# Mutual Information

Definition (Mutual Information): Let $(X,Y) \sim P(\mathcal{X} \times \mathcal{Y})$. The mutual information between $X$ and $Y$ is

$$I(X;Y) \triangleq D_{KL}\left(P_{XY} \,\|\, P_X \otimes P_Y\right)$$

where $P_X$ and $P_Y$ are the $X$ and $Y$ marginals of $P_{XY}$ and $P_X \otimes P_Y$ is the induced product measure.

Remark:

(i) MI is a "fundamental" measure of dependence between random variables

(ii) A typical interpretation of mutual information is a quantification of the amount of "info" that $X$ and $Y$ convey about eachother.

Proposition (Basic Properties of MI):

① $I(X;Y) \geq 0$ w/ equality iff $X \perp\!\!\!\perp Y$

② $I(X;Y) = D_{KL}\left(P_{Y|X} \,\|\, P_Y \,|\, P_X\right)$

③ $I(X;Y) = I(Y;X)$

④ $I(X;Y) \geq I(X;f(Y))$ for any deterministic function w/ equality iff $f$ is a bijection.

⑤ $\underbrace{I(X,Y;Z)}_{D_{KL}(P_{XYZ} \| P_{XY} \otimes P_Z)} \geq I(X;Z)$ ← More information

w/ equality iff $Z \perp\!\!\!\perp Y|X$

**Proof**

① by definition

② Chain rule of KL divergence

$$D_{KL}(P_{XY} \| Q_{XY}) = D_{KL}(P_X \| Q_X) + D_{KL}(P_{Y|X} \| Q_{Y|X} | P_X)$$

- For us $Q_{XY} = P_X \otimes P_Y$

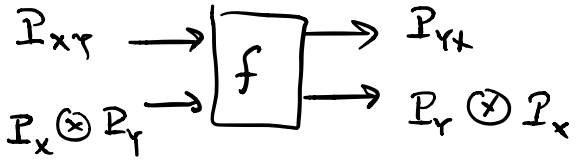
$$Q_X = P_X \quad Q_{Y|X} = P_Y$$

$$\Rightarrow D_{KL}(P_X \| Q_X) + D_{KL}(P_{Y|X} \| Q_{Y|X} | P_X)$$

$$= D_{KL}(P_X \| P_X) + D_{KL}(P_{Y|X} \| P_Y | P_X)$$

$$= 0 + D_{KL}(P_{Y|X} \| P_Y | P_X)$$

③ Let $f(x,y) = (y,x)$ apply in **BOTH** directions



④ Establish later

⑤ Let $f(X,Y,Z) = (X,Z)$ & use DPI

Proposition (I vs. H)

①

$$I(X;X) = \begin{cases} H(X) & , \ X \text{ discrete} \\ \infty & , \ \text{o/w} \end{cases}$$

② Discrete X

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

③ Continuous X

$$\begin{aligned} I(X;Y) &= h(X) + h(Y) - h(X,Y) \\ &= h(X) - h(X|Y) \\ &= h(Y) - h(Y|X) \end{aligned}$$

Proof:

① $D_{KL}(X;X) = D_{KL}(P_{X|X} \| P_X | P_X)$

$$P_{X|X}(\cdot|x) = \delta_x(\cdot), \quad x \in \text{supp}(P_X)$$

(i) Discrete: $\delta_x \ll P_X$

$$I(X;X) = D_{KL}(P_{X|X} \| P_X | P_X)$$

$$= \sum_{x \in \mathcal{X}} p_X(x) D_{KL}(\underbrace{P_{X|X}(\cdot|x)}_{\delta_x(\cdot)} \| P_X)$$

$$= \sum_{x \in \mathcal{X}} p_X(x) \sum_{x' \in \mathcal{X}} \delta_x(x') \frac{\log(\delta_x(x'))}{p_X(x')}$$

$$= \sum_{x \in \mathcal{X}} p_X(x) \log\left(\frac{1}{p_X(x)}\right) = H(X)$$

(ii) Continuous: $I(X;X) = D_{KL}(P_{X,X} \| P_X \otimes P_X)$

and we will show $P_{XX} \not\ll P_X \otimes P_X$

Let

$$\Delta = \{(x,x) \mid x \in \mathcal{X}\}$$

$$P_{XX}(\Delta) = \int_{\Delta} dP_{XX} = \int_{\mathcal{X}} \int_{\mathcal{X}} dP_{XX}(x,x') \mathbb{1}_{\{x=x'\}} = \int_{x \in \mathcal{X}} dP_X(x) \int_{x' \in \mathcal{X}} dP_{X|X}(x'|x) \mathbb{1}_{\{x=x'\}}$$

$$= \int_{x \in \mathcal{X}} dP_X(x) \int_{x' \in \mathcal{X}} d\delta_x(x') = 1 > 0$$

$$\int_{\{x\}} d\delta_x(x') = \delta_x(\{x\}) = 1$$

However

$$P_X \otimes P_X(\Delta) = \int_{\Delta} dP_X \otimes P_X(x, x')$$

$$= \int_{\mathcal{X}} dP_X(x) \int_{\mathcal{X}} dP_X(x') \, \mathbb{1}_{\{x = x'\}} = 0$$

$$\int_{\{x\}} dP_X(x') = P_X(\{x\}) = 0$$

∎

This concludes the proof for continuous $X$ case.

For an arbitrary non-discrete variable $X$:

$$A := \{x \in \mathcal{X} \mid P_X(\{x\}) > 0\}$$

$$\Delta_A := \{(x, x) \mid x \in A^c\}$$

and repeat continuous argument.

② $$I(X;Y) = \sum_{x,y} P_{XY}(x,y) \log \frac{P_{XY}(x,y)}{P_X(x) P_Y(y)}$$

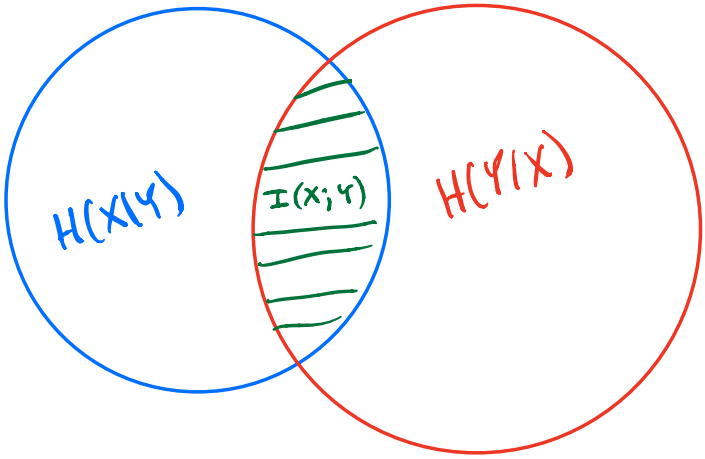
$$= \sum_{xy} P_{XY}(x,y) \log \frac{P_Y(y) P_{X|Y}(x|y)}{P_X(x) P_Y(y)}$$

$$= \sum_{xy} P_{XY}(x,y) \log \frac{1}{P_X} - \sum_{xy} P_{XY}(x,y) \log \frac{1}{P_{X|Y}(x|y)}$$

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

∎

Illustration



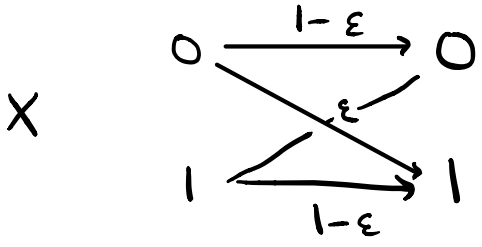

## Examples

① BSC $\quad P_X = X \sim \text{Ber}(\tfrac{1}{2}) \perp\!\!\!\perp P_Z = Z \sim \text{Ber}(\varepsilon), \quad \varepsilon \in [0, 1/2]$

and

$$Y = X \oplus Z \pmod 2 = \begin{cases} x \oplus 0 = x, & Z = 0 \text{ w.p. } 1-\varepsilon \\ x \oplus 1 = \bar{x}, & Z = 1 \text{ w.p. } \varepsilon \end{cases}$$

⇒ So




is a binary symmetric channel w/ parameter $\varepsilon$.

$$I(X;Y) = \underbrace{H(Y)}_{\text{(a)}} - \underbrace{H(Y|X)}_{\text{(b)}}$$

(a) Start by computing pmf $P_Y$

$$p_Y(0) = p_X(0)\, p_{Y|X}(0|0) + p_X(1)\, p_{Y|X}(0|1) = \tfrac{1}{2}(1-\varepsilon) + \tfrac{1}{2}\varepsilon = \tfrac{1}{2}$$

$$\Rightarrow Y \sim \text{Ber}(\tfrac{1}{2}) \Rightarrow H_b(\tfrac{1}{2}) = 1$$

(b) Safe way

$$H(Y|X) = p_X(0)\,H(Y|X=0) + p_X(1)\,H(Y|X=1) = \cdots$$

→ Instead, consider the following.

$$H(Y|X) = \sum_{x\in\{0,1\}} p_X(x)\,H(Y|X=x)$$

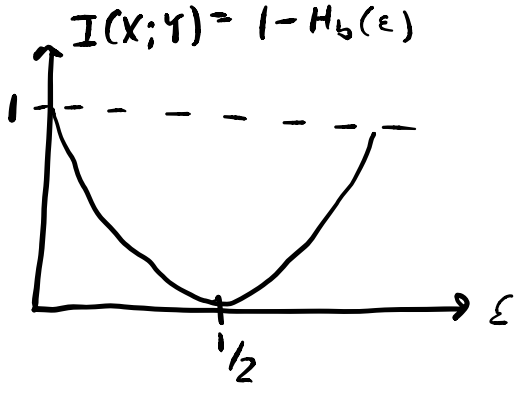
$$= \sum_{x\in\{0,1\}} p_X(x)\,H(X\oplus Z \mid X=x)$$

$$= \sum_{x\in\{0,1\}} p_X(x)\,H(x\oplus Z \mid X=x)$$

$$H(x\oplus Z\oplus x \mid X=x) = H(Z|X=x)$$

$$= \sum_{x\in\{0,1\}} p_X(x)\,H(Z|X=x) = H(Z|X) = H(Z) = H_b(\varepsilon)$$

$$\Rightarrow I(X;Y) = 1 - H_b(\varepsilon)$$



(2) Bivariate Gaussian

$$(X,Y) \sim \mathcal{N}\left(\begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}1 & \rho\\ \rho & 1\end{bmatrix}\right)$$

$$I(X;Y) = h(X) + h(Y) - h(X,Y)$$

$$= \frac{1}{2}\log(2\pi e) + \frac{1}{2}\log(2\pi e) - \frac{1}{2}\log\left((2\pi e)^2(1-\rho)^2\right)$$

$$= \frac{1}{2}\log\left(\frac{1}{(1-\rho)^2}\right)$$