Module-1
Information Measures

Dr. Markkandan S

School of Electronics Engineering (SENSE)
Vellore Institute of Technology
Chennai

Outline

Introduction to Information Theory

Introduction to Information Theory

- Information Theory, developed by Claude Shannon in 1948, is a mathematical framework for quantifying information, studying data compression, error correction, and cryptography.
- It deals with the fundamental limits on data compression and communication, providing tools to measure information content, uncertainty, and the relationships between variables.
- Key concepts include entropy, mutual information, channel capacity, and coding theory, with applications in communication systems, data science, and cryptography.
- Information Theory has profound implications in various fields, from telecommunications to neuroscience, shaping modern technologies and communication systems.

Review of Probability Theory

Review of Probability Theory

- Probability theory provides the foundation for Information Theory, describing the behavior of uncertain events and random variables.
- Random Variables (RVs) are variables whose values depend on the outcome of a random experiment.
- Probability Distributions describe the likelihood of different outcomes of a random variable.
- Expectation (E), also known as the mean, represents the average value of a random variable.
- Variance (Var) measures the dispersion of values around the mean.
- Covariance (Cov) measures the degree to which two random variables change together.

Review of Probability Theory

- **Sample Space**: The set of all possible outcomes of a random experiment, denoted by Ω.
- **Events**: Subsets of the sample space, denoted by $A, B, C, \ldots$.
- **Probability Axioms**:
  1. Non-negativity: $P(A) \geq 0$ for all events A.
  2. Normalization: $P(\Omega) = 1$.
  3. Additivity: For mutually exclusive events A and B, $P(A \cup B) = P(A) + P(B)$.
- **Probability Distributions**:
  - **Discrete**: Assigns probabilities to individual outcomes.
  - **Continuous**: Described by probability density functions (PDFs).

Probability Distributions

- **Discrete Probability Distribution**:
  - Assigns probabilities to individual outcomes of a discrete random variable.
  - Probability mass function (PMF): $P(X = x)$ for each possible value x.
  - Example: Coin toss, dice roll.
- **Continuous Probability Distribution**:
  - Describes probabilities over intervals for continuous random variables.
  - Probability density function (PDF): $f(x)$ such that $P(a \leq X \leq b) = \int_a^b f(x)\,dx$.
  - Example: Normal distribution, exponential distribution.
- **Expectation and Variance**:
  - **Expectation**: Average value of a random variable, denoted by $\mathbb{E}[X]$.
  - **Variance**: Measure of dispersion, denoted by $\mathrm{Var}[X]$.
  - Calculation: $\mathbb{E}[X] = \sum_x xP(X = x)$ for discrete, $\mathbb{E}[X] = \int_{-\infty}^{\infty} xf(x)\,dx$ for continuous.

Conditional Probability and Bayes' Theorem

- **Conditional Probability**:
  - Probability of an event A given that another event B has occurred, denoted by $P(A|B)$.
  - Calculated as $P(A|B) = \frac{P(A \cap B)}{P(B)}$ if $P(B) > 0$.
  - Example: Probability of rain given cloudy skies.
- **Bayes' Theorem**:
  - Relates conditional probabilities of events.
  - States: $P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}$.
  - Useful for updating beliefs based on new evidence.
  - Widely used in statistics, machine learning, and data analysis.

Self-Information and Average Information

Self-Information and Average Information

- **Self-Information** ($I(x)$) measures the amount of surprise or uncertainty associated with a specific outcome x of a random variable X.
- It is defined as:

$$I(x) = -\log(P(x))$$

  where $P(x)$ is the probability of occurrence of outcome x.
- For example, if the probability of raining tomorrow is 0.2, then the self-information associated with this event is $-\log(0.2) \approx 2.32$ bits.
- **Average Information** ($H(X)$) represents the expected amount of information conveyed by a random variable X.
- It is calculated as the expected value of self-information over all possible outcomes:

$$H(X) = E[I(x)] = -\sum_{i=1}^{n} P(x_i) \log(P(x_i))$$

  where x_i are the possible outcomes of X.
- Average information provides a measure of the uncertainty

Mutual Information

Mutual Information

- **Definition**: Mutual information measures the amount of information that one random variable contains about another random variable.
- **Formula**: For two discrete random variables X and Y, mutual information $I(X;Y)$ is given by:

$$I(X;Y) = \sum_{x_i} \sum_{y_j} P(X = x_i, Y = y_j) \log \left(\frac{P(X = x_i, Y = y_j)}{P(X = x_i) P(Y = y_j)} \right)$$

- **Properties**:
  - Symmetric: $I(X;Y) = I(Y;X)$
  - Non-negative: $I(X;Y) \geq 0$
  - Zero mutual information implies independence

Mutual Information

- For continuous random variables, the summation is replaced by integration:

$$I(X;Y) = \iint f(x,y)\log\left(\frac{f(x,y)}{f_X(x)f_Y(y)}\right)\,dx\,dy$$

- Mutual information provides insights into the statistical dependence between variables and is used in various applications such as feature selection and clustering.

Example Problem: Mutual Information

Consider two random variables X and Y with a joint probability distribution given by:

$$P(X, Y) = \begin{cases} \frac{1}{4}, & \text{for } X = 0 \text{ and } Y = 0, \\ \frac{1}{4}, & \text{for } X = 1 \text{ and } Y = 0, \\ \frac{1}{4}, & \text{for } X = 0 \text{ and } Y = 1, \\ \frac{1}{4}, & \text{for } X = 1 \text{ and } Y = 1. \end{cases}$$

Calculate the mutual information $I(X; Y)$.

Solution:

$$I(X; Y) = -\sum_{x,y} P(x, y) \log \left(\frac{P(x, y)}{P(x)P(y)} \right)$$

$$I(X; Y) = -\left(\frac{1}{4} \log \left(\frac{1}{4} \right) + \frac{1}{4} \log \left(\frac{1}{4} \right) + \frac{1}{4} \log \left(\frac{1}{4} \right) + \frac{1}{4} \log \left(\frac{1}{4} \right) \right)$$

$$I(X; Y) = -4 \times \frac{1}{4} \log \left(\frac{1}{4} \right) = -\log \left(\frac{1}{4} \right) = 2 \text{ bits}$$

Entropy

Entropy

- **Definition**: Entropy is a measure of uncertainty or randomness in a probability distribution.
- It is defined as the average amount of information generated by a random variable:
- **Shannon's Entropy**: For a discrete random variable X with probability mass function $P(X = x_i)$, Shannon's entropy is defined as:

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

where $P(x_i)$ is the probability of the i-th outcome of X.

- **Properties**:
  - Non-negativity: $H(X) \geq 0$
  - Maximum entropy: Uniform distribution
  - Minimum entropy: Deterministic distribution

Entropy

- **Entropy** ($H(X)$) is a fundamental concept in Information Theory, quantifying the uncertainty or randomness associated with a random variable X.
- Entropy is measured in bits for discrete random variables and nats or bits for continuous random variables.
- High entropy implies high uncertainty, while low entropy indicates low uncertainty or high predictability.
- For example, a fair coin toss has maximum entropy ($H(X) = 1$ bit), indicating maximum uncertainty, whereas a biased coin has lower entropy, reflecting higher predictability.
- Entropy plays a crucial role in various applications, including data compression, cryptography, and communication systems.

Example Problems and Solutions

Example Problem 1:

Suppose we have a random variable X representing the outcome of rolling a fair six-sided die. Calculate the self-information and average information associated with each outcome.

Solution:

Given that the die is fair, each outcome has a probability of $P(x_i) = \frac{1}{6}$.

1. Self-Information:

$$I(x_i) = -\log\left(\frac{1}{6}\right) = \log(6) \approx 2.58 \text{ bits}$$

2. Average Information:

$$H(X) = -\sum_{i=1}^{6} \frac{1}{6} \log\left(\frac{1}{6}\right) = \log(6) \approx 2.58 \text{ bits}$$

Example Problems and Solutions

Example Problem 2:
Consider a biased coin with a probability of heads (H) being 0.8 and tails (T) being 0.2. Calculate the entropy of the coin.
Solution:
Given $P(H) = 0.8$ and $P(T) = 0.2$.
Entropy is given by:

$$H(X) = -P(H)\log(P(H)) - P(T)\log(P(T))$$

$$H(X) = -(0.8)\log(0.8) - (0.2)\log(0.2)$$

$$H(X) \approx -(0.8)(-0.322) - (0.2)(-1.609)$$

$$H(X) \approx 0.257 + 0.322$$

$$H(X) \approx 0.579 \text{ bits}$$

Example Problem 3:
Consider a continuous random variable X with a probability density function (pdf) given by $f(x) = \frac{1}{\pi(1+x^2)}$ for $-\infty < x < \infty$. Calculate the entropy of X.
Solution:
The entropy for a continuous random variable is given by:

$$H(X) = -\int_{-\infty}^{\infty} f(x) \log(f(x))\, dx$$

Calculating the integral may involve more advanced techniques such as substitution or integration by parts, making it a challenging problem.

Example Problems and Solutions (continued)

Example Problem 4:

A communication channel has three possible symbols: A, B, and C. The probabilities of transmitting these symbols are $P(A) = 0.4$, $P(B) = 0.3$, and $P(C) = 0.3$, respectively. Calculate the entropy of the communication channel.

Solution:

Given probabilities $P(A) = 0.4$, $P(B) = 0.3$, and $P(C) = 0.3$.

The entropy is calculated as:

$$H(X) = -\sum_{i=1}^{3} P(x_i) \log(P(x_i))$$

$$H(X) = -(0.4)\log(0.4) - (0.3)\log(0.3) - (0.3)\log(0.3)$$

$$H(X) \approx -(0.4)(-0.699) - (0.3)(-1.204) - (0.3)(-1.204)$$

$$H(X) \approx 0.279 + 0.361 + 0.361$$

$$H(X) \approx 1.001 \text{ bits}$$

Example Problems and Solutions (continued)

Example Problem 5:

A continuous random variable X has a uniform distribution on the interval $[0, 1]$. Calculate the entropy of X.

Solution:

Given that X follows a uniform distribution on $[0, 1]$, the probability density function is $f(x) = 1$ for $0 \leq x \leq 1$.

The entropy for a continuous random variable is given by:

$$H(X) = -\int_0^1 f(x) \log(f(x))\, dx$$

$$H(X) = -\int_0^1 1 \cdot \log(1)\, dx$$

$$H(X) = -\int_0^1 0\, dx$$

$$H(X) = 0$$

The entropy of X is 0 bits, indicating no uncertainty as X is completely determined.

Marginal Entropy

Marginal Entropy

- **Marginal Entropy** $H(X)$ represents the average uncertainty of a single random variable without considering any other variables.
- For a discrete random variable X with joint probability distribution $P(X, Y)$, the marginal entropy is computed by summing the entropy over all possible values of X:

$$H(X) = -\sum_{x} P(x) \sum_{y} P(y|x) \log P(y|x)$$

- For a continuous random variable X with joint probability density function $f(x, y)$, the marginal entropy is computed by integrating over all possible values of X:

$$H(X) = -\int_{x} \int_{y} f(x, y) \log f(x, y)\, dy\, dx$$

- Marginal entropy provides insights into the uncertainty associated with individual variables, disregarding their relationship with other variables.

Example Problem: Marginal Entropy

Example Problem:
Consider two random variables X and Y with a joint probability distribution given by:

$$P(X,Y) = \begin{cases} \frac{1}{4}, & \text{for } X = 0 \text{ and } Y = 0, \\ \frac{1}{4}, & \text{for } X = 1 \text{ and } Y = 0, \\ \frac{1}{4}, & \text{for } X = 0 \text{ and } Y = 1, \\ \frac{1}{4}, & \text{for } X = 1 \text{ and } Y = 1. \end{cases}$$

Calculate the marginal entropy $H(X)$.

Example Problem: Marginal Entropy (Solution)

To calculate the marginal entropy $H(X)$, we need to sum over all possible values of Y and compute the entropy for each value of X.
For $X = 0$:

$$P(X = 0) = P(X = 0, Y = 0) + P(X = 0, Y = 1) = \frac{1}{4} + \frac{1}{4} = \frac{1}{2}$$

For $X = 1$:

$$P(X = 1) = P(X = 1, Y = 0) + P(X = 1, Y = 1) = \frac{1}{4} + \frac{1}{4} = \frac{1}{2}$$

Now, we can calculate the entropy for each value of X:

$$H(X = 0) = -\frac{1}{2}\log\left(\frac{1}{2}\right) - \frac{1}{2}\log\left(\frac{1}{2}\right) = 1 \text{ bit}$$

$$H(X = 1) = -\frac{1}{2}\log\left(\frac{1}{2}\right) - \frac{1}{2}\log\left(\frac{1}{2}\right) = 1 \text{ bit}$$

Finally, the marginal entropy $H(X)$ is the sum of the entropies for each value of X:

$$H(X) = 1 + 1 = 2 \text{ bits}$$

Joint Entropy and Conditional Entropy

Joint Entropy and Conditional Entropy

- **Joint Entropy** $H(X, Y)$ measures the uncertainty associated with two random variables X and Y considered together.
- It is calculated similarly to marginal entropy but with the joint probability distribution or density function.
- **Conditional Entropy** $H(Y|X)$ measures the average uncertainty of Y given the value of X.
- It is calculated as the difference between the joint entropy and the marginal entropy of X:

$$H(Y|X) = H(X, Y) - H(X)$$

- Conditional entropy quantifies the remaining uncertainty in Y after observing X.
- These measures play a crucial role in understanding the relationships and dependencies between random variables.

Joint Entropy

- **Joint Entropy** $H(X, Y)$ measures the uncertainty associated with two random variables X and Y considered together.
- It is calculated similarly to marginal entropy but with the joint probability distribution or density function.

$$H(X, Y) = -\sum_{x}\sum_{y} P(x, y) \log(P(x, y))$$

(for discrete case)

$$H(X, Y) = -\int_{x}\int_{y} f(x, y) \log f(x, y) \, dy \, dx$$

(for continuous case)

- Joint entropy provides insights into the total uncertainty associated with the joint behavior of two variables.

Example Problems: Joint Entropy

Consider two random variables X and Y with a joint probability distribution given by:

$$P(X,Y) = \begin{cases} \frac{1}{4}, & \text{for } X = 0 \text{ and } Y = 0, \\ \frac{1}{4}, & \text{for } X = 1 \text{ and } Y = 0, \\ \frac{1}{4}, & \text{for } X = 0 \text{ and } Y = 1, \\ \frac{1}{4}, & \text{for } X = 1 \text{ and } Y = 1. \end{cases}$$

Calculate the joint entropy $H(X,Y)$.

Solution:

The joint entropy $H(X,Y)$ is calculated using the formula:

$$H(X,Y) = -\sum_x \sum_y P(x,y) \log(P(x,y))$$

$$H(X,Y) = -\left(\frac{1}{4}\log\left(\frac{1}{4}\right) + \frac{1}{4}\log\left(\frac{1}{4}\right) + \frac{1}{4}\log\left(\frac{1}{4}\right) + \frac{1}{4}\log\left(\frac{1}{4}\right)\right)$$

$$H(X,Y) = -4 \times \frac{1}{4}\log\left(\frac{1}{4}\right) = -\log\left(\frac{1}{4}\right) = 2 \text{ bits}$$

Conditional Entropy

- **Conditional Entropy** $H(Y|X)$ measures the average uncertainty of Y given the value of X.
- It is calculated as the difference between the joint entropy and the marginal entropy of X:

$$\boxed{H(Y|X) = H(X,Y) - H(X)}$$

- Conditional entropy quantifies the remaining uncertainty in Y after observing X.
- These measures play a crucial role in understanding the relationships and dependencies between random variables.

Example Problems: Conditional Entropy

Example Problem:
Consider two random variables X and Y with a joint probability distribution given by:

$$P(X,Y) = \begin{cases} \frac{1}{4}, & \text{for } X = 0 \text{ and } Y = 0, \\ \frac{1}{4}, & \text{for } X = 1 \text{ and } Y = 0, \\ \frac{1}{4}, & \text{for } X = 0 \text{ and } Y = 1, \\ \frac{1}{4}, & \text{for } X = 1 \text{ and } Y = 1. \end{cases}$$

Calculate the conditional entropy $H(Y|X)$.
Solution:
To calculate $H(Y|X)$, we need to find $H(X,Y)$ and $H(X)$ first. From previous slides, $H(X) = 2$ bits.

$$H(Y|X) = H(X,Y) - H(X) = 2 - 2 = 0 \text{ bits}$$

Relationship between Entropy and Mutual Information

Relationship between Entropy and Mutual Information

- **Entropy and Mutual Information**:
  - Entropy measures the average uncertainty in a single random variable.
  - Mutual information measures the amount of information shared between two random variables.
- **Connection**:
  - Mutual information $I(X; Y)$ between X and Y can be expressed in terms of their entropies as:

$$I(X; Y) = H(X) + H(Y) - H(X, Y)$$

- **Interpretation**:
  - Mutual information quantifies the reduction in uncertainty about one variable due to the knowledge of the other variable.

Markov Statistical Model for Information Source

Markov Statistical Model for Information Source

- **Markov Chain**:
  - A stochastic model describing a sequence of possible events in which the probability of each event depends only on the state attained in the previous event.
  - Characterized by a transition probability matrix P, where P_{ij} represents the probability of transitioning from state i to state j.
- **Markov Source Model**:
  - Extension of the Markov chain concept to information theory.
  - Each symbol emitted by the source depends only on the previous symbol.
  - Provides a way to model correlated data sources.

Markov Process

A discrete stochastic process $X_1, X_2, \ldots$ is said to be Markov Chain or a Markov Process if, for n=1,2,...

$$P(X_{n+1} = x_{n+1} | X_n = x_n, X_{n-1} = x_{n-1}, \ldots, X_1 = x_1) = P(X_{n+1} = x_{n+1} | X_n = x_n)$$

for all $x_1, x_2, ...x_n, x_{n+1} \in X$.
The probability density function of a Markov process can be written as

$$p(x_1, x_2, \ldots x_n) = p(x_1)p(x_2|x_1)p(x_3|x_2)\ldots p(x_n|x_{n-1})$$

Entropy Rate

The Entropy Rate of a stochastic process X is given by

$$H(X) = \lim_{n\to\infty} \frac{1}{n} H(X_1, X_2, \ldots X_n)$$

provided the limit exists.

Example: Two-State Markov Chain

Consider a two-state Markov chain with a probability transition matrix

$$P = \begin{bmatrix} 1-p_1 & p_1 \\ p_2 & 1-p_2 \end{bmatrix}$$

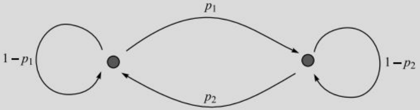

Fig. 1.18 The state transition graph of a two-state Markov chain.

For stationary distribution, the net probability distribution across any cut set in the state transition graph should be zero. Let α and β be the stationary probabilities of the two states. Thus, the stationary distribution is given by

$$\alpha = \frac{p_2}{p_1 + p_2} \quad \text{and} \quad \beta = \frac{p_1}{p_1 + p_2} \tag{1.85}$$

Note that $\alpha + \beta = 1$. The entropy of the state X_n at time n will be

$$H(X_n) = H\left(\frac{p_2}{p_1 + p_2}, \frac{p_1}{p_1 + p_2}\right) = H(\alpha, \beta) \tag{1.86}$$

The entropy rate of this two-state Markov chain is given by

$$H(X) = H(X_2 \mid X_1) = \frac{p_2}{p_1 + p_2} H(p_1) + \frac{p_1}{p_1 + p_2} H(p_2) \tag{1.87}$$

Entropy and Information Rate of Markov Source

- **Markov Source** is a type of information source where the probability of generating a symbol depends only on the previous symbol or symbols.
- The entropy H of the Markov source is given by:

$$H = -\sum_{i} \pi_i \sum_{i,j} P_{ij} \log P_{ij}$$

where p_i is the transition probability of symbol i. π_i is stationary distribution

- The **information rate** of a Markov source is the average rate of information production and is given by:

$$R = H(X)/T$$

where T is the time interval.

Example Problem:
Consider a binary information source with the following transition probabilities:

$$P = \begin{pmatrix} 0.7 & 0.3 \\ 0.4 & 0.6 \end{pmatrix}$$

Calculate the entropy H of the Markov source.
Solution:
The stationary distribution π of the Markov chain can be obtained by solving $\pi P = \pi$. Solving the equations, we get $\pi = \begin{pmatrix} 0.5714 & 0.4286 \end{pmatrix}$.
The entropy H of the Markov source is given by:

$$H = -\sum_{i} \pi_i \sum_{i,j} P_{ij} \log P_{ij}$$

Substituting the values, we get:

$$H = -(0.5714 \times 0.7 \log 0.7 + 0.5714 \times 0.3 \log 0.3 + 0.4286 \times 0.4 \log 0.4 + 0.4286 \times$$

$$\approx 0.895 \text{ bits}$$

Example Problem: Markov Source Model

Example Problem:
Consider a binary Markov source with transition probabilities $P(0|0) = 0.7$, $P(1|0) = 0.3$, $P(0|1) = 0.4$, and $P(1|1) = 0.6$. Find the entropy of the source.

Solution:
To find the entropy of the Markov source, we first need to determine the stationary distribution π of the Markov chain. Then, we can use the entropy formula to calculate the entropy.
The stationary distribution π satisfies the equation $\pi = \pi P$, where P is the transition probability matrix. we find that $\pi = (0.5714, 0.4286)$.
The entropy of the Markov source is then given by:

$$\begin{aligned} H(X) &= -\sum_i \pi_i \sum_j P(i|j) \log_2 P(i|j) \\ &= -(0.5714 \cdot 0.7 \log_2 0.7 + 0.5714 \cdot 0.3 \log_2 0.3 \\ &\quad + 0.4286 \cdot 0.4 \log_2 0.4 + 0.4286 \cdot 0.6 \log_2 0.6) \\ &\approx 0.931 \text{ bits per symbol} \end{aligned}$$

Constructing the Transition Probability Matrix

Given Probabilities:

- P(0—0) = 0.7
- P(1—0) = 0.3
- P(0—1) = 0.4
- P(1—1) = 0.6

Matrix Structure:

$$P = \begin{pmatrix} P_{00} & P_{01} \\ P_{10} & P_{11} \end{pmatrix}$$

Resulting Matrix:

$$P = \begin{pmatrix} 0.7 & 0.3 \\ 0.4 & 0.6 \end{pmatrix}$$

Properties:

- Each row sums to 1
- P_ij is probability of transitioning from state i to j
- Describes Markov chain behavior

Usage:

- Find stationary distribution
- Calculate entropy
- Analyze long-term behavior

Example Problem: Entropy and Information Rate

Consider a Markov source with two symbols A and B where $P(A|B) = 0.3$ and $P(B|A) = 0.4$. The initial symbol A is chosen with probability 0.5 and B with probability 0.5. Calculate the entropy $H(X)$ and information rate R.

Solution:

First, we calculate the stationary probabilities $\pi(A)$ and $\pi(B)$ for the Markov source. Let $\pi(A) = p$ and $\pi(B) = 1 - p$. Using the equilibrium condition:

$$\pi(A) = P(A|A)\pi(A) + P(A|B)\pi(B)$$
$$\pi(B) = P(B|B)\pi(B) + P(B|A)\pi(A)$$

Given $P(A|A) = 0.6$ and $P(B|B) = 0.7$:

$$p = 0.6p + 0.3(1 - p)$$
$$p = 0.6p + 0.3 - 0.3p$$
$$p - 0.3p = 0.3$$
$$0.7p = 0.3$$
$$p = \frac{3}{7}$$
$$\pi(A) = \frac{3}{7}, \quad \pi(B) = \frac{4}{7}$$

Now, we calculate the entropy $H(X)$:

$$H(X) = -[\pi(A)P(A|B)\log P(A|B) + \pi(B)P(B|A)\log P(B|A) + \pi(A)P(A|A)\log P(A|A) + \pi(B)P(B|B)\log P(B|B)]$$

Substituting the values, we get:

$$H(X) = -\left[\frac{4}{7} \times 0.3\log_2 0.3 + \frac{3}{7} \times 0.4\log_2 0.4 + \frac{3}{7} \times 0.6\log_2 0.6 + \frac{4}{7} \times 0.7\log_2 0.7\right]$$
$$\approx 0.9177 \text{ bits per symbol}$$

Next, we calculate the information rate R:

$$R = H(X)/T = 0.9177 \text{ bits per symbol}/1 \text{ second} = 0.9177 \text{ bits per second}$$

Information Measures of Continuous Random Variables

Entropy for Continuous Random Variables

- **Definition**:
  - Extension of Shannon's entropy to continuous probability distributions.
  - Defined using probability density functions (PDFs) instead of probability mass functions (PMFs).
- **Formula**:
  - Entropy $H(X)$ for a continuous random variable X with PDF $f(x)$ is

$$H(X) = -\int_{-\infty}^{\infty} f(x) \log f(x)\, dx$$

Mutual Information for Continuous Random Variables

- **Definition**:
  - Extension of mutual information to continuous random variables.
  - Similar formula as for discrete random variables, but integrals are used instead of summations.
- **Formula**:
  - Mutual information $I(X;Y)$ between continuous random variables X and Y with PDFs $f(x)$ and $g(y)$ is

$$I(X;Y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)\log\left(\frac{f(x,y)}{f(x)f(y)}\right) dx\, dy$$

Example Problem: Entropy for Continuous Random Variable

Example Problem:
Let X be a continuous random variable with PDF, Calculate the entropy $H(X)$ of X.

$$f(x) = \begin{cases} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{x^2}{2\sigma^2}}, & \text{for } -\infty < x < \infty, \\ 0, & \text{otherwise.} \end{cases}$$

Solution:

$$H(X) = -\int_{-\infty}^{\infty} f(x) \log f(x)\, dx$$

Substituting the given PDF, we have:

$$H(X) = -\int_{-\infty}^{\infty} \left(\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{x^2}{2\sigma^2}}\right) \log\left(\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{x^2}{2\sigma^2}}\right) dx$$

$$= -\int_{-\infty}^{\infty} \left(\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{x^2}{2\sigma^2}}\right) \left[\log\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) + \log\left(e^{-\frac{x^2}{2\sigma^2}}\right)\right] dx$$

$$= -\int_{-\infty}^{\infty} \left(\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{x^2}{2\sigma^2}}\right) \left[-\frac{1}{2}\log(2\pi\sigma^2) - \frac{x^2}{2\sigma^2}\right] dx$$

$$= \frac{1}{2}\log(2\pi\sigma^2) \int_{-\infty}^{\infty} \left(\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{x^2}{2\sigma^2}}\right) dx + \frac{1}{2\sigma^2} \int_{-\infty}^{\infty} x^2 \left(\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{x^2}{2\sigma^2}}\right) dx$$

$$= \frac{1}{2}\log(2\pi\sigma^2) \cdot 1 + \frac{1}{2\sigma^2} \cdot \sigma^2 = \frac{1}{2}\log(2\pi\sigma^2) + \frac{1}{2}$$

Therefore, the entropy $H(X) = \frac{1}{2}\log(2\pi\sigma^2) + \frac{1}{2}$.

Conditional Entropy for Continuous Random Variables

- **Definition**:
  - Conditional entropy measures the average uncertainty of a random variable given the value of another random variable.
  - It quantifies the remaining uncertainty about one variable after observing the other variable.
- **Formula**:
  - Conditional entropy $H(Y|X)$ for continuous random variables X and Y with joint PDF $f(x,y)$ and marginal PDF $f(x)$ is given by:

$$H(Y|X) = -\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)\log\left(\frac{f(x,y)}{f(x)}\right)\,dx\,dy$$

- **Interpretation**:
  - Conditional entropy measures the average uncertainty in Y given the value of X.
  - It is a measure of the remaining uncertainty in Y after X has been observed.

Example Problem: Mutual Information for Continuous Random Variables

Example Problem:
Let X and Y be continuous random variables with joint PDF:

$$f(x,y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} \exp\left(-\frac{1}{2(1-\rho^2)}\left[\frac{x^2}{\sigma_X^2} + \frac{y^2}{\sigma_Y^2} - \frac{2\rho xy}{\sigma_X\sigma_Y}\right]\right)$$

Calculate the mutual information $I(X;Y)$ between X and Y.
Solution:
The mutual information $I(X;Y)$ for continuous random variables X and Y with joint PDF $f(x,y)$, and marginal PDFs $f(x)$ and $g(y)$ is given by:

$$I(X;Y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)\log\left(\frac{f(x,y)}{f(x)f(y)}\right)\, dx\, dy$$

The solution involves numerical evaluation, and for jointly Gaussian distributions, the mutual information can be simplified:

$$I(X;Y) = -\frac{1}{2}\log(1-\rho^2)$$

where ρ is the correlation coefficient between X and Y.

Example Problem: Information Measures

Let X and Y be continuous random variables with joint PDF $f(x, y) = 2e^{-x-y}$ for $0 < x < \infty$ and $0 < y < \infty$. Calculate the mutual information $I(X; Y)$.

Solution:

The marginal PDFs are given by:

$$f_X(x) = \int_0^\infty f(x, y)\, dy = \int_0^\infty 2e^{-x-y}\, dy = 2e^{-x} \int_0^\infty e^{-y}\, dy = 2e^{-x}$$

$$f_Y(y) = \int_0^\infty f(x, y)\, dx = \int_0^\infty 2e^{-x-y}\, dx = 2e^{-y} \int_0^\infty e^{-x}\, dx = 2e^{-y}$$

Thus, the mutual information is:

$$\begin{aligned}
I(X; Y) &= \iint f(x, y) \log\left(\frac{f(x, y)}{f_X(x) f_Y(y)}\right) dx\, dy = \iint 2e^{-x-y} \log\left(\frac{2e^{-x-y}}{(2e^{-x})(2e^{-y})}\right) dx\, dy \\
&= \iint 2e^{-x-y} \log\left(\frac{2e^{-x-y}}{2e^{-x} \cdot 2e^{-y}}\right) dx\, dy = \iint 2e^{-x-y} \log\left(\frac{2e^{-x-y}}{4e^{-x-y}}\right) dx\, dy \\
&= \iint 2e^{-x-y} \log\left(\frac{1}{2}\right) dx\, dy = \iint 2e^{-x-y}(-1)\, dx\, dy = -2 \iint e^{-x-y}\, dx\, dy
\end{aligned}$$

We split the double integral into two single integrals:

$$-2 \int_0^\infty \int_0^\infty e^{-x-y}\, dx\, dy = -2 \left(\int_0^\infty e^{-x}\, dx\right) \left(\int_0^\infty e^{-y}\, dy\right)$$

$$\int_0^\infty e^{-x}\, dx = \left[-e^{-x}\right]_0^\infty = 1$$

$$\int_0^\infty e^{-y}\, dy = \left[-e^{-y}\right]_0^\infty = 1$$

$$I(X; Y) = -2 \cdot 1 \cdot 1 = -2$$

This calculation shows that mutual information $I(X; Y)$ is -2. However, mutual information should not be negative; hence, there's likely a need for reevaluation in the problem or setup. Mutual information is expected to be non-negative.

Relative Entropy (Kullback-Leibler Divergence)

- **Definition**:
  - Relative entropy, also known as Kullback-Leibler (KL) divergence, measures the difference between two probability distributions.
  - It quantifies the information lost when one distribution is used to approximate another.
- **Formula**:
  - Relative entropy $D_{\mathrm{KL}}(P\|Q)$ between probability distributions P and Q is given by:

$$D_{\mathrm{KL}}(P\|Q) = \sum_i P(i) \log\left(\frac{P(i)}{Q(i)}\right)$$

for discrete distributions, and:

$$D_{\mathrm{KL}}(P\|Q) = \int_{-\infty}^{\infty} p(x) \log\left(\frac{p(x)}{q(x)}\right)\, dx$$

for continuous distributions.

Properties of Relative Entropy

- **Non-negativity**:
  - Relative entropy $D_{\mathrm{KL}}(P\|Q)$ is always non-negative.
  - $D_{\mathrm{KL}}(P\|Q) \geq 0$ for all probability distributions P and Q.
  - The minimum value of $D_{\mathrm{KL}}(P\|Q)$ is 0, attained if and only if $P = Q$.
- **Asymmetry**:
  - Relative entropy is not symmetric: $D_{\mathrm{KL}}(P\|Q) \neq D_{\mathrm{KL}}(Q\|P)$ in general.
  - This property indicates that the divergence between P and Q may not be the same as the divergence between Q and P.

Example Problem: Relative Entropy for Discrete Distributions

Let P and Q be two discrete probability distributions defined as follows:

$$P = \left\{\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right\}, \quad Q = \left\{\frac{1}{2}, \frac{1}{4}, \frac{1}{4}\right\}$$

Calculate the relative entropy $D_{\mathrm{KL}}(P\|Q)$.

Solution:

The relative entropy $D_{\mathrm{KL}}(P\|Q)$ between two discrete distributions P and Q is given by:

$$D_{\mathrm{KL}}(P\|Q) = \sum_i P(i) \log_2 \left(\frac{P(i)}{Q(i)}\right)$$

Substituting the given distributions, we have:

$$D_{\mathrm{KL}}(P\|Q) = \frac{1}{3} \log_2 \left(\frac{1/3}{1/2}\right) + \frac{1}{3} \log_2 \left(\frac{1/3}{1/4}\right) + \frac{1}{3} \log_2 \left(\frac{1/3}{1/4}\right)$$

Calculating each term separately:

$$\frac{1}{3} \log_2 \left(\frac{1/3}{1/2}\right) = \frac{1}{3} \log_2 \left(\frac{2}{3}\right) \approx \frac{1}{3}(-0.58496) = -0.19499$$

$$\frac{1}{3} \log_2 \left(\frac{1/3}{1/4}\right) = \frac{1}{3} \log_2 \left(\frac{4}{3}\right) \approx \frac{1}{3}(0.41504) = 0.13835$$

$$\frac{1}{3} \log_2 \left(\frac{1/3}{1/4}\right) = \frac{1}{3} \log_2 \left(\frac{4}{3}\right) \approx \frac{1}{3}(0.41504) = 0.13835$$

Summing these values:

$$D_{\mathrm{KL}}(P\|Q) = -0.19499 + 0.13835 + 0.13835 \approx 0.08171$$

Therefore, the relative entropy $D_{\mathrm{KL}}(P\|Q) \approx 0.08171$.

Example: Relative Entropy-Continuous Distributions

Let $P(x)$ and $Q(x)$ be two continuous probability density functions (PDFs) defined as follows, Calculate the relative entropy

$$P(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}, \quad Q(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{x^2}{2\sigma^2}}$$

Solution:
The relative entropy $D_{\mathrm{KL}}(P\|Q)$ between two continuous distributions $P(x)$ and $Q(x)$ is given by:

$$D_{\mathrm{KL}}(P\|Q) = \int_{-\infty}^{\infty} P(x) \log\left(\frac{P(x)}{Q(x)}\right) dx$$

$$D_{\mathrm{KL}}(P\|Q) = \int_{-\infty}^{\infty} \left(\frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}\right) \log\left(\frac{\frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}}{\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{x^2}{2\sigma^2}}}\right) dx$$

Simplifying the logarithm term:

$$\frac{\frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}}{\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{x^2}{2\sigma^2}}} = \frac{1}{\sqrt{\sigma^2}} e^{\frac{x^2}{2\sigma^2} - \frac{x^2}{2}} = \frac{1}{\sigma} e^{\frac{x^2}{2}\left(\frac{1}{\sigma^2} - 1\right)}$$

So the integral becomes:

$$D_{\mathrm{KL}}(P\|Q) = \int_{-\infty}^{\infty} \left(\frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}\right) \left(\log\left(\frac{1}{\sigma}\right) + \log\left(e^{\frac{x^2}{2}\left(\frac{1}{\sigma^2} - 1\right)}\right)\right) dx$$

$$= \int_{-\infty}^{\infty} \left(\frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}\right) \left(-\log(\sigma) + \frac{x^2}{2}\left(\frac{1}{\sigma^2} - 1\right)\right) dx$$

Example: Relative Entropy-Continuous Distributions

Splitting the integral, we get:

$$D_{\mathrm{KL}}(P\|Q) = -\log(\sigma)\int_{-\infty}^{\infty}\left(\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}\right)dx + \frac{1}{2}\left(\frac{1}{\sigma^2}-1\right)\int_{-\infty}^{\infty}x^2\left(\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}\right)dx$$

The first integral evaluates to 1 because the integral of a PDF over its entire range is 1:

$$\int_{-\infty}^{\infty}\left(\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}\right)dx = 1$$

The second integral is a known result for the Gaussian distribution:

$$\int_{-\infty}^{\infty}x^2\left(\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}\right)dx = 1$$

Therefore:

$$D_{\mathrm{KL}}(P\|Q) = -\log(\sigma) + \frac{1}{2}\left(\frac{1}{\sigma^2}-1\right)$$

Simplifying further:

$$D_{\mathrm{KL}}(P\|Q) = -\log(\sigma) + \frac{1}{2\sigma^2} - \frac{1}{2}$$

Hence, the relative entropy is:

$$D_{\mathrm{KL}}(P\|Q) = \log(\sigma) + \frac{1}{2} - \frac{1}{2\sigma^2}$$

Information Theory in Communication Systems

- **Overview**:
  - Information theory plays a crucial role in the design and analysis of communication systems.
  - It provides tools to quantify and optimize the transmission of information over noisy channels.
- **Applications**:
  - Error-correcting codes: Information theory guides the design of efficient codes that can detect and correct errors during transmission.
  - Channel capacity: Information theory helps determine the maximum rate at which information can be reliably transmitted over a communication channel.
  - Source coding: It addresses the problem of efficiently representing information for transmission or storage.

Entropy Coding

- **Definition**:
  - Entropy coding is a technique used in information theory for lossless data compression.
  - It exploits the statistical properties of the source data to assign shorter codes to more frequent symbols and longer codes to less frequent symbols.
- **Example Algorithms**:
  - Huffman coding: A widely used entropy coding technique that constructs prefix codes based on the probability of symbols.
  - Arithmetic coding: An alternative to Huffman coding that represents a message as a single floating-point number in the interval [0, 1].
- **Performance**:
  - Entropy coding achieves compression efficiency close to the entropy of the source data.
  - It is widely used in data compression applications such as image and video compression, file compression, and communication systems.