Entropy, Relative Entropy and Mutual Information

Chapter 2 note taking of "*Elements of Information Theory*"

Entropy (Self-Information)

Entropy measures the uncertainty of a random variable, formulated as follows:

$$\begin{aligned} H(X) &= -\sum_{x \in \mathcal{X}} p(x) \log p(x) \\ &= \mathbb{E}_p \left[\frac{1}{\log p(X)} \right] \end{aligned}$$

where X is a discrete random variable, and $p(x) = p_X(x) = \Pr\{X = x\}$ is a probability mass function.

> **Remark 2.1:** Note that entropy is about the probability distribution, which does not depend on the actual values (such as vectors) taken by the random variable X.

Intuitively, the amount of information (entropy) is related to the probability of an event. For instance, if something happens with 100% certainty, it doesn't carry any useful information, and the entropy is 0.

Some lemmas related to the definition of entropy:

- **Lemma 1:** $H(X) \geq 0$
- **Lemma 2:** $H_b(X) = \log_b a \cdot H_a(X) = \log_b a \cdot \log_a p$

Joint Entropy and Conditional Entropy

Since $H(X)$ defines the entropy of a single random variable, it can be extended to a pair of random variables (X, Y), called ***Joint Entropy***, defined as follows:

$$\begin{aligned} H(X, Y) &= -\sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \log p(x, y) \\ &= \mathbb{E} \left[\frac{1}{\log p(X, Y)} \right] \end{aligned}$$

Moreover, I can define entropy conditioned on a second variable from the pair (X, Y), called ***Conditional Entropy***, as follows:

$$
\begin{aligned}
H(X \mid Y) &= \sum_{x \in \mathcal{X}} p(x) H(Y \mid X = x) \\
&= -\sum_{x \in \mathcal{X}} p(x) \sum_{y \in \mathcal{Y}} p(y \mid x) \log p(y \mid x) \\
&= -\sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \log p(y \mid x) \\
&= \mathbb{E}\left[\log \frac{1}{p(Y \mid X)}\right]
\end{aligned}
$$

Then, based on the definitions of joint and conditional entropy, I can prove the following theorem:

$$
H(X, Y) = H(X) + H(Y \mid X)
$$

Proofs:

$$
\begin{aligned}
H(X, Y) &= -\sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \log p(x, y) \\
&= -\sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \log p(x) p(y \mid x) \\
&= -\sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \log p(x) - \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \log p(y \mid x) \\
&= -\sum_{x \in \mathcal{X}} p(x) \log p(x) - \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \log p(y \mid x) \\
&= H(X) + H(Y \mid X)
\end{aligned}
$$

Relative Entropy and Mutual Information

Relative Entropy, denoted as $D(p \mid\mid q)$, measures the difference betIen two probability distributions. HoIver, it is **not** a true distance metric, since it is **not symmetric** and does **not satisfy the triangle inequality**. Therefore, it's better to think of $D(p \mid\mid q)$ as a measure of the "gap" or divergence betIen two distributions p and q.

This measure is known as the ***Kullback-Leibler Divergence (KL Divergence)***, and it is defined as:

$$
\begin{aligned}
D(p \mid\mid q) &= \sum_{x \in \mathcal{X}} p(x) \log \frac{p(x)}{q(x)} \\
&= \mathbb{E}_p \log \frac{p(X)}{q(X)}
\end{aligned}
$$

KL Divergence is always **non-negative (mostly positive)**, and it equals zero **if and only if** $p = q$.

> **Remark 2.2:** In the field of deep learning, the **distribution** p is often assumed to be the **true but unknown** distribution I aim to approximate, while **distribution** q is a **known and tractable** distribution, such as a Gaussian Distribution. This concept is widely used in generative models (e.g., VAE, diffusion models, flow matching), where minimizing KL divergence helps the model learn to resemble the true data distribution.

Then, I can measure the amount of information that one random variable contains about another random variable based on KL Divergence. I define this as ***Mutual Information***, denoted as $\boldsymbol{I(X;Y)}$. This is essentially the relative entropy between the joint distribution and the product of their marginal distributions, defined as follows:

$$\begin{aligned} I(X;Y) &= \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log\frac{p(x,y)}{p(x)p(y)} \\ &= D(p(x,y)\;||\;p(x)p(y)) \\ &= \mathbb{E}_{p(x,y)}\left[\log\frac{p(X,Y)}{p(X)p(Y)}\right] \end{aligned}$$

To understand this term intuitively, it's helpful to derive it in reverse. For instance, let's first consider the meaning of $D(p(x,y)\;||\;p(x)p(y))$. According to the definition of KL Divergence, $D(p(x,y)\;||\;p(x)p(y))$ represents *how far the joint distribution is from independence* between $p(x)$ and $p(y)$. If $p(x)$ and $p(y)$ are completely independent, then $p(x,y)=p(x)p(y)$, and we get $D(p(x,y)\;||\;p(x)p(y))=0$, leading to $I(X;Y)=0$. This means there is no shared information between X and Y.

Relationship Between Entropy and Mutual Information

It can be derived mutual information in terms of entropy as follows:

Proofs:

$$\begin{aligned} I(X;Y) &= \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log\frac{p(x,y)}{p(x)p(y)} \\ &= \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log\frac{p(x\mid y)\,\cancel{p(y)}}{p(x)\,\cancel{p(y)}} \\ &= \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log\frac{p(x\mid y)}{p(x)} \\ &= -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log p(x) + \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log p(x\mid y) \\ &= -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log p(x) - \left(-\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log p(x\mid y)\right) \\ &= \underbrace{-\sum_{x\in\mathcal{X}} p(x)\log p(x)}_{H(X)} - H(X\mid Y) \\ &= H(X) - H(X\mid Y)\square \end{aligned}$$

As previously explained, the meaning of $I(\cdot;\cdot)$ can also be intuitively understood through the concept of entropy.

For example, suppose a *transmitter* is trying to communicate a message, which we represent as a random variable X. The *receiver* observes a signal, represented as another random variable Y, that is influenced by X. In this context, the entropy $H(X)$ measures the total amount of uncertainty — or potential information — in the message being transmitted. Once the receiver observes Y, they can try to infer the original message X. If Y provides full information about X, the uncertainty in X after observing Y becomes minimal. If Y is noisy or incomplete, the uncertainty remains higher.

This remaining uncertainty is captured by the conditional entropy $H(X \mid Y)$. In other words, a larger $H(X \mid Y)$ indicates more uncertainty remains about the original message, meaning the receiver didn't recover it fully. A smaller $H(X \mid Y)$ implies the receiver was able to infer X more precisely. Therefore, mutual information $\boldsymbol{I(X;Y) = H(X) - H(X \mid Y)}$ quantifies the **reduction in uncertainty** about the sender's message (X) due to the receiver's observation (Y).

Back to the notation, the relationship between entropy and mutual information can be summarized as:

1. $I(X;Y) = H(X) - H(X \mid Y)$
2. $I(X;Y) = H(Y) - H(Y \mid X)$
3. $I(X;Y) = H(X) + H(Y) - H(X,Y)$
4. $I(X;Y) = I(Y;X)$
5. $I(X;X) = H(X)$

Chain Rules of Entropy, Relative Entropy, and Mutual Information

Joint entropy can be expressed as the sum of conditional entropies:

$$H(X_1, X_2, \ldots, X_n) = \sum_{n=1}^{N} H(X_n \mid X_{n-1}, \ldots, X_1)$$

Based on this property, we define ***conditional mutual information*** as:

$$\begin{aligned} I(X;Y \mid Z) &= H(X \mid Z) - H(X \mid Y, Z) \\ &= \mathbb{E}_{p(x,y,z)} \left[\log \frac{p(X,Y \mid Z)}{p(X \mid Z) p(Y \mid Z)} \right] \end{aligned}$$

This leads us to the chain rule of mutual information:

$$I(X_1, \ldots, X_n \mid Y) = \sum_{n=1}^{N} I(X_n; Y \mid X_{n-1}, \ldots, X_1)$$

> **Remark 2.3:** These definitions are grounded in the fundamental property of probability distributions, known as the ***chain rule***.

Additionally, we can define ***conditional KL divergence*** as:

$$D(p(y \mid x) \,||\, q(y \mid x)) = \sum_{x \in \mathcal{X}} p(x) \sum_{y \in \mathcal{Y}} p(y \mid x) \log \frac{p(y \mid x)}{q(y \mid x)}$$
$$= \mathbb{E}_{p(x,y)} \log \frac{p(Y \mid X)}{q(Y \mid X)}$$

Jensen's Inequality and Its Consequences

In information theory, properties of convex functions play a fundamental role. A function is defined as convex as follows:

$$\text{A function } f(x) \text{ is convex over an interval } (a, b) \text{ if, for every } x_1, x_2 \text{ and } 0 \leq \lambda \leq 1,$$
$$f(\lambda x_1 + (1 - \lambda)x_2) \leq \lambda f(x_1) + (1 - \lambda) f(x_2)$$

The function f is ***strictly convex*** if the equality holds only when $\lambda = 0$ or $\lambda = 1$, and $-f$ is then referred to as ***concave***.

To build intuition, consider a quadratic function: $f(x) = ax^2 + bx + c$. If the second derivative is non-negative, i.e., $a \geq 0$, then the function is convex (or strictly convex if $a > 0$). More generally, convexity can be analyzed using a Taylor series expansion around a point x_0:

Proof:

$$f(x) = f(x_0) + f'(x_0)(x - x_0) + \frac{f''(x^*)}{2}(x - x_0)^2$$

where x^* lies between x_0 and x. Assume that $f''(x^*) \geq 0$, and let $x_0 = \lambda x_1 + (1 - \lambda)x_2$. Setting $x = x_1$, we get:

$$\begin{aligned} f(x_1) &= f(x_0) + f'(x_0)(x_1 - \lambda x_1 + (1 - \lambda)x_2) + \frac{f''(x^*)}{2}(x_1 - x_0)^2 \\ &= f(x_0) + f'(x_0)((1 - \lambda)(x_1 - x_2)) + \frac{f''(x^*)}{2}(x_1 - x_0)^2 \\ &\geq f(x_0) + f'(x_0)((1 - \lambda)(x_1 - x_2)) \end{aligned}$$

Similarly, by taking $x = x_2$:

$$f(x_2) \geq f(x_0) + f'(x_0)(\lambda(x_2 - x_1))$$

Now, multiplying $f(x_1)$ by λ and $f(x_2)$ by $(1 - \lambda)$, then summing:

$$\begin{aligned} \lambda f(x_1) + (1 - \lambda) f(x_2) &\geq f(x_0) + \underbrace{\lambda f'(x_0)((1 - \lambda)(x_1 - x_2)) + (1 - \lambda) f'(x_0)(\lambda(x_2 - x_1))}_{= -\lambda(1-\lambda) f'(x_0)(x_2 - x_1) + \lambda(1-\lambda) f'(x_0)(x_2 - x_1) = 0} \\ &= f(\lambda x_1 + (1 - \lambda)x_2) \end{aligned}$$

This recovers the definition of a convex function, thereby completing the proof. □

Next, we introduce one of the most widely used inequalities across various fields such as machine learning, deep learning, and information theory — ***Jensen's Inequality***:

$$\mathbb{E}\left[f(X)\right] \geq f(\mathbb{E}(X))$$

where f is a convex function and X is a random variable.

Thanks to Jensen's Inequality, quantities such as KL divergence, mutual information, and their conditional forms between two probability distributions p and q are guaranteed to be non-negative (and often strictly positive). Equality holds if and only if $p = q$.

Log-Sum Inequality and Data-Processing Inequality

Now, we explore an important consequence of the concavity of the logarithm function, known as the **log-sum inequality**. This inequality plays a crucial role in proving that entropy is concave:

$$\sum_{i=1}^{n} a_i \log \frac{a_i}{b_i} \geq \left(\sum_{i=1}^{n} a_i\right) \log \frac{\sum_{i=1}^{n} a_i}{\sum_{i=1}^{n} b_i}$$

with equality if and only if $\frac{a_i}{b_i} = C_i$, a constant for all i.

Proof:

Let $a_i > 0$ and $b_i > 0$, and define the function $f(t) = t \log t$. This function is strictly convex because its second derivative is

$$f''(t) = \frac{1}{t} \log e > 0 \quad \text{for all } t > 0. \tag{1}$$

Applying Jensen's Inequality:

$$\sum_{i=1}^{n} \alpha_i f(t_i) \geq f\left(\sum_{i=1}^{n} \alpha_i t_i\right)$$

where $\alpha_i \geq 0$ and $\sum_i \alpha_i = 1$. Now, choose

$$\alpha_i = \frac{b_i}{\sum_j b_j}, \quad \text{and} \quad t_i = \frac{a_i}{b_i}. \tag{2}$$

Then the inequality becomes:

$$\sum_{i=1}^{n} \frac{a_i}{\sum_{j=1}^{n} b_j} \log \frac{a_i}{b_i} \geq \left(\sum_{i=1}^{n} \frac{a_i}{\sum_{j=1}^{n} b_j}\right) \log \left(\sum_{i=1}^{n} \frac{a_i}{\sum_{j=1}^{n} b_j}\right)$$

Multiplying both sides by $\sum_{j=1}^{n} b_j$, we recover the log-sum inequality. □

Next, The **Data-Processing Inequality (DPI)** is a fundamental result in information theory that formalizes the intuition that "no clever manipulation of data can increase information." Specifically, it states that if a random variable X influences Z only through an intermediate variable Y, then the mutual information between X and Z cannot exceed that between X and Y.

Formally, consider a Markov chain:

$$X \to Y \to Z$$

which means that X and Z are conditionally independent given Y, i.e.,

$$p(z \mid x, y) = p(z \mid y).$$

Then, the **Data-Processing Inequality** states with Markov chain $X \to Y \to Z$:

$$I(X; Z) \leq I(X; Y).$$

In other words, processing data (from $Y \to Z$) cannot increase the amount of information Z has about X.