Information Theory

김상우

Introduction

Information Theory

Scientific study of data as it is transferred, stored, retrieved
정보의 불확실성을 다루는 확률적인 이론

목적

- ▷ Ultimate Limit of Data Compression
- ▷ Ultimate Limit of Reliable Data Transmission
- ▷ Rate-Distortion Theory

용어 정리

Information: 앞으로 일어날 가능성이 있는 Event

Information Source: 정보를 발생시키는 근원

$X \sim p(x)$: $p(x)$ is the prob. dist. function for X

Ex. Information Source: 주사위 던지기

Information: 주사위를 던졌을 때 발생할 수 있는 event

Uncertainty

Information은 불확정성 (uncertainty)와 밀접한 관련이 있다

나올 수 있는 결과들이 많을수록, uncertainty가 높고 정보량이 많다

- 주사위 던지기
- 1~100까지 쓰인 카드에서 뽑기
- 로또 뽑기

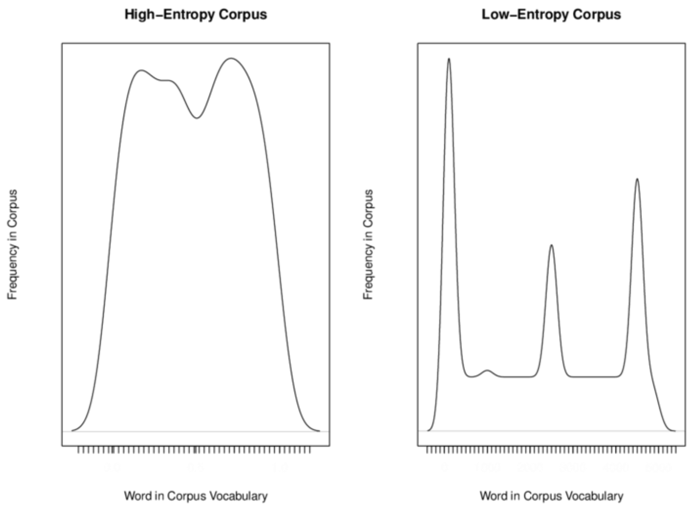
정보 이론에서는, '새롭게 얻을 수 있는 데이터'가 정보이다

Entropy

Entropy

정보량을 구체적으로 측정하는 척도
RV이 가지고 있는 확률분포의 불확정성을 측정하는 척도

Entropy
~ Uncertainty
~ Self Information
~ Information Content
~ Average Surprise



Entropy

Intuition

- 작은 확률: 높은 엔트로피를 가지므로 반비례 관계
- 단순히 $\frac{1}{p(x)}$을 사용하면 $p(x)$와 지워진다
- 다른 증가함수인 log함수를 취해준다
  (꼭 log 함수일 필요는 없다)

$$H(X) = \sum_x p(x) \log \frac{1}{p(x)}$$
$$= -\sum_x p(x) \log p(x)$$
$$= E_p \log \frac{1}{p(x)}$$

하나의 x에 대해 p(x)가 1인 경우 (delta function)
$\lim_{p \to 0} p \log \frac{1}{p} = 0$ 이므로, entropy = 0

Entropy

Properties

▷ H(X) is shift invariant

Proof)

$$\text{Suppose } Y = X + a$$

$$p_Y(y) = p_X(x)$$

$$H(Y) = \sum_y p_Y(y) \log \frac{1}{p_Y(y)}$$

$$= \sum_x p_X(x) \log \frac{1}{p_X(x)}$$

Entropy

Properties

- ▷ H(X)≥0

Proof)

$$\log \frac{1}{p(x)} \geq 0 \; when \; 0 \leq p(x) \leq 1$$

$$\therefore H(X) = \sum_{x} p(x) \log \frac{1}{p(x)} \geq 0$$

Entropy

Properties

- If X is uniform with n values, H(X)=log(n)   (n: cardinality)

Proof)

$$H(X) = \sum_{i=1}^{n} \frac{1}{n} \log \frac{1}{1/n} = \log n$$

A fair coin toss : H = 1
A fair die : H = $\log_2(6)$

Entropy

Properties

- ▷ $H_b(X) = (\log_b a) H_a(X)$

Proof)

$$\log_b p = (\log_b a) \log_a p$$

Joint Entropy

Definition

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

- If X and Y are independent, H(X, Y) = H(X) + H(Y)

Joint Entropy

Properties

- If X and Y are independent, H(X, Y) = H(X) + H(Y)

Proof)

$$
\begin{aligned}
H(X,Y) &= -E[\log p(x,y)] \\
&= -E[\log p(x) + \log p(y)] \\
&= -E[\log p(x)] - E[\log p(y)] \\
&= H(X) + H(Y)
\end{aligned}
$$

Definition 6.1 *Let X and Y be two* ***discrete*** *random variables.*

The ***joint probability distribution of X and Y*** *is given by*

$$p_{X,Y}(x, y) = P(X = x, Y = y)$$

Here $p_{X,Y}(x, y)$ is defined for all real numbers x and y.

The ***marginal distribution of X*** **is**

$$p_X(x) = P(X = x) = \sum_{\text{all } y} p_{X,Y}(x, y).$$

Similarly, the ***marginal distribution of Y*** **is**

$$p_Y(y) = P(Y = y) = \sum_{\text{all } x} p_{X,Y}(x, y).$$

Conditional Entropy

Definition

$$H(Y|X=x) = \sum_y p(y|x) \log \frac{1}{p(y|x)}$$

$$\begin{aligned} H(Y|X) &= E_{p(x)} H(Y|X=x) \\ &= \sum_x p(x) \sum_y p(y|x) \log \frac{1}{p(y|x)} \\ &= -E_{\boxed{p(x,y)}} \log \boxed{p(y|x)} \end{aligned}$$

Mismatch!

Conditional Entropy

Properties

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

Proof)

$$\begin{aligned} H(X,Y) &= E\left[\log\frac{1}{p(x,y)}\right] \\ &= E\left[\log\frac{1}{p(x)p(y|x)}\right] \\ &= E\left[\log\frac{1}{p(x)}\right] + E\left[\log\frac{1}{p(y|x)}\right] \\ &= H(X) + H(Y|X) \end{aligned}$$

Conditional Entropy

Properties

- H(X, Y) = H(X) + H(YlX)

Proof)

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

Relative Entropy (KL Divergence)

Definition

$$D(p(x)\|q(x)) = \sum_x p(x)\log\frac{p(x)}{q(x)}$$
$$= E_{p(x)}[\log\frac{p(x)}{q(x)}]$$

- Entropy와 Mutual Information의 중간 단계
- 2가지 distribution 간의 거리를 재는 방법

Relative Entropy (KL Divergence)

Maximum Likelihood Estimation

- A technique used to find the optimal parameters of a distribution that best describes a set of data

$$\theta_{MLE} = \arg\max_{\theta} P(x|\theta)$$

Kullback-Leibler Divergence

- Measures the dissimilarity between two probability distributions

$$\begin{aligned} D_{KL}(P \parallel Q) &= \mathbb{E}_{x \sim P(x)} \left[\log \frac{P(x)}{Q(x)} \right] \\ &= \int_{-\infty}^{\infty} P(x) \log \frac{P(x)}{Q(x)} \, dx \end{aligned}$$

Relative Entropy (KL Divergence)

Proof)

KL Divergence의 정의에 의해

$$\begin{aligned}\theta_{\text{min KL}} &= \arg\min_{\theta} D_{KL}\left[P(x|\theta^*) \parallel P(x|\theta)\right] \\ &= \arg\min_{\theta} \mathbb{E}_{x\sim P(x|\theta^*)}\left[\log \frac{P(x|\theta^*)}{P(x|\theta)}\right] \\ &= \arg\min_{\theta} \mathbb{E}_{x\sim P(x|\theta^*)}\left[\log P(x|\theta^*) - \log P(x|\theta)\right]\end{aligned}$$

θ^*은 target distribution의 parameters이므로 고정

$$\begin{aligned}\theta_{\text{min KL}} &= \arg\min_{\theta} \mathbb{E}_{x\sim P(x|\theta^*)}\left[-\log P(x|\theta)\right] \\ &= \arg\max_{\theta} \mathbb{E}_{x\sim P(x|\theta^*)}\left[\log P(x|\theta)\right]\end{aligned}$$

Relative Entropy (KL Divergence)

Proof)

$$
\begin{aligned}
\theta_{\min \mathrm{KL}} &= \arg\max_{\theta} \mathbb{E}_{x \sim P(x|\theta^*)} \left[\log P(x|\theta)\right] \\
&= \arg\max_{\theta} \lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^{n} \log P(x_i|\theta) \\
&= \arg\max_{\theta} \log P(x|\theta) \\
&= \arg\max_{\theta} P(x|\theta) \\
&= \theta_{MLE}
\end{aligned}
$$

Law of Large Numbers

여기서부터 x는 probability가 아니고, data samples $x = \{x_1, x_2, \dots, x_n\}$ 를 의미

Cross Entropy

Definition

$$CE(p,q) = E_p[-log q] = -\sum_{x\in X} p(x) log q(x) = H(p) + \boxed{D_{KL}(p||q)}$$

$$\boxed{\begin{aligned} D_{KL}(P \parallel Q) &= \mathbb{E}_{x\sim P(x)}\left[\log \frac{P(x)}{Q(x)}\right] \\ &= \int_{-\infty}^{\infty} P(x) \log \frac{P(x)}{Q(x)}\, dx \end{aligned}}$$

Mutual Information

Mutual Information

Definition

$$I(X;Y) = \sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)}$$
$$= D(p(x,y) \| p(x)p(y))$$

Joint Distribution

Marginal Distribution

- X, Y가 independent 하면 mutual information = 0
- 두 RV 간의 correlation을 측정하므로, correlation이 없으면 0이 된다
- 반대로, X, Y가 같아지면 I(X;Y)는 최대가 된다 (H(X)와 같아진다)

Mutual Information

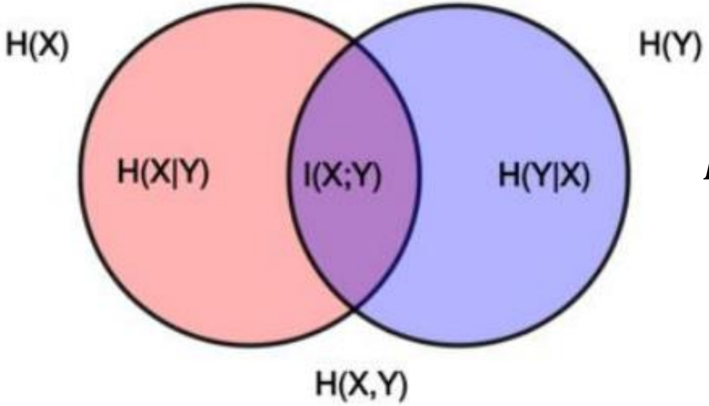
Properties

- I(X;Y) = H(X) – H(X|Y)

Proof)

$$
\begin{aligned}
I(X;Y) &= E\left[\log \frac{p(x,y)}{p(x)p(y)}\right] \\
&= E\left[\log \frac{p(x|y)p(y)}{p(x)p(y)}\right] \\
&= E\left[\log \frac{p(x|y)}{p(x)}\right] \\
&= E\left[\log \frac{1}{p(x)}\right] - E\left[\log \frac{1}{p(x|y)}\right] \\
&= H(X) - H(X|Y)
\end{aligned}
$$

Mutual Information



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

Figure A. The Venn diagram depicting the relationship between individual ($H(X)$, $H(Y)$), joint ($H(X,Y)$), and conditional ($H(X|Y)$, $H(Y|X)$) entropies. The intersection of the circles is the mutual information $I(X;Y)$.

Convex Functions

Convex Functions

Definition

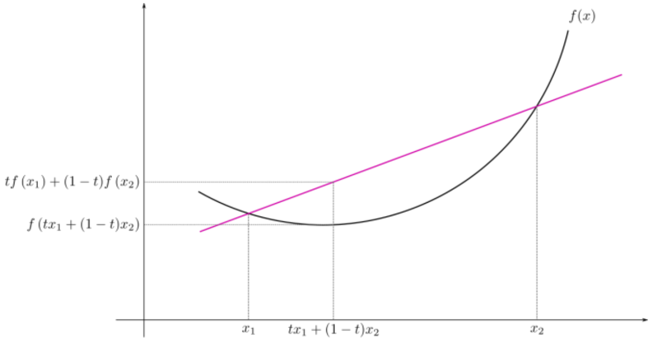
$f(x)$ is convex on $a < x < b$ iff

x_1과 x_2가 같아질 경우, 등호 성립

$$f\big((1-\lambda x_1) + \lambda x_2\big) \le (1-\lambda) f(x_1) + \lambda f(x_2)$$

for $\forall x_1, x_2 \ s.t. \ a < x_1, x_2 < b$

- f(x) is concave if −f(x) is convex
- If ∃f''(x) & f''(x)≥0, f(x) is convex



Jensen Inequality

Definition

For any convex $\mathrm{f}(x)$,
$\mathrm{f}(\mathrm{E}[\mathrm{X}]) \leq \mathrm{E}[\mathrm{f}(\mathrm{X})]$

▷ 수학적 귀납법으로 증명

2 points: convex function의 정의에 의해 성립
k-1 points: $p_i' = \frac{p_i}{\sum_{j=1}^{k-1} p_j} = \frac{p_i}{1-p_k}$를 두어 증명

Jensen Inequality

For any convex $\mathrm{f}(x)$,
$\mathrm{f}(\mathrm{E}[\mathrm{X}]) \leq \mathrm{E}[\mathrm{f}(\mathrm{X})]$

Properties

▷ X가 deterministic 할 경우 (random variable이 아닐 경우), 등호 성립

Information Inequality

Definition

$$D(p(x)\|q(x)) \geq 0 \text{ with equality iff } p(x) = q(x)$$

▷ Relative Entropy는 항상 양수임을 Jensen Inequality로 증명 가능

Information Inequality

Proof)

$$
\begin{aligned}
D(p\|q) &= \sum_x p(x)\log\frac{p(x)}{q(x)} \\
&= \sum_x p(x)\{-\log\frac{q(x)}{p(x)}\} \\
&\geq -\log\left\{\sum_x p(x)\frac{q(x)}{p(x)}\right\} \\
&= -\log\sum_x q(x) = -\log 1 = 0
\end{aligned}
$$

- log 함수는 convex 하다

$\frac{q(x)}{p(x)}$가 constant일 경우, 등호 성립

p와 q는 확률 분포이므로,
p(x)=q(x)일 때만 등호 성립

Information Inequality

Properties

▷ I(X;Y) ≥ 0 : Mutual Information은 항상 positive하다

Proof)

$$I(X;Y) = \sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)}$$
$$= D(p(x,y) \| p(x)p(y)) \geq 0$$

Information Inequality

Properties

- H(X) ≤ log|χ| : 모든 entropy는 log(χ)를 넘을 수 없다 (χ: cardinality)
- H(X)는 uniform distribution일 때 최대이다

Proof)

$$\text{Let } u(x) = \frac{1}{n} \text{ be the uniform distribution}$$

$$\begin{aligned} 0 &\le D(p(x) \| u(x)) \\ &= \sum_x p(x) \log \frac{p(x)}{1/n} \\ &= \log n - H(X) \\ &= \log|\boldsymbol{\chi}| - H(X) \end{aligned}$$

p(x)가 uniform distribution일 경우, 등호 성립

Information Inequality

Properties

- ▷ I(X;Y) = 0 iff X & Y are independent

Proof)

$$I(X;Y) = \sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)}$$

$$= D(p(x,y) || p(x)p(y))$$

RHS is 0 iff $p(x,y) = p(x)p(y)$

Information Inequality

Properties

- Conditioning reduces entropy (Theorem)

  H(X|Y) ≤ H(X)
  with equality iff X & Y are independent

Proof)

$$I(X;Y) = H(X) - H(X|Y) \geq 0$$
$$\therefore H(X|Y) \leq H(X)$$
$$I(X;Y) = 0 \text{ iff } X\&Y \text{ are independent}$$

Jensen-Shannon Distance

$$D_{JS}(p||q) = \frac{1}{2}D_{KL}(p||\frac{p+q}{2}) + \frac{1}{2}D_{KL}(q||\frac{p+q}{2})$$

KL Divergence

- Positivity (O)
- Symmetry (X)
- Triangle Inequality (X)

Jensen-Shannon Distance

- Positivity (O)
- Symmetry (O)
- Triangle Inequality (O)

Conditional Mutual Information

Conditional Mutual Information

Definition

$$I(X;Y|Z) = \sum_{z} p(z) I(X;Y|Z=z)$$
$$= \sum_{x,y,z} p(x,y,z) \log \frac{p(x,y|z)}{p(x|z)p(y|z)}$$
$$= \sum_{x,y,z} p(x,y,z) \log \frac{p(y|x,z)}{p(y|z)}$$
$$= H(Y|Z) - H(Y|X,Z)$$

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

Examples

Joint PDF

WX \ YZ	00	01	10	11
00	$\frac{1}{8}$	0	$\frac{1}{8}$	0
01	0	$\frac{1}{8}$	0	$\frac{1}{8}$
10	$\frac{1}{8}$	0	$\frac{1}{8}$	0
11	0	$\frac{1}{8}$	0	$\frac{1}{8}$

$H(W) = H(X) = H(Y) = H(Z) = 1$

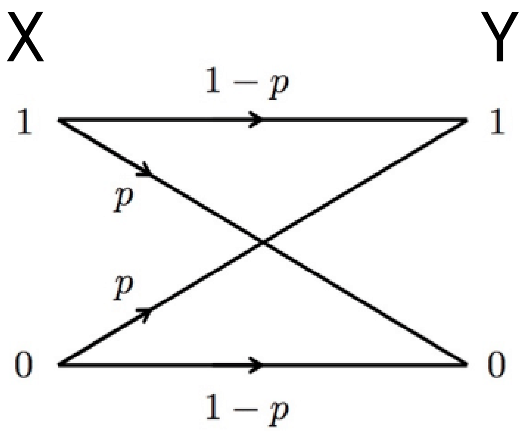
$H(WX) = H(YZ) = H(XY) = \cdots = 2$

$H(WXY) = H(WXZ) = H(WYZ) = H(XYZ) = 3$

$H(Z \mid WXY) = 0$

$H(WXYZ) = H(WXY) + H(Z \mid WXY) = 3$

Binary Symmetric Channel



$$p(y|x) = \begin{cases} 1-p & \text{if } y = x, \\ p & \text{if } y \neq x. \end{cases}$$

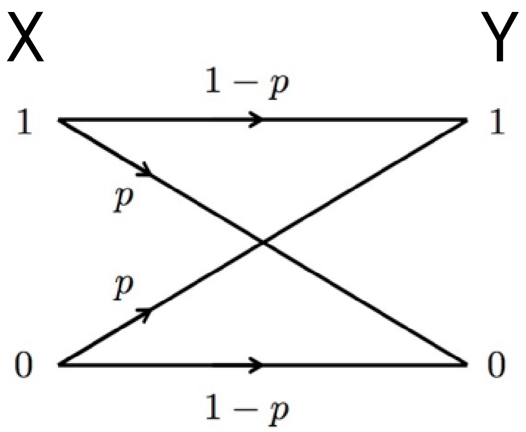
$$H(Y|X=0) = H(\{1-p, p\}) = H(p)$$
$$H(Y|X=1) = H(\{p, 1-p\}) = H(p)$$

$$\begin{aligned} I(X;Y) &= H(Y) - H(Y|X) \\ &= H(\{p_Y(0), p_Y(1)\}) - H(p) \end{aligned}$$

$$p_Y(0) = (1-p) * p_X(0) + p * p_X(1)$$
$$p_Y(1) = p * p_X(0) + (1-p) * p_X(1)$$

$$\max I(X;Y) = 1 - H(p)$$

Binary Symmetric Channel



$$p(y|x) = \begin{cases} 1 - p & \text{if } y = x, \\ p & \text{if } y \neq x. \end{cases}$$

$$I(X;Y) = H(\{P_Y(0), P_Y(1)\}) - H(p)$$

$$P_Y(0) = (1-p)\cdot(1-q) + p\cdot q$$

$$P_Y(1) = p\cdot(1-q) + (1-p)\cdot q$$

$$\left[\begin{array}{l} P_X(0) = 1-q \\ P_X(1) = q \end{array}\right]$$

$$P_Y(0) = 2pq - p - q + 1 = (2p-1)q + (1-p) = Aq + B$$

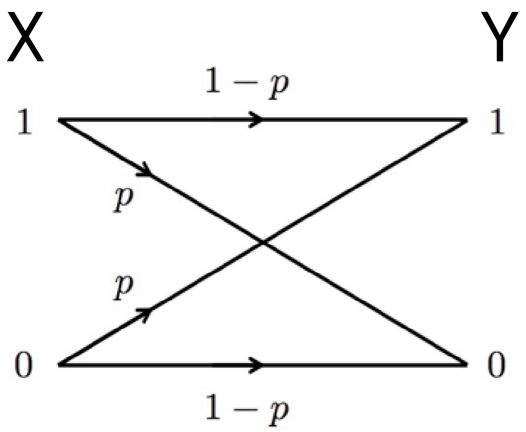
$$P_Y(1) = p + q - 2pq = (1-2p)q + p = Cq + D$$

$$H(Y) = H(\{P_Y(0), P_Y(1)\})$$

$$= -P_Y(0)\log P_Y(0) - P_Y(1)\log P_Y(1)$$

$$= -(Aq+B)\log(Aq+B) - (Cq+D)\log(Cq+D)$$

Binary Symmetric Channel



$$p(y|x) = \begin{cases} 1-p & \text{if } y = x, \\ p & \text{if } y \neq x. \end{cases}$$

$$\frac{d}{dq}[H(q)] = -A\log(Aq+B) - \frac{A\cdot(Aq+B)}{Aq+B}$$

$$- C\log(Cq+D) - \frac{C\cdot(Cq+D)}{Cq+D}$$

$$= -A\log(Aq+B) - C\log(Cq+D) - A - C$$

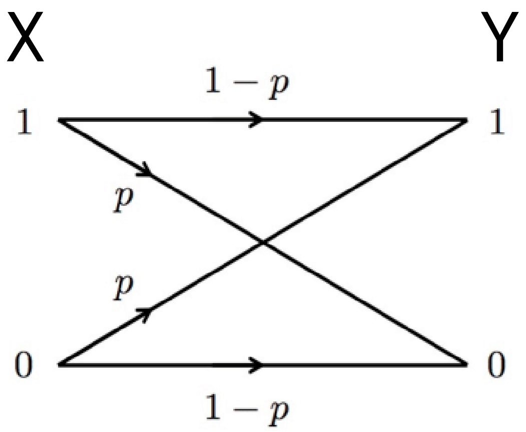
$$= -A\log(Aq+B) + A\log(-Aq+1-B)$$

$$(\because C = -A,\ D = 1-B)$$

$$= A\log\frac{1-(Aq+B)}{Aq+B}$$

$$= A\log\left(\frac{1}{Aq+B} - 1\right)$$

Binary Symmetric Channel



$$p(y|x) = \begin{cases} 1-p & \text{if } y = x, \\ p & \text{if } y \neq x. \end{cases}$$

$\frac{1}{Aq+B}-1=1 \Rightarrow Aq+B=\frac{1}{2}.$

$(2p-1)q+(1-p)=\frac{1}{2}$

$(4p-2)q+(2-2p)=1$

$q=\frac{2p-1}{4p-2}=\frac{2p-1}{2(2p-1)}=\frac{1}{2}.$

$q=\frac{1}{2}$ 일 때, $P_Y(0)=P_Y(1)=\frac{1}{2}$

(uniform distribution)

$\max I(X;Y)=1-H(p)$

Thank You