2. Information Measure

Vaughan Sohn

October 6, 2024

Contents

Entropy

Mutual Information

KL-Divergence

Remarks about Informations measures

Convexity and Concavity of Information Measures

Data Processing Inequality and Fano's Inequality

Entropy

Entropy

Represent INFORMATION

어떤 사건 E가 발생했을 때, 그 사건이 매우 *희귀하다면* 우리에게 많은 정보를 제공해주겠지만, 매우 흔한 사건이라면 별 다른 정보를 제공해주지 않을 것이다.
$\Rightarrow$ 이러한 직관에 기반하여, entropy를 다음과 같이 정의해보자.

Definition 1 (entropy on event)

For the event E, we define a measure of information, **entropy** $H(E) \in \mathbb{R}^+$ that satisfies the following properties:

- Function of $P(E)$
- Continuous in $P(E)$
- If $P(E)$ is increasing, then entropy $H(E)$ is decreasing
- If $E_1 \perp E_2$, then joint entropy is just addition of each entropy

$$H(E_1 \cap E_2) = H(E_1) + H(E_2)$$

Therefore, the entropy can be defined by the following function,

$$H(E) \triangleq -\log P(E)(\text{bits}).$$

Entropy

일반화하면, 특정한 event E가 아니라 random variable; experiment에 대한 entropy도 다음과 같이 정의할 수 있다.
$\Rightarrow$ Average amount of information by observing the realization of X. (i.e., $x \in \mathcal{X}$)

Definition 2 (entropy on random variable)

The entropy $H(X)$ of a discrete random variable X is defined by

$$H(X) = -\sum_{x \in \mathcal{X}} P_X(x) \log P_X(x)$$

By definition, entropy has following properties:

- $H(P) \geq 0$, with equality iff P is deterministic.
- $H(P)$ is continuous in $P \in \mathbb{R}^{|\mathcal{X}|}$
- H is *divisible* with successive choices
  Example:
  $$H([1/2, 1/3, 1/6]) = H([1/2, 1/2]) + \frac{1}{2} H([2/3, 1/3])$$

Examples

Examples:

- Binary random variable $X \sim$ Bernoulli(p)인 r.v.에 대해 entropy를 구하라.

  $$H_B(p) \triangleq$$

- Random variable uniformly distributed over a finite set r.v. U에 대한 sample space 가 $\mathcal{U} = \{1, 2, \cdots, M\}$이고 uniform distribution을 따를때, entropy를 구하라.

  $$H(U) \triangleq$$

Multivariable entropy

Definition 3 (joint entropy)

The **joint entropy** $H(X,Y)$ of a pair of discrete random variables (X,Y) with a joint distribution $P_{X,Y}(x,y)$ is defined as

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

Definition 4 (conditional entropy on observable)

The **conditional entropy** of Y, conditioned on $X = x$ is defined as

$$H(Y|X=x) = -\sum_{y\in\mathcal{Y}} P_{Y|X}(y|x)\log P_{Y|X}(y|x).$$

Definition 5 (conditional entropy on r.v.)

The **conditional entropy** of Y, conditioned on X is defined as

$$H(Y|X) = \mathbb{E}_{P_X}[H(Y|X=x)] = \mathbb{E}_{P_{XY}}[-\log P_{Y|X}(Y|X)]$$

✓meaning: Random variable X에 대한 entropy $H(X)$가 가능한 outcome $x \in \mathcal{X}$의 entropy $H(X=x)$의 expectation으로 정의되는 것처럼, random variable X자체에 대한 conditioned entropy $H(Y|X)$는 X가 가질 수 있는 모든 outcome $x \in \mathcal{X}$에 대한 conditioned entropy$H(Y|X=x)$의 expectation으로 정의할 수 있다.

∗ Proof: Joint probability에 대한 표현으로 전환하는 과정을 기술하면 다음과 같다.
$\Rightarrow$

Multivariable Entropy

Theorem 6 (chain rule)

$$\begin{aligned} H(X,Y) &= H(X) + H(Y|X) \\ &= H(Y) + H(X|Y) \end{aligned}$$

$*$ Proof:

$\Rightarrow$

Examples

Example: 다음의 joint probability가 주어졌을 때, 각각의 entropy를 구하라.

	X = 0	X = 1
Y = 0	1/2	1/3
Y = 1	0	1/6

- $H(X), H(Y)$
- $H(X, Y)$
- $H(Y|x=0),\ H(Y|x=1)$
- $H(Y|X)$

$\Rightarrow$

Mutual Information

Mutual Information

Definition 7 (mutual information)

The **mutual information** $I(X;Y)$ is defined as

$$I(X;Y) \triangleq H(X) - H(X|Y)$$

✓meaning: $H(X)$가 X에 대한 정보[$*$], $H(X|Y)$가 Y를 알았을 때 X에 대해 남아있는 정보이므로 $I(X;Y)$는 Y를 알게됨으로서 얻은 X에 대한 정보로 해석할 수 있다.

- Dependency on channel $W = P_{Y|X}$,
  - 채널이 완전하다면 $I(X;Y) = H(X) \leftrightarrow H(X|Y) = 0$
  - 채널이 불완전하여 손실되는 정보가 있다면 $I(X;Y) = 0 \leftrightarrow H(X|Y) = H(X)$
- By definition, mutual information has following properties:
  - independence:

$$X \perp Y \ \rightarrow \ I(X;Y) = 0$$

  - symmetry relation: $(*)$

$$I(X;Y) = H(Y) - H(Y|X) = H(X) - H(X|Y)$$

Conditional mutual information

Definition 8 (conditional mutual information)

The **conditional mutual information** $I(X;Y|Z)$ is defined as

$$\begin{aligned} I(X;Y|Z) &\triangleq H(X|Z) - H(X|Y,Z) \\ &= H(Y|Z) - H(Y|X,Z) \end{aligned}$$

- Conditional mutual information을 정의하기 위해 conditioned r.v. Z의 모든 realization에 대한 expectation을 취하여 계산할 수 있다.

$$I(X;Y|Z) = \mathbb{E}_Z[I(X;Y|Z=z)]$$

- 예를 들어, conditioned r.v. Z는 channel을 이용한 transmission에서 어떤 channel을 사용할 지 결정하는 요소가 될 수 있다.

Chain rule

Theorem 9 (chain rule)

$$I(\underbrace{X_1, X_2}_{joint\ r.v.}; Y) = I\left(X_1; Y\right) + I\left(X_2; Y | X_1\right)$$
$$= I(X_2; Y) + I(X_1; Y | X_2)$$

* Proof:

$\Rightarrow$

KL-Divergence

KL-Divergence

Definition 10 (KL-divergence)

The *relative entropy* or **Kullback-Leibler distance** between two PMF $P_X(x)$ and $Q_X(x)$ is defined as

$$D(P||Q) \triangleq \sum_{x \in \mathcal{X}} P_X(x) \log \frac{P_X(x)}{Q_X(x)} = \mathbb{E}_{P_X}\left[\log \frac{P(X)}{Q(X)}\right]$$

- KL divergence는 동일한 sample space에 대한 서로 다른 확률분포간의 차이; *distance*를 정량화한다.
- $||$의 오른쪽에 위치한 PMF Q는 분모에 위치하기 때문에, KL divergence가 잘 정의되기 위해서는 다음 조건을 만족해야한다.
    - P is dominated by Q ($P << Q$) [$*$]
    - If $Q(X) = 0$, then $P(X) = 0$ (for all $x \in \mathcal{X}$).

KL-Divergence

By definition, KL-Divergence has following properties:

- KL divergence is *not symmetric*

$$D(P||Q) \neq D(Q||P)$$

- when $P = Q$, then KL divergence is $D(P||Q) = 0$
- **information inequality**

Theorem 11 (information inequality)

KL divergence is non-negative

$$D(P||Q) \geq 0 \textit{ with equality iff } P = Q.$$

$*$ Proof:

Remarks about Informations measures

Useful facts for Information Measures

Theorem 12

$H(X) \leq H(U)$, where U is the uniform distribution over $\mathcal{X}$ (equality iff U is uniformly distributed)

$*$ Proof:

Useful facts for Information Measures

Theorem 13

$H(Y) \geq H(Y|X)$, *with equality iff* $X \perp Y$.

Theorem 14

$I(X;Y) \geq 0$, *with equality iff* $X \perp Y$.

$*$ Proof:

Another definition of mutual information

KL-divergence를 사용하면 mutual information을 다른 관점으로 해석할 수 있다.

Definition 15 (mutual informtion)

Consider two random variables X and Y with a joint PMF $P_{X,Y}(x,y)$ and marginal PMF $P_X(x)$ and $P_Y(y)$. The **mutual information** $I(X;Y)$ is the relative entropy *between the joint distribution* $P_{X,Y}(x,y)$ *and the product distribution* $P_X(x) \cdot P_Y(y)$.

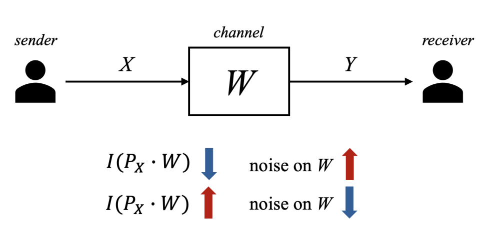
$$I(X;Y) \triangleq D(P_{X,Y}(x,y) || P_X(x)P_Y(y)) = \sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} P_{X,Y}(x,y) \log \frac{P_{X,Y}(x,y)}{P_X(x)P_Y(y)}$$

$*$ Proof: 첫 번째 정의와 두 번째 정의가 동일함을 보이자.

$\Rightarrow$

Application of Information Measures

- Entropy: Data compression (e.g, Huffman code)
- Mutual information: Data transmission



- KL divergence: Hypothesis testing [$*$]
  - 어떤 sample space에 대해서 서로 다른 두 hypothesis를 가정해 볼 수 있다.
    H_0: $X \sim P, H_1 : X \sim Q$
  - X를 여러번 관측하여 얻은 realization value를 사용하여얻은 empirical distribution 으로부터 실제 probability distribution을 유추할 수 있기 때문에, 어떤 hypothesis가 진실인지 추정할 수 있다.
  - Hypothesis testing에서 추정이 틀렸을 확률은 다음과 같이 정의된다.

$$P(*) \approx \exp[-nD(P||Q)]$$

  - 따라서 두 PMF의 KL divergence를 알고있다면 얼마나 많은 데이터($= n$)가 있어야 우리가 원하는 수준의 error-rate를 달성할 수 있는지를 계산할 수 있다.

Convexity and Concavity of Information Measures

Convexity and Concavity

Definition 16 (convexity)

Definition 17 (concavity)

-

Prerequisites: Log-sum inequality

Lemma 18 (log-sum inequality)

* Proof:

Theorem 19

-

✓ meaning:

Convexity of KL-divergence

* Proof:

Concavity of Entropy

Corollary 20

✓ meaning:

∗ Proof:

Concavity and Concavity of Mutual Information

Corollary 21

✓ meaning:

∗ Proof:

Concavity and Concavity of Mutual Information

Corollary 22

✓ meaning:

∗ Proof:

Data Processing Inequality and Fano's Inequality

Definition 23 (Markov chain)

-

Data Processing Inequality

Theorem 24 (data processing inequality)

✓ meaning:
∗ Proof:

Data Processing Inequality

Corollary 25 (Data Processing Inequality on entropy)

Corollary 26 (Data Processing Inequality on KL-divergence)

Fano's Inequality

System:

Theorem 27 (Fano's inequality)

✓ meaning:

Fano's Inequality

* Proof:

Appendix

Notations

- entropy of r.v. $X \sim P_X$: $H(X), H(P_X)$
- conditional entropy of Y conditioned on $X = x$: $H(Y|X = x), H(P_{Y|X}(\cdot|x))$
- conditional entropy of Y conditioned on X: $H(Y|X), H(P_{Y|X})$
- mutual information of X, Y: $I(X;Y), I(P_X \cdot P_{Y|X}), I(P_X, P_{Y|X})$
  주로 $W = P_{Y|X}$로 두고 $I(P_X \cdot W)$와 같이 표기하여 사용한다.
- KL divergence between two distribution P and Q: $D(P||Q), D(P_P||P_Q)$

References

- T. M. Cover and J. A. Thomas. Elements of Information Theory, Wiley, 2nd ed., 2006.
- Lecture notes for EE623: Information Theory (Fall 2024)