B) = P(A|B) * P(B)$ [Conditional Probability]
Absolute independence: A and B are independent if $P(A \cap B) = P(A)P(B)$; equivalently, $P(A) = P(A|B)$ and $P(B) = P(B|A)$.
A and B are conditionally independent given C if $P(A \cap B|C) = P(A|C)P(B|C)$.
This lets us decompose the joint distribution: $P(A \cap B \cap C) = P(A|C)P(B|C)P(C)$.
Are Male viewers and GOT independent? Ans:

$$\begin{aligned} P(M \cap GOT) &= 0.16 \\ P(M) &= 0.46 \\ P(GOT) &= 0.40 \\ P(M) * P(GOT) &= 0.46 * 0.40 = 0.184 \end{aligned}$$

Since, $P(M \cap GOT) \neq P(M) * P(GOT)$ so, not independent.

p(smart ∧ study ∧ 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 1.5: 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?

$$\begin{aligned} P(Sm \cap Pr|Sd) &= P(Sm|Sd) * P(Pr|Sd) \\ P(Sm \cap Pr|Sd) &= P(Sm \cap Pr \cap Sd)/P(Sd) \\ &= 0.432/0.6 \\ &= 0.72 \end{aligned}$$

1.8 Law of Total Probability

The law of total probability allows us to compute the probability of an event by considering all possible conditions (or outcomes) of a related variable. It states that the probability of an event Y is the sum of the conditional probabilities of Y given each possible outcome of another variable Z, weighted by the probability of each outcome of Z. For example,

$$P(a) = P(a \cap b) + P(a \cap \neg b) \tag{1.27}$$
$$= P(a|b)p(b) + P(a|\neg b)\text{using conditional probability} \tag{1.28}$$

More generally,

$$P(Y = i) = \sum_z P(Y = i|Z = z)P(z)$$

Example: Let Y be the event of carrying an umbrella, and Z be the weather condition. The possible values of Z are:

- $Z = 1$: Sunny
- $Z = 2$: Cloudy
- $Z = 3$: Rainy

The conditional probabilities are:

$$P(Y = 1 \mid Z = 1) = 0.1, \quad P(Y = 1 \mid Z = 2) = 0.3, \quad P(Y = 1 \mid Z = 3) = 0.9$$

The probabilities of the weather conditions are:

$$P(Z = 1) = 0.4, \quad P(Z = 2) = 0.3, \quad P(Z = 3) = 0.3$$

We can compute $P(Y = 1)$ using the law of total probability:

$$P(Y = 1) = P(Y = 1 \mid Z = 1)P(Z = 1) + P(Y = 1 \mid Z = 2)P(Z = 2) + P(Y = 1 \mid Z = 3)P(Z = 3)$$

Substituting the values:

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

$$P(Y = 1) = 0.04 + 0.09 + 0.27 = 0.4$$

Thus, the probability that the person carries an umbrella is $P(Y = 1) = 0.4$.