2강 : Entropy, Relative, Mutual

✳ Entropy

— $H(X) = -\sum_{x \in X} p(x) \log p(x) = E[-\log p(x)]$

- 특정 함수 g에 대해 $g(X)$도 R.V.

1) Non-negativity : $H(X) \geq 0$

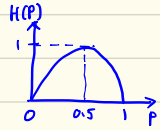
pf) $-\log p(x) > 0$

2) $H_b(X) = \log_b a \; H_a(X)$

pf) $\log_b p(x) = \log_a p(x) \cdot \log_b a$

— Example : Bernoulli RV's entropy

- $X = \begin{cases} 0 & \text{확률 } p \\ 1 & \text{〃 } 1-p \end{cases}$
- $H(X) = -p \log p - (1-p)\log(1-p) = H(p)$



binary entropy function

✳ Joint Entropy

— R.V. X, Y에 대한 Joint Entropy

- $H(X,Y) = -E[\log p(X,Y)] = -\sum_x \sum_y p(x,y) \log p(x,y)$
- 의미 : 여러 개의 R.V.가 모여, 하나의 RV가 되는 것.

— Marginal Distribution

- $\sum_{y \in Y} p(x,y) = p(x)$
- ex)

x \ y	0	1	p(x)
0	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{3}{4}$
1	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{4}$

marginal 꼴.

✳ Conditional Entropy

— $(X,Y) \sim p(x,y)$ 일 때 $-H(Y|X=x)$

- conditional probability : $p(y|x) = \frac{p(x,y)}{p(x)}$

- $H(Y|X) = -\sum_x p(x) \sum_y p(y|x) \log p(y|x)$

$= -\sum_x \sum_y p(x,y) \log p(y|x) = -E_{x,y}[\log p(y|x)]$

ex)

Y \ X	1	2
1	$\frac{1}{2}$	$\frac{1}{4}$
2	$\frac{1}{8}$	$\frac{1}{8}$

, $H(Y|X)$ 은

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

① $H(Y|X) = -\sum_{x,y} p(x,y) \log p(y|x)$

$p(y=1|x=1) = \frac{\frac{1}{2}}{\frac{1}{2}+\frac{1}{8}} = \frac{4}{5}$, $p(y=1|x=2) = \frac{\frac{1}{4}}{\frac{1}{4}+\frac{1}{8}} = \frac{2}{3}$

$p(y=2|x=1) = \frac{\frac{1}{8}}{\frac{1}{2}+\frac{1}{8}} = \frac{1}{5}$ $p(y=2|x=2) = \frac{1}{3}$

$H(Y|X) = -\left(\frac{1}{2}\log\frac{4}{5} + \frac{1}{8}\log\frac{1}{5} + \frac{1}{4}\log\frac{2}{3} + \frac{1}{8}\log\frac{1}{3}\right)$

② $H(Y|X) = \sum_x p(x) H(Y|X)$

$= \left(p(x=1) H(Y|x=1) + p(x=2) H(Y|x=2)\right)$

$= \left(\frac{5}{8} H\left(\frac{4}{5}\right) + \frac{3}{8} H\left(\frac{1}{3}\right)\right)$

$H(Y|X) = \sum_x p(x) H(Y|X=x)$

* Chain Rule

$H(X,Y) = H(X) + H(Y|X)$ log 이므로 덧셈

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

$= \log p(x) \cdot p(y|x)$

$= \log p(x) + \log p(y|x) = H(X) + H(Y|X)$

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

$= H(Y) + H(X|Y)$ $\Rightarrow H(Y|X) \neq H(X|Y)$

✳ Relative Entropy = KL divergence

- 두 probability distribution의 distance 측정.
- RV의 분포를 p라고 가정했을 때, length = $H(p)$
  그러나 실제 분포를 우리가 모르니까, q라고 가정해버리면, 필요 bit = $H(q) = H(p) + D(p\|q)$ (penalty)
- KL Divergence
  - $D(p\|q) = \sum_x p(x) \log \frac{p(x)}{q(x)} = E_p\left[\log \frac{p(x)}{q(x)}\right]$
  - if $p=q$, $D(p\|q)=0$
  - $D(p\|q) \neq D(q\|p)$; Asymmetric

✳ Mutual Information = $I(X;Y)$

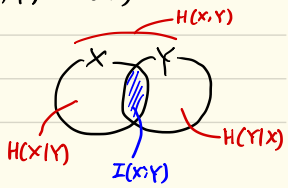
- 2개의 RV X, Y에 대해 Joint prob $p(x,y)$와 marginal prob $p(x), p(y)$를 가질때,
  - $I(X;Y) = D\left(p(x,y) \| p(x)p(y)\right)$
    $= \sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)}$
- $I(X;Y) = \sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)}$
  $= \log \frac{p(x)p(y|x)}{p(x)} = \log \frac{p(y|x)}{p(y)} = \log p(y|x) - \log p(y)$

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

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

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

$$= H(X) + H(Y) - H(X,Y) \Rightarrow$$



✳ Chain rule for entropy

— 여러 개의 RV $X_1, \cdots, X_n \sim p(x_1, \cdots, x_n)$ 에 대해

(∵ $H(X) = H(X|Y) + I(X;Y)$)

- $H(X_1, \cdots, X_n) = \sum_{i=1}^{n} H(X_i | X_1, \cdots, X_{i-1}) \leq H(X_1) + H(X_2) + \cdots + H(X_n)$

  확률은 곱셈. 엔트로피는 덧셈. ← 등호는 independent.

-

✳ Conditional Mutual Information

— Z가 주어졌을 때, X와 Y의 conditional MI.

- $I(X;Y|Z) = H(X|Z) - H(X|Y,Z)$ ← $I(X;Y)$보다 클수도, 작을수도 있음.

— Chain rule of MI

- $I(X_1, X_2, \cdots, X_n ; Y) = \sum_{i=1}^{n} I(X_i ; Y | X_1, \cdots, X_{i-1})$

  X_i

✳ Conditional Relative Entropy

— $p(y|x)$ 와 $q(y|x)$ 사이 relative entropy.

- $D(p(y|x) \,\|\, q(y|x)) = \sum_x E_x [D(p(y|X=x) \,\|\, q(y|X=x))]$

  $= \sum_x \sum_y p(x) p(y|x) \log \frac{p(y|x)}{q(y|x)}$

  $= \sum_{x,y} p(x,y) \log \frac{p(y|x)}{q(y|x)} = E_{x,y}\left(\log \frac{p(y|x)}{q(y|x)}\right)$

  똑같음.

—

$D(p(x,y) \,\|\, q(x,y)) = D(p(x) \,\|\, q(x)) + D(p(y|x) \,\|\, q(y|x))$

pf) $\frac{p(x,y)}{q(x,y)} = \frac{p(x)p(y|x)}{q(x)q(y|x)} = \frac{p(y|x)}{q(y|x)} \cdot \frac{p(x)}{q(x)}$

↑ log → 덧셈.

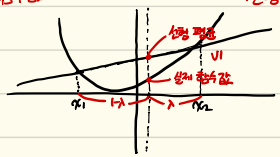
아래 볼록

✳ Convex functions

— 모든 $x_1, x_2 \in (a,b)$ 와 $0 \le \lambda \le 1$ 에 대해, 아래 만족시 $f(x)$는 (a,b)에서 convex function 이다.

• $f(\lambda x_1 + (1-\lambda)x_2) \le \lambda f(x_1) + (1-\lambda) f(x_2)$ --- Theorem 2.6.1.

함수값 / 선형 평균



• $\lambda = 0$ or 1 일 때 equal.

위로 볼록

— $f(x)$가 convex면 $-f(x)$는 concave.

— $f(x)$의 second derivative가 positive면, $f(x)$는 convex.

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

→ equality를 만족하는 x^*가 존재.

$\sim\sim\sim\sim\sim \ge 0$

$x_0 = \lambda x_1 + (1-\lambda)x_2$ 라 하면

1) $x = x_1$ 이면 $f(x_1) \ge f(x_0) + f'(x_0)[(1-\lambda)(x_1 - x_2)]$ --- ①

2) $x = x_2$ 이면 $f(x_2) \ge f(x_0) + f'(x_0)[\lambda(x_2 - x_1)]$ --- ②

① × λ + ② × $(1-\lambda)$: $\lambda f(x_1) + (1-\lambda) f(x_2) \ge f(x_0) = f(\lambda x_1 + (1-\lambda)x_2)$ □

✳ Jensen's Inequality

— $f(x)$가 convex function, X가 RV 면

• $E[f(X)] \ge f[E(X)]$

위의 식에서 λ, $1-\lambda$가 RV의 probability 면 됨.

pf) X의 값: x_1, x_2 , 확률 : p_1, p_2 라 하자.

2.6.1

1. $p_1 f(x_1) + p_2 f(x_2) \ge f(p_1 x_1 + p_2 x_2)$, $p_1 + p_2 = 1$ 인가? $p_1 = \lambda$면 2.6.1에 의해 O.

2. $\sum_{i=1}^{k} p_i f(x_i) \ge f\left(\sum_{i=1}^{k} p_i x_i\right)$, $\sum_{i=1}^{k} p_i = 1$ 이라 하자.

$$\sum_{i=1}^{k} p_i f(x_i) = p_k f(x_k) + (1-p_k)\sum_{i=1}^{k-1} \frac{p_i}{1-p_k} f(x_i) \quad \left(\because \sum_{i=1}^{k-1} p_i = 1-p_k\right)$$

$$= p_k f(x_k) + (1-p_k)\sum_{i=1}^{k-1} p_i' f(x_i) \quad \left(p_i' = \frac{p_i}{1-p_k},\ \sum_{i=1}^{k-1} p_i' = 1\right)$$

$$\ge p_k f(x_k) + (1-p_k) f\left(\sum p_i' x_i\right)$$

$$\ge f\left(p_k x_k + (1-p_k)\sum_{i=1}^{k-1} p_i' x_i\right) = f\left(\sum_{i=1}^{k} p_i x_i\right)$$ □

✳ Information Inequality

- 두 Probability distribution $p(x), q(x)$에 대해

  • $D(p\|q) \geq 0$. 등호는 $p(x)=q(x)$에서 성립.

pf) Using Jensen's Inequality

- log ; concave function

-

$$-D(p\|q) = -\sum_x p(x) \log \frac{p(x)}{q(x)}$$

$$= \sum_x p(x) \log \frac{q(x)}{p(x)}$$

$$\leq \log \sum_x p(x) \cdot \frac{q(x)}{p(x)} = \log \sum q(x) = \log 1 = 0 \text{ .}$$

(∵ log 가 concave function. $\sum p(x)=1$)

- Non-negativity of MI Corollary (특별 case) 증명

  • $I(X;Y) = D(p(x,y) \| p(x)p(y)) \geq 0$, 등호는 $p(x), p(y)$가 independent할때.

  pf) $D(p\|q)=0 \iff p=q$ 이므로

  $I(X;Y)=0 \iff p(x,y)=p(x)p(y)$ ∴ X, Y가 서로 independent.

  • $I(X;Y|Z) \geq 0$ 등호 ; X, Y independent given Z.

$H(X) = \log|\mathcal{X}| - D(p\|u)$

✳ Uniform distribution

RV X에 의해 생성되는 symbol 종류를 갖는 집합.

- $|\mathcal{X}|$; $\mathcal{X}$의 원소의 개수라 하면

  • $H(X) \leq \log|\mathcal{X}|$. 등호 : X가 $\mathcal{X}$에 대해 uniform distribution.

pf) $u(x) = \frac{1}{|\mathcal{X}|}$: uniform 확률 분포.

$p(x)$: X에 대한 확률 분포 라 하면

$$D(p\|u) = \sum p(x) \log \frac{p(x)}{u(x)} = \sum p(x) \log p(x) - \sum p(x) \log u(x)$$

$$= -H(X) - \sum p(x) \log \frac{1}{|\mathcal{X}|}$$

$$= -H(X) + \log|\mathcal{X}| \underbrace{\sum p(x)}_{=1}$$

$$= -H(X) + \log|\mathcal{X}| \geq 0$$

(information inequality. (등호 : $p=u$))

∴ $H(X) \leq \log|\mathcal{X}|$. 등호는 $p=u$ (uniform) .

* (conditional Entropy reduce.

— $H(X|Y) \leq H(X)$. 등호 : X, Y independent.

pf) $I(X;Y) = H(X) - H(X|Y) \geq 0$.

— 특정 RV Y가 known 이면, entropy가 줄어듦.

* Independence Bound and entropy

— $X_1, \cdots, X_n \sim p(x_1, \cdots, x_n)$ 이면

· $H(X_1, \cdots, X_n) \leq \sum_{i=1}^{n} H(X_i)$. 등호 : x_i 들이 independent

pf) $H(X_1, \cdots, X_n) = \sum_{i=1}^{n} H(X_i | X_1, \cdots, X_{i-1}) \leq \sum_{i=1}^{n} H(X_i)$ //

* Log sum inequality

— Non-negative number $a_1, \cdots, a_n$ 과 $b_1, \cdots, b_n$ 에 대하

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

pf) $f(t) = t \log t \rightarrow f''(t) = \frac{1}{t} \log e > 0$ for all $t > 0$.

$\Rightarrow \sum \alpha_i f(t_i) = f(\sum \alpha_i t_i)$, $\alpha_i = \frac{b_i}{\sum b_i}$, $t_i = \frac{a_i}{b_i}$, $\sum b_i = \lambda$ 이라.

$\Rightarrow \sum_i \frac{b_i}{\sum_j b_j} \frac{a_i}{b_i} \log\left(\frac{a_i}{b_i}\right) = \left(\sum_i \frac{b_i}{\sum_j b_j} \cdot \frac{a_i}{b_i}\right) \log \left(\sum_i \frac{b_i}{\sum_j b_j} \cdot \frac{a_i}{b_i}\right)$

$\Rightarrow \sum_i \frac{a_i}{\lambda} \log \frac{a_i}{b_i} = \sum_i \frac{a_i}{\lambda} \log \frac{\sum_i a_i}{\sum_j b_i}$

$\Rightarrow \therefore \sum_i a_i \log \frac{a_i}{b_i} = \left(\sum_i a_i\right) \log \frac{\sum_i a_i}{\sum_i b_i}$

✳ Convexity of Relative Entropy

- $D(p\|q)$ is convex in pair (p,q) : $(p_1,q_1),(p_2,q_2)$ 에 대해

$$= D(\lambda p_1 + (1-\lambda)p_2 \,\|\, \lambda q_1 + (1-\lambda)q_2) = \lambda D(p_1\|q_1) + (1-\lambda) D(p_2\|q_2)$$

pf) log sum equality ; $\sum a_i \log \frac{a_i}{b_i} \geq \sum a_i \log \frac{\sum a_i}{\sum b_i}$ --- ①

$a_1 = \lambda p_1$, $a_2 = (1-\lambda)p_2$, $b_1 = \lambda q_1$, $b_2 = (1-\lambda)q_2$ 라 하면

① : $(\lambda p_1 + (1-\lambda)p_2) \log \frac{\lambda p_1 + (1-\lambda)p_2}{\lambda q_1 + (1-\lambda)q_2} \leq \lambda p_1 \log \frac{\lambda p_1}{\lambda q_1} + (1-\lambda)p_2 \log \frac{(1-\lambda)p_2}{(1-\lambda)q_2}$

$$= \lambda p_1 \log \frac{p_1}{q_1} + (1-\lambda)p_2 \log \frac{p_2}{q_2}$$

위의 식을 모든 x에 대해 더함 : $\sum_x$ $\longrightarrow$ $\sum_x p(x) \log \frac{q(x)}{p(x)} = D(p\|q)$

$\Rightarrow D(\lambda p_1 + (1-\lambda)p_2 \,\|\, \lambda q_1 + (1-\lambda)q_2) \leq \lambda D(p_1\|q_1) + (1-\lambda) D(p_2\|q_2)$ ∎

✳ Concavity of Entropy

- $H(p)$는 concave function이다. 즉,

- $H(\lambda p_1 + (1-\lambda)p_2) \geq \lambda H(p_1) + (1-\lambda) H(p_2)$

pf) $H(p) = \underbrace{\log|\mathcal{X}|}_{\text{concave}} - \underbrace{D(p\|u)}_{\text{convex}}$ $\Rightarrow$ concave function.

✳ Mutual Information and Input distribution

— $(X,Y) \sim p(x,y) = p(x)\,p(y|x)$ 이라 하면

① $p(y|x)$가 주어졌을 때, $I(X;Y)$는 $p(x)$에 대한 concave function 이다.
(Channel transaction probability) (변화 가능) = 최대 값이 있다.

pf) $I(X;Y) = H(Y) - H(Y|X) = H(Y) - \sum_x p(x)\,H(Y|x)$

$p(y|x)$ fix 이고 $H(Y)$는 $p(y)$에 대한 concave function 이므로 $H(Y)$는 $p(x)$에 대해 concave. ($\because p(y) = \sum_x p(x)\,p(y|x)$)

$\sum_x p(x)\,H(Y|x)$ → linear function of $p(x)$

따라서 $I(X;Y)$는 concave □

② $p(x)$가 주어졌을 때, $I(X;Y)$는 $p(y|x)$에 대한 convex function 이다.
(채널의 변화)

pf) RV: Y_1 Y_2 → Y는 Y_1, Y_2의 mixed.

2개의 channel $p_1(y|x)$, $p_2(y|x)$ 이 있다고 하면,
RV Z에 의해 한 channel이 선택된다 하자. (λ로 p_1, $1-\lambda$로 p_2)

$$I(X;Y,Z) = I(X;Y|Z) + I(X;Z)$$
$$= I(X;Z|Y) + I(X;Y)$$

Z는 X와 상관 없음. $I(X;Z) = 0$

$\therefore I(X;Y) \le I(X;Y|Z) = \lambda I(X;Y_1) + (1-\lambda) I(X;Y_2)$ $\therefore$ convex.

✳ Markov Chain

— 3개의 RV X, Y, Z 에 대해 Z는 Y의 영향만 받고, Y는 X의 영향만 받을 때, 즉

• $p(x,y,z) = p(x)\,p(y|x)\,p(z|y,x) \rightarrow p(x)\,p(y|x)\,p(z|y)$

이면 X, Y, Z는 그 순으로 Markov Chain 이며, $X \rightarrow Y \rightarrow Z$ 로 표기.

— $p(x,z|y) = p(x|y)\,p(z|y)$ Y가 주어지면 X, Z는 independent.

— $X \rightarrow Y \rightarrow Z \Leftrightarrow Z \rightarrow Y \rightarrow X$

— $Z = f(Y)$면 $X \rightarrow Y \rightarrow Z$.

✳ Data Processing Inequality

갈수록 MI가 줄어든다.

- $X \to Y \to Z$ 면, $I(X;Y) \geq I(X;Z)$

pf) $I(X;Y,Z) = I(X;Y|Z) + I(X;Z)$

$= I(X;Z|Y) + I(X;Y)$

여기서 $X \to Y \to Z$ 이므로 $I(X;Z|Y) = 0$ ∵ Y가 주어지면 X, Z는 independent

그러므로 $I(X;Y) = I(X;Z) + I(X;Y|Z) \geq I(X;Z)$ □

또한 $I(X;Y) \geq I(X;Y|Z)$

✳ Fano's Inequality

— X를 Y로 estimate한 후에 error boundary를 제공.

— X (sent) → Y (observed)

- $\hat{X} = g(Y)$: decoding (복원)
- $P_e = Pr(\hat{X} \neq X)$

— Error

- $E = \begin{cases} 1 & \text{if } \hat{X} \neq X \\ 0 & \text{if } \hat{X} = X \end{cases}$

— $\hat{X} = g(Y)$ 이므로 $X \to Y \to \hat{X}$

- Fano's Inequality : $H(P_e) + P_e \log|\mathcal{X}| \geq H(X|\hat{X}) \geq H(X|Y)$

$\Rightarrow P_e \geq \dfrac{H(X|Y) - 1}{\log|\mathcal{X}|}$

$\because I(X;\hat{X}) \leq I(X;Y)$, $H(X) - H(X|\hat{X}) \geq H(X) - H(X|Y)$

pf) $X \to Y \to \hat{X}$, E 또한 RV.

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

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

- $H(E|\hat{X}) \leq H(E) = H(P_e)$
- $H(E|X,\hat{X})$: X, $\hat{X}$ 면 E를 알수 있음. $= 0$.
- $H(X|E,\hat{X}) = Pr(E=0)\,H(X|\hat{X},E=0) + Pr(E=1)\,H(X|\hat{X},E=1)$

$\leq (1-P_e)\cdot 0 + P_e \cdot \log|\mathcal{X}|$

(예제서면, $H(X|\hat{X}) \leq H(X) \leq \log|\mathcal{X}|$)

$\therefore H(X|\hat{X}) \leq H(P_e) + P_e \log|\mathcal{X}|$, $H(X|Y) \leq H(P_e) + P_e \log|\mathcal{X}|$ □

✳ Fano 정리

— RV X, Y에 대해 $p = P_r(X \neq Y)$ 라 하면

• $H(X|Y) \leq H(p) + p \log |\mathcal{X}|$

— Corollary : More strict

• $\hat{X} : Y \to X$, $P_e = P_r(X \neq \hat{X})$ 라 하면

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

(1)아야 나면 그 자리 1개 빼졌다는 소리.