ECE 5630: Information Theory for Data Transmission, Security and Machine Learning 3/5/20

Lecture 10: Mutual Information

Lecturer: Prof. Ziv Goldfeld

Scriber: Isay Katsman, *Net ID*: isk22

Assistant Editor: Kia Khezeli

10.1 Mutual Information

Definition 10.1 (Mutual Information) *Let $(X,Y) \sim P_{XY} \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$. The mutual information between X and Y is defined as*

$$I(X;Y) := D_{\mathsf{KL}}(P_{XY}||P_X \otimes P_Y),$$

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 10.1 (Comments) *Note the following:*

(i) Mutual information is a fundamental measure of dependence between random variables: it is invariant to invertible transformations of the random variables, nullifies if and only if random variables are independent, and emerges as a solution to operational data compression and transmission questions.

(ii) We interpret $I(X;Y)$ as the amount of information that X and Y convey about each other.

Proposition 10.1 (Basic Properties of Mutual Information) *Mutual information satisfies the following properties:*

1. *$I(X;Y) \geq 0$ with equality if and only if $X \perp\!\!\!\perp Y$.*
2. *$I(X;Y) = D_{\mathsf{KL}}(P_{Y|X}||P_Y|P_X)$.*
3. *$I(X;Y) = I(Y;X)$.*
4. *$I(X;Y) \geq I(X;f(Y))$ for any deterministic function, with equality if and only if f is a bijection.*
5. *$I(X,Y;Z) \geq I(X;Z)$. Note that $I(X,Y;Z) = D_{\mathsf{KL}}(P_{XYZ}||P_{XY} \otimes P_Z)$.*

Proof:

1. Clear by definition (derives from non-negativity of KL divergence for probability measures).
2. Let $Q_{XY} = P_X \otimes P_Y$ and observe that $Q_X = P_X$ and $Q_{Y|X} = P_Y$. From the chain rule for KL divergences, we have

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

Thus,

$$D_{\mathsf{KL}}(P_{XY}||P_X \otimes P_Y) = \overbrace{D_{\mathsf{KL}}(P_X||P_X)}^{0} + D_{\mathsf{KL}}(P_{Y|X}||P_Y|P_X) = D_{\mathsf{KL}}(P_{Y|X}||P_Y|P_X).$$

3. Let $g(x,y) = (y,x)$ and consider the transition kernel induced by g. Passing $P_{X,Y}$ and $P_X \otimes P_Y$ through g produces $P_{Y,X}$ and $P_Y \otimes P_X$, respectively. Applying the KL divergence DPI to this setup we obtain $D_f(P_{XY}||P_X \otimes P_Y) \geq D_f(P_{YX}||P_Y \otimes P_X)$. Reversing the role of X and Y completes the proof.

4. The proof follows by the mutual information DPI. As will be shown in the next lecture, if $X \to Y \to Z$ forms a Markov chain, then $I(X;Y) \geq I(X;Z)$ with equality if and only if $X \to Z \to Y$. Clearly, $X \to Y \to f(Y)$, and if f is a bijection, then we also have $X \to f(Y) \to Y$.

5. Let $g(x,y,z) = (x,z)$ and consider the induced transition kernel. Passing $P_{X,Y,Z}$ and $P_{X,Y} \otimes P_Z$ through g produces $P_{X,Z}$ and $P_X \otimes P_Z$, respectively. Applying the KL divergence DPI produces the result.

∎

Proposition 10.2 (Mutual Information and Entropy)

1. *$I(X;X) = \begin{cases} H(X), & \text{discrete } X, \\ \infty, & \text{otherwise.} \end{cases}$*

2. *For discrete X: $I(X;Y) = H(X) + H(Y) - H(X,Y) = H(X) - H(X|Y) = H(Y) - H(Y|X)$.*

3. *For continuous X: $I(X;Y) = h(X) + h(Y) - h(X,Y) = h(X) - h(X|Y) = h(Y) - h(Y|X)$.*

Proof:

1. We consider discrete and continuous cases separately. Finally, we extend the derivation for the continuous case to arbitrary non-discrete case.

   (i) **Discrete:** From the definition $I(X,X) = D_{\mathsf{KL}}(P_{X|X} || P_X | P_X)$ where $P_{X|X}(\cdot|x) = \delta_x(\cdot)$. Note that $\delta_x \ll P_X$, for any $x \in \operatorname{supp}(P_X)$. Then,

$$I(X;X) = D_{\mathsf{KL}}(P_{X|X}||P_X|P_X) = \sum_{x\in\mathcal{X}} p_X(x) D_{\mathsf{KL}}(\underbrace{P_{X|X}(\cdot|x)}_{\delta_x(\cdot)}||P_X)$$
$$= \sum_{x\in\mathcal{X}} p_X(x) \sum_{x'\in\mathcal{X}} \delta_x(x') \log \frac{\delta_x(x')}{p_X(x')} = \sum_{x\in\mathcal{X}} p_X(x) \log \frac{1}{p_X(x)} = H(X).$$

   (ii) **Continuous:** Assume $P_X \ll \lambda$ where λ is the Lebesgue measure. From the definition $I(X;X) = D_{\mathsf{KL}}(P_{XX}||P_X \otimes P_X)$. We will show that $P_{XX} \not\ll P_X \otimes P_X$, thereby implying that KL divergence diverges, as claimed. Define the diagonal set $\Delta := \{(x,x) : x \in \mathcal{X}\}$. Then,

$$P_{XX}(\Delta) = \int_\Delta \mathsf{d}P_{XX}(x,x) = \int_\mathcal{X}\int_\mathcal{X} \mathbb{1}_{\{x=x'\}} \mathsf{d}P_{XX}(x,x) = \int_\mathcal{X} \mathsf{d}P_X(x) \int_\mathcal{X} \mathbb{1}_{\{x=x'\}} \mathsf{d}P_{X|X}(x'|x)$$
$$= \int_\mathcal{X} \mathsf{d}P_X(x) \int_\mathcal{X} \mathbb{1}_{\{x=x'\}} \mathsf{d}\delta_x(x') = \int_\mathcal{X} \delta_x(x) \mathsf{d}P_X(x) = 1.$$

   However,

$$P_X \otimes P_X(\Delta) = \int_\Delta \mathsf{d}P_X \otimes P_X(x,x') = \int_\mathcal{X}\int_\mathcal{X} \mathbb{1}_{\{x=x'\}} \mathsf{d}P_X \otimes P_X$$
$$= \int_\mathcal{X} \mathsf{d}P_X(x) \int_\mathcal{X} \mathbb{1}_{\{x=x'\}} \mathsf{d}P_X(x') = \int_\mathcal{X} P_X(x) \mathsf{d}P_X(x) = 0,$$

   where the last equality follows from the fact that $P_X(x) = 0$ for all $x \in \mathcal{X}$ because $P_X \ll \lambda$. Thus, $P_{XX} \not\ll P_X \otimes P_X$ as $P_{XX}(\Delta) > 0$ while $P_X \otimes P_X(\Delta) = 0$.

   (iii) **Non-discrete:** The continuous distribution argument trivially extends to an arbitrary non-discrete scenario. In particular, define $\mathcal{A} := \{x \in \mathcal{X} : P_X(\{x\}) > 0\}$ and $\Delta_\mathcal{A} := \{(x,x) : x \in \mathcal{A}^c\}$. Repeating the above proof for $\Delta_\mathcal{A}$ instead of Δ produces the general result.

2. By definition, $I(X;Y) = \sum_{x,y} P_{XY}(x,y) \log \frac{P_{XY}(x,y)}{P_X(x)P_Y(y)}$ and $P_{XY}(x,y) = P_Y(y)P_{X|Y}(x|y)$. Then,

$$\begin{aligned} I(X;Y) &= \sum_{x,y} P_{XY}(x,y) \log \frac{P_Y(y)P_{X|Y}(x|y)}{P_X(x)P_Y(y)} \\ &= \sum_{x,y} P_{XY}(x,y) \log \frac{1}{P_X(x)} - \sum_{x,y} P_{XY}(x,y) \log \frac{1}{P_{X|Y}(x|y)} = H(X) - H(X|Y). \end{aligned}$$

   By repeating the above argument using $P_{XY}(x,y) = P_X(x)P_{Y|X}(y|x)$ we get $I(X;Y) = H(Y) - H(Y|X)$. Additionally recall from the definition of conditional entropy that $H(Y|X) = H(X,Y) - H(X)$, so we have $I(X;Y) = H(Y) - H(Y|X) = H(Y) + H(X) - H(X,Y)$.

3. The derivation for the continuous case is analogous to the discrete case, and is thus omitted.

■

Remark 10.2 (Illustration) *The relationship between mutual information and entropy is illustrated in Figure 1.*

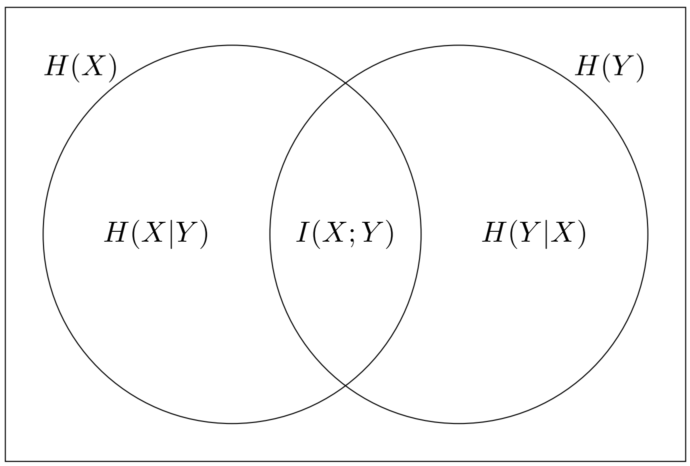

Figure 1: The relationship between mutual information and entropy.

Example 10.1

- <u>*Binary Symmetric Channel (BSC):*</u> *Let $X \sim \mathrm{Ber}(1/2)$ and $Y = X \oplus Z$ (addition modulo 2) where $Z \sim \mathrm{Ber}(\epsilon)$, with $\epsilon \in [0, 1/2]$ independent of X. The BSC is depicted in Figure 2.*

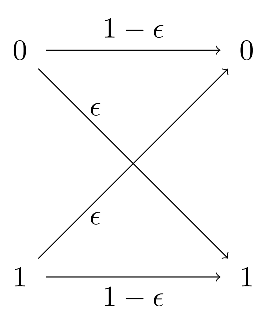

Figure 2: Binary symmetric channel with flip parameter ϵ.

First observe that

$$Y = \begin{cases} X \oplus 0, & Z = 0, \\ X \oplus 1, & Z = 1. \end{cases} = \begin{cases} X, & w.p.\ 1-\epsilon, \\ 1-X, & w.p.\ \epsilon. \end{cases}$$

To find $I(X;Y)$, we compute $H(Y)$ and $H(Y|X)$, separately. For $H(Y)$, we first find the PMF of Y. Consider:

$$P_Y(0) = P_X(0) \cdot P_{Y|X}(0|0) + P_X(1) \cdot P_{Y|X}(0|1) = \frac{1}{2}(1-\epsilon) + \frac{1}{2}\epsilon = \frac{1}{2}.$$

Thus, $Y \sim \mathrm{Ber}(1/2)$, and so $H(Y) = H_b(1/2) = 1$.

For $H(Y|X)$, we have

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

By independence of X and Z, we have

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

where the last equality follows from the fact that entropy is invariant to bijection. Then,

$$H(Y|X) = \sum_{x\in\{0,1\}} p_X(x)H(X\oplus Z|X=x) = \sum_{x\in\{0,1\}} p_X(x)H(Z) = H(Z) = H_b(\epsilon).$$

This gives us that $I(X;Y) = 1 - H_b(\epsilon)$ for the BSC. Figure 3 depicts the mutual information as a function of ϵ.

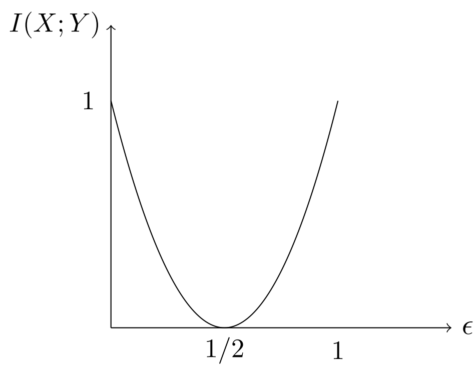

Figure 3: Mutual information of a BSC as a function of its parameter ϵ.

Notice that, in the "worst" case, $\epsilon = 1/2$ and we have $I(X;Y) = 0$, i.e., we cannot pass any information through the BSC.

- *<u>Bivariate Gaussian</u>: Let $(X,Y) \sim \mathcal{N}\left(\begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}1 & \rho\\ \rho & 1\end{bmatrix}\right)$. Recall that for a d-dimensional Gaussian we have $h\big(\mathcal{N}(\mu,\Sigma)\big) = \frac{1}{2}\log((2\pi e)^d \det K)$. Thus*

$$I(X;Y) = h(X) + h(Y) - h(X,Y) = \frac{1}{2}\log(2\pi e) + \frac{1}{2}\log(2\pi e) - \frac{1}{2}\log((2\pi e)^2(1-\rho^2)) = \frac{1}{2}\log\frac{1}{1-\rho^2}$$

Note that $I(X;Y) = \infty$ when $\rho = 1$. One could equivalently see that for $X = Y$, we have $I(X;Y) = \infty$ from Proposition 10.1. Moreover, $\rho = 0$ implies that X and Y are uncorrelated. For Gaussian random variables uncorrelation is equivalent to independence, which, in turn, is equivalent to $I(X;Y) = 0$.