ECE 5630: Information Theory for Data Transmission, Security and Machine Learning 03/10/2020

Lecture 11: Conditional Mutual Information and Letter Typical Sequences

Lecturer: Prof. Ziv Goldfeld

Scriber: Zhilu Zhang, *Net ID*: zz452

Assistant Editor: Kia Khezeli

11.1 Conditional Mutual Information

We next define the conditional mutual information between two random variables, X and Y, given a third variable Z. As a building block, we need the conditional mutual information given the event $\{Z = z\}$. Let $P_{XYZ} \in \mathcal{P}(\mathcal{X} \times \mathcal{Y} \times \mathcal{Z})$ and consider the induced conditional distribution $P_{XY|Z}(\cdot|z) \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$, for $z \in Z$. Denoting by $P_{X|Z}(\cdot|z)$ and $P_{Y|Z}(\cdot|z)$ the corresponding marginals, we set

$$I(X;Y|Z=z) := D_{\mathsf{KL}}\left(P_{XY|Z}(\cdot|z)\middle\|P_{X|Z} \otimes P_{Y|Z}(\cdot|z)\right).$$

Definition 11.1 (Conditional MI) *For $P_{XYZ} \in \mathcal{P}(\mathcal{X}\times\mathcal{Y}\times\mathcal{Z})$, the conditional mutual information between X and Y given Z is defined as*

$$I(X;Y|Z) := D_{\mathsf{KL}}\left(P_{XY|Z}\middle\|P_{X|Z} \otimes P_{Y|Z}|P_Z\right) = \mathbb{E}_{z\sim P_Z}\big[I(X;Y|Z=z)\big].$$

Remark 11.1

1. *$I(X;Y|Z)$ is a functional of P_{XYZ}, and not just the conditional probability law $P_{XY|Z}$.*
2. *It is straightforward to verify that*

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

In order to study the properties of conditional mutual information, we first review the related concept of Markov chains.

Definition 11.2 (Markov chain) *Let $(X,Y,Z) \sim P_{XYZ}$. We say that $X \to Y \to Z$ forms a Markov chain if*

$$P_{XYZ} = P_X P_{Y|X} P_{Z|Y}.$$

Example 11.1 *Let X, Y and Z be three mutually independent random variables. Clearly, $X \to Y \to Z$ forms a Markov chain.*

Example 11.2 *Define*

$$\begin{aligned} Y_1 &= X + Z_1, \\ Y_2 &= X + Z_1 + Z_2 = Y_1 + Z_2. \end{aligned}$$

then $X \to Y_1 \to Y_2$ forms a Markov chain (exercise).

Proposition 11.1 (Equivalent condition of Markov chain) *The following statements are equivalent.*

$$\begin{aligned} X \to Y \to Z &\iff P_{XYZ} = P_X P_{Y|X} P_{Z|Y} \\ &\iff P_{XZ|Y} = P_{X|Y} P_{Z|Y} \\ &\iff P_{Z|XY} = P_{Z|Y} \\ &\iff X \perp\!\!\!\perp Z|Y \\ &\iff Z \to Y \to X. \end{aligned}$$

We are now ready to study additional properties of mutual information and its conditional version.

Proposition 11.2 (More properties of MI) *Let $(X, Y, Z) \sim P_{XYZ}$. Then,*

1. *Non-negativity: $I(X;Y|Z) \geq 0$, with equality if and only if (iff) $X \to Z \to Y$.*
2. *Chain rule:*
   (a) *Small: $I(X,Y;Z) = I(X;Z) + I(Y;Z|X) = I(Y;Z) + I(X;Z|Y)$.*
   (b) *Full: $I(X_1, X_2, ..., X_n; Y) = I(X_1;Y) + \sum_{i=2}^n I(X_i;Y|X_{i-1}, X_{i-2}, ..., X_1)$.*
3. *Data processing inequality: If $X \to Y \to Z$, then $I(X;Y) \geq I(X;Z)$, with equality iff $X \to Z \to Y$.*
4. *If f is a bijection, then $I(X;Y) = I(X;f(Y))$.*
5. *Concavity/convexity: For $(X,Y) \sim P_{XY}$, denote $I(X;Y)$ as $I(P_X, P_{Y|X})$. Then,*
   (a) *For fixed $P_{Y|X}$, $P_X \to I(P_X, P_{Y|X})$ is concave.*
   (b) *Fox fixed P_X, $P_{Y|X} \to I(P_X, P_{Y|X})$ is convex.*

Proof:

1. Follows by definition.
2.
$$\begin{aligned} I(X_1, X_2, ..., X_n; Y) &= H(X_1, X_2, ..., X_n) - H(X_1, X_2, ..., X_n|Y) \\ &= \sum_{i=1}^n H(X_i|X_{i-1}, ..., X_1) - \sum_{i=1}^n H(X_i|X_{i-1}, ..., X_1, Y) \\ &= \sum_{i=1}^n I(X_i; Y|X_{i-1}, X_{i-2}, ..., X_1) \end{aligned}$$
3. First, observe that $I(X;Y,Z) = I(X;Y) + I(X;Z|Y) = I(X;Z) + I(X;Y|Z) \geq I(X;Z)$. Note that $I(X;Z|Y) = 0$ as $X \to Y \to Z$. By definition, $I(X;Y|Z) \geq 0$, and hence $I(X;Y) \geq I(X;Z)$, with equality if $X \to Z \to Y \iff I(X;Y|Z) = 0$.
4. For any deterministic function f, $X \to Y \to f(Y)$. By the *DPI*, $I(X;Y) \geq I(X;f(Y))$. But when f is a bijection, then $X \to f(Y) \to Y$ also holds. Applying the *DPI* again yields $I(X;f(Y)) \geq I(X;Y)$.
5. (a) It suffices to show that for any $\lambda \in [0,1]$, we have
$$I(\lambda P_X^{(0)} + (1-\lambda) P_X^{(1)}, P_{Y|X}) \geq \lambda I(P_X^{(0)}, P_{Y|X}) + (1-\lambda) I(P_X^{(1)}, P_{Y|X}).$$
   Let $\Theta \sim \mathsf{Ber}(\lambda)$. Define $P_{X|\Theta}(\cdot|0) = P_X^{(0)}$ and $P_{X|\Theta}(\cdot|1) = P_X^{(1)}$. By the law of total probability we have $P_X = \lambda P_X^{(0)} + (1-\lambda) P_X^{(1)}$ and by definition $\Theta \to X \to Y$. Thus,
$$I(X;Y) = I(X,\Theta;Y) = I(\Theta;Y) + I(X;Y|\Theta) \geq I(X;Y|\Theta).$$
   (b) Follows because $(P,Q) \to D_{\mathsf{KL}}(P \| Q)$ is convex in (P,Q) and that $I(P_X, P_{Y|X}) = D_{\mathsf{KL}}\left(P_{Y|X} \| P_Y | P_X\right)$.

■

11.2 Letter Typical Sequences

11.2.1 Introduction for binary alphabets

Let $\mathcal{X} = \{0, 1\}$, and consider its n-folds extension $\mathcal{X}^n$, i.e., $\mathcal{X}^n$ is the set of all binary sequences of length n. Element of $\mathcal{X}^n$ are denoted as $x^n := (x_1, ..., x_n) \in \mathcal{X}^n$. Clearly, there are $|\mathcal{X}^n| = |\mathcal{X}|^n = 2^n$ sequences in $\mathcal{X}^n$.

Now, let $P \in \mathcal{P}(\mathcal{X})$, i.e., $P = \mathsf{Ber}(\alpha)$, for some $\alpha \in (0, 1)$. Let $\{X_i\}_{i=1}^{\infty}$ be a sequence of random variables independently and identically distributed according to P. In other words, for all $n \in \mathbb{N}$, we have $(X_1, ..., X_n) \sim P^{\otimes n}$, where $P^{\otimes n}$ denotes the n-fold product measure induced by P, i.e., $P^{\otimes n}(x^n) := \prod_{i=1}^{n} P(\{x_i\})$.

Note that for any $x^n \in \mathcal{X}^n$, we have $P^{\otimes n}(x^n) \geq 0$. More specifically, if the sequence x^n contains $k \leq n$ ones (and $n - k$ zeros) then $P^{\otimes n}(x^n) = \alpha^k (1 - \alpha)^{n-k} > 0$. Despite the fact that all sequences have positive probability, clearly they are not all equiprobable. A natural question to ask is:

Question: What are the most probable sequences in $\mathcal{X}^n$ with respect to i.i.d. draws from $P = \mathsf{Ber}(\alpha)$?

Answer: We expect that a typical sequence will have roughly $n\alpha$ ones and $n(1 - \alpha)$ zeros.

Based on the above observation, the goal is to define a subset of $\mathcal{X}^n$ that is much smaller than $\mathcal{X}^n$ in cardinality, but that absorbs most of the probably mass (with respect to $P^{\otimes n}$). Calling this subset $\mathcal{T}^{(n)}(P)$ (for now), we would like it to satisfy

1. the set is "small", i.e., $|\mathcal{T}^{(n)}(P)| \ll |\mathcal{X}^n|$ in the sense that $\lim_{n\to\infty} \frac{|\mathcal{T}^{(n)}(P)|}{|\mathcal{X}^n|} = 0$.
2. the set "absorbs most of the probability", i.e., $\lim_{n\to\infty} P^{\otimes n}\left(\mathcal{T}^{(n)}(P)\right) = 1$.

To formalize this idea and define the desired set, we introduce the notion of empirical frequency.

Definition 11.3 (Empirical frequency) *Let $\mathcal{X}$ be discrete. For any $x^n \in \mathcal{X}^n$ and $a \in \mathcal{X}$, the number of occurrences of a in x^n is $N_{x^n}(a) := \sum_{i=1}^{n} \mathbb{1}_{\{x_i = a\}}$. The empirical frequency $\nu_{x^n}(a)$ of x^n is defined as*

$$\nu_{x^n}(a) := \frac{1}{n} N_{x^n}(a), \quad \forall a \in \mathcal{X}$$

Note that $\nu_{x^n}(a)$ is a valid PMF on $\mathcal{X}$. In the next lecture, we will define $\mathcal{T}^{(n)}(P)$ as the set that contains all sequences whose empirical frequency is roughly equal to the PMF of P.