2. Information Measure

Vaughan Sohn

October 6, 2024

Contents

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

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

Multivariable Entropy

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

Definition 7 (mutual information)

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

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

✓<u>meaning:</u> $H(X)$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:

Chain rule

Theorem 12 (chain rule)

$$D(P_{X,Y}||Q_{X,Y}) = D(P_X||Q_X) + D(P_{Y|X}||Q_{Y|X})$$

$*$ Proof:

$\Rightarrow$

Remarks about Informations measures

Another definition of mutual information

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

Definition 13 (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.

$\Rightarrow$

Useful facts for Information Measures

Fact

Here are some easy facts:

$$H(P_X) = H(P_U) - D(P_X||P_U) \quad (*)$$
$$I(P_X \cdot W) = D(P_{XY}||(P_X \cdot P_Y))$$

$*$ Proof:
$\Rightarrow D(P_X||P_U)$를 전개하자.

Useful facts for Information Measures

Corollary 14

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

Corollary 15 (onditioning reduces entropy)

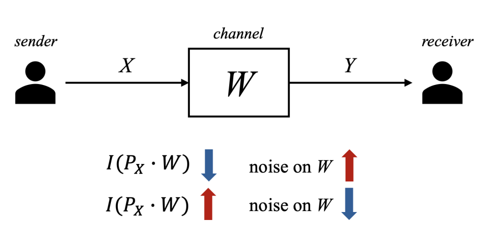
$H(Y) \geq H(Y|X)$, with equality iff $X \perp Y$.

Corollary 16

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

$*$ Proof: By the facts, we can easily prove that

- 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 17 (convexity)

A function $f(x)$ is said to be **convex** over an interval (a, b) if for every $x_1, x_2 \in (a, b)$ and $0 \leq \theta \leq 1$,

$$f((1-\theta)x_1 + \theta x_2) \leq (1-\theta)f(x_1) + \theta f(x_2)$$

A function f is said to be **strictly convex** if equality holds only if $\theta = 0$ or $\theta = 1$.

Definition 18 (concavity)

A function $f(x)$ is said to be **concave** over an interval (a, b) if for every $x_1, x_2 \in (a, b)$ and $0 \leq \theta \leq 1$,

$$f((1-\theta)x_1 + \theta x_2) \geq (1-\theta)f(x_1) + \theta f(x_2)$$

A function f is said to be **strictly convex** if equality holds only if $\theta = 0$ or $\theta = 1$.

- A function f is concave if $-f$ is convex.
- convex, concave는 함수가 function value의 linear combination과 비교하여 어떻게 동작하는지를 판단한다.
- 어떤 함수가 convex라면, local minimum이 global minimum과 동일하기 때문에 최적화를 쉽게 할 수 있다.

Prerequisites: Log-sum inequality

Lemma 19 (log-sum inequality)

Given $a_i, b_i \geq 0, i = 1, 2, \cdots, n$, *let* $a = \sum_{i=1}^{n} a_i$ *and* $b = \sum_{i=1}^{n} b_i$. *Then we have*

$$\sum_i a_i \log \frac{a_i}{b_i} \geq a \log \frac{a}{b},$$

with equality iff $\frac{a_i}{a} = \frac{b_i}{b}, \ \forall i$.

$*$ Proof: Consider $k_a, k_b > 0$, s.t. $k_a \cdot a = 1, \ k_b \cdot b = 1$

$\Rightarrow$

Convexity of KL-divergence

Theorem 20

$D(P||Q)$ is **convex** *in the pair (P, Q); That is if (P_1, Q_1) and (P_2, Q_2) are any two pairs of PMF, for all $0 \leq \theta \leq 1$ then*

$$D(P_\theta||Q_\theta) \leq (1-\theta)D(P_1||Q_1) + \theta D(P_2||Q_2)$$

where $P_\theta = (1-\theta)P_1 + \theta P_2$ and $Q_\theta = (1-\theta)Q_1 + \theta Q_2$.

✓meaning:기존의 PMF를 θ라는 동일한 파라미터에 의해 mixing된 PMF는 각 PMF P, Q가 가지고 있던 특징이 줄어들어서 두 분포가 유사하게 변화한다. ($D \Downarrow$)

$*$ Proof: By log-sum inequality,

$\Rightarrow$

Concavity of Entropy

Corollary 21

$H(P)$ is a concave function of P

$$H(P_\theta) \geq (1-\theta)H(P_1) + \theta H(P_2)$$

✓meaning: PMF를 mixing 할수록 점점 uniform distribution에 가깝게 변한다. ($H \Uparrow$)

$*$ Proof:

- (pf.1) D가 convex이므로 다음의 관계를 이용하여 쉽게 증명할 수 있다.

$$H(P) = H(U) - D(P||U)$$

- (pf.2) 또는 transmission system에서 source data X의 distribution P_1, P_2을 결정하는 랜덤 스위치 $Z \sim B(\theta)$를 생각해볼 수 있다.
- 따라서 source X는 $B(p)$에 따라 mixed distribution P_θ을 따른다.
- 이때, X의 entropy $H(P_\theta)$는 랜덤 스위치 Z의 값을 관측했을 때의 entropy $H(P_\theta|Z)$보다 당연히 크다는 것을 직관적으로 이해할 수 있다.
  $\Rightarrow$

Concavity and Concavity of Mutual Information

Corollary 22

The mutual information $I(P_X \cdot W)$ *is a* **concave** *function of* $P_X(x)$ *for fixed* W

$$I(P_\theta \cdot W) \geq (1-\theta)I(P_1 \cdot W) + \theta I(P_2 \cdot W)$$

✓meaning: Source data의 distribution을 mixing 할 수록 channel을 통과한 뒤 얻을 수 있는 정보의 양이 더 많아진다.

∗ Proof: 다음의 system을 가정하자.

- one channel W, two source distribution P_1, P_2 depended on $Z \sim B(\theta)$.

$$X \sim P_\theta = (1-\theta)P_1 + \theta P_2$$

- (hint: chain rule)
  $\Rightarrow$

Concavity and Concavity of Mutual Information

Corollary 23

The mutual information $I(P_X \cdot W)$ is a **convex** *function of W for fixed $P_X(x)$*

$$I(P_X \cdot W_\theta) \leq (1-\theta)I(P_X \cdot W_1) + \theta I(P_X \cdot W_2)$$

✓meaning: Channal의 distribution을 mixing 할 수록 channel을 통과한 뒤 얻을 수 있는 정보의 양이 더 줄어든다.

∗ Proof: 다음의 system을 가정하자.

- one source distribution P_X, two channel W_1, W_2 depended on $Z \sim B(\theta)$.

$$W_\theta = (1-\theta)W_1 + \theta W_2$$

- (hint: chain rule)
  $\Rightarrow$

Conditioning effect on I

Entropy는 conditioning을 취하면 항상 그 값이 작아지지만 mutual information은 여기서 보인 것처럼 Z를 어떻게 설정하는지에 따라 더 커질수도 있고 더 작아질수도 있다.

Data Processing Inequality and Fano's Inequality

Prerequisites: Markov chain

Definition 24 (Markov chain)

Random variablesX, Y, Z are said to form a **Markov chain** in that order (denoted by $X \to Y \to Z$) if the conditional distribution of Z depends *only* on Y and is conditionally independent of X. $(*)$

$$P_{XYZ}(x, y, z) = P_X(x)P_{Y|X}(y|x)P_{Z|Y}(z|y)$$

- Markov chain은 서로 직접적으로 영향받는 변수가 정해져있기 때문에, 아무런 영향을 주지 않는 r.v.에 대한 conditioning은 무시할 수 있다.

$$P_{Z|YX} = P_{Z|Y}$$

- $X \to Y \to$ is equivalent to $X \to Y \to X$
  - $*$ Proof:

Data Processing Inequality

Theorem 25 (data processing inequality)

If $X \to Y \to Z$*, then*

$$I(X;Y) \geq I(X;Z)$$

with equality iff $I(X;Y|Z) = 0$*, or equivalently*

- $X \to Z \to Y$.
- *given* z, $X \perp Y$
- $I(X;Y) = I(X;Z)$

✓meaning: (data processing) 원본 데이터 X에 대해 processing을 수행하면 할 수록 점차 원본과의 관계; mutual information은 줄어든다. ($\leftrightarrow$ 증가할 수 없다.)

$*$ Proof: (By chain rule,)
$\Rightarrow$

Data Processing Inequality

Corollary 26 (Data Processing Inequality on entropy)

If $X \to Y \to Z$, *then*

$$H(X|Y) \leq H(X|Z)$$

with equality iff $I(X;Y|Z) = 0$

Corollary 27 (Data Processing Inequality on KL-divergence)

P_X *and* Q_X *is two input distributions and* P_Y *and* Q_Y *is two corresponding output distributions, then*

$$D(P_Y||Q_Y) \leq D(P_X||Q_X)$$

where $P_Y(y) = \sum_x P_X(x)W(y|x)$, $Q_Y(y) = \sum_x Q_X(x)W(y|x)$

✓meaning: 서로 다른 분포를 따르는 source였어도, 동일한 channel을 통과하면 data 간의 차이가 줄어든다.

∗ Proof:

Fano's Inequality

System: 우리가 구하고 싶은 원본 데이터 X 대신, 실제로 관측가능한 데이터 Y를 이용하여 X의 값을 추정하는 상황을 가정해보자. ($X \to Y \to \hat{X}(Y)$)

- X: (hide) data, Y: observable data
- $\hat{X}(Y)$: estimate function
- indicator:

$$\mathbf{1}_E := \begin{cases} 0, & \hat{X}(Y) \neq X \\ 1, & \hat{X}(Y) = X \end{cases}$$

- error probabilistic:

$$P_e := P(\mathbf{1}_E = 0)$$

Theorem 28 (Fano's inequality)

For any estimator $\hat{X}$, we have error probabilistic's lower bound

$$H(X|Y) \leq \log 2 + P_e \cdot \log(|\mathcal{X}| - 1) \Leftrightarrow \frac{H(X|Y) - \log 2}{\log(|\mathcal{X}| - 1)} \leq P_e$$

✓meaning: $H(X|Y)$의 값이 충분히 작아져야 error probabilistic도 작아질 수 있다. 따라서 $H(X|Y)$의 값을 측정하여 원하는 error rate를 얻기 위해서 얼마나 많은 양의 sample data Y^n가 필요할 지 유추할 수 있다.

Fano's Inequality

∗ <u>Proof</u>: (using chain rule, uniform bound and *data processing inequality*)

$H(X, \mathbf{1}_E | \hat{X}(Y)) =$

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

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)