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

$$P_{XY}(x,y) = P_{Y|X}(y)P_X(x), \tag{1}$$

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 $g : \mathcal{Y} \to \mathcal{X}$ be the estimator and denoted by $\hat{X}$ the estimation of X. That is,

$$\hat{X} = g(Y). \tag{2}$$

Hence, the probability of error is

$$\Pr\left[g(Y) \neq X\right], \tag{3}$$

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

   $$E = \begin{cases} 1 & \text{if} \quad g(Y) \neq X \\ 0 & \text{if} \quad g(Y) = X \end{cases} \tag{4}$$

   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 $P_{X\boldsymbol{Y}_1\boldsymbol{Y}_2\cdots\boldsymbol{Y}_n,Z} : \mathcal{X} \times \mathcal{Y}_1^{k_1} \times \mathcal{Y}_2^{k_2} \times \cdots \times \mathcal{Y}_n^{k_n} \times \mathcal{Z} \to [0, 1]$, with $\mathcal{X}$, $\mathcal{Y}_1$, $\mathcal{Y}_2 \cdots$, $\mathcal{Y}_n$ and $\mathcal{Z}$ finite sets.

- Compare the following values $I(X; \boldsymbol{Y}_1)$, $I(X; \boldsymbol{Y}_2)$, $\cdots$, $I(X; \boldsymbol{Y}_n)$ 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 $\boldsymbol{Y}_i$ maximizes $I(X; \boldsymbol{Y}_i)$ while it minimizes $I(\boldsymbol{Y}_i; \boldsymbol{Y}_{i+1})$. 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)