Homework 3: Information Measures

CR07: Selected Topics Information Theory (Fall 2019). ENS de Lyon
samir.perlaza@inria.fr – Deadline: October 13, 2019 at 11:59 p.m.

1 Information Measures - Entropy

Let $X$ be a discrete random variable with probability mass function $P_X : \mathcal{X} \rightarrow [0, 1]$, with $\mathcal{X}$ a finite set. Let also $Y$ be a discrete random variable correlated to $X$. Assume that the joint probability mass function $P_{XY} : \mathcal{X} \times \mathcal{Y} \rightarrow [0, 1]$, with $\mathcal{Y}$ a countable set, factorizes as

$$P_{XY}(x, y) = P_{Y|X}(y)P_X(x),$$

for some given distribution $P_{Y|X}$.

Given a joint realization $(x, y) \in \mathcal{X} \times \mathcal{Y}$, the objective consists in estimating the value of $x$ using the value of $y$. Let the function function $g : \mathcal{Y} \rightarrow \mathcal{X}$ be the estimator and denoted by $\hat{X}$ the estimation of $X$. That is,

$$\hat{X} = g(Y).$$

Hence, the probability of error is

$$\Pr [g(Y) \neq X],$$

where the probability is with respect to the joint distribution $P_{XY}$ in (1).

1. Provide a lower bound on the error probability (3) for any estimator function $g$.

**HINT 1:** Prove that $H(X|\hat{X}) > H(X|Y)$.

**HINT 2:** Consider the Bernoulli random variable

$$E = \begin{cases} 1 & \text{if } g(Y) \neq X \\ 0 & \text{if } g(Y) = X \end{cases}$$

and use the properties of entropy on $H(E, X, \hat{X})$.

2. Evaluate your bound for the cases in which $H(Y|X) = H(Y)$ or $H(Y|X) = 0$. Comment on these cases.
2 Information Measures - Mutual Information

Consider a deep neural network (DNN) with \( n \) layers and \( k_i \) neurons in layer \( i \), with \( i \in \{1, 2, \ldots, n\} \). Let \( X \) be a discrete random variable representing the input of the network and let \( Y_i \) be the output at the \( i \)-th layer. Let \( Z \) be the output of the (DNN). Assume also the joint probability \( P_{X Y_1 Y_2 \ldots Y_n Z} : \mathcal{X} \times \mathcal{Y}_1^{k_1} \times \mathcal{Y}_2^{k_2} \times \cdots \times \mathcal{Y}_n^{k_n} \times Z \rightarrow [0, 1] \), with \( \mathcal{X}, \mathcal{Y}_1, \mathcal{Y}_2, \ldots, \mathcal{Y}_n, Z \) finite sets.

- Compare the following values \( I(X; Y_1), I(X; Y_2), \ldots, I(X; Y_n) \) and \( I(X; Z) \). Provide a formal mathematical argument.

- Some DNN are designed in such a way that the output of the \( i \)-th layer \( Y_i \) maximizes \( I(X; Y_i) \) while it minimizes \( I(Y_i; Y_{i+1}) \). Explain your intuitions behind such a rule. Provide a formal analysis (half a page maximum).

- Given the previous discussions what is your suggestion (in terms of information measures) for designing DNN in terms of number of layers and neurons per layers? (half a page maximum)

2