Data Science / Analyse et Traitement de l'Information

TP7: Entropy and Detection Theory (70 pts).

1 Quantifiers of information

In information theory, the amount of information present in a random variable (r.v.) is quantified with entropy which is a function, only of the probability mass function. Intuitively, entropy indicated how surprising the outcome of an experiment, or equivalently the observed value of an r.v. would be.

Entropy The uncertainty or entropy of a discrete random variable (RV) X that takes value in the set $\mathcal{X}$ (also called alphabet $\mathcal{X}$) is defined as:

$$H(X) = \mathbb{E}_{P_X}\left[-\log_2 P_X(X)\right] = -\sum_{x\in\mathcal{X}} P_X(x)\log_2 P_X(x),$$

where $P_X(\cdot)$ denotes the probability mass function (PMF) of the RV X.

Joint Entropy The uncertainty or entropy of a discrete random vector $(X,Y)^T$ is defined as:

$$H(X,Y) = \mathbb{E}_{P_{X,Y}}\left[-\log_2 P_{X,Y}(X,Y)\right] = -\sum_{x\in\mathcal{X},y\in\mathcal{Y}} P_{X,Y}(x,y)\log_2 P_{X,Y}(x,y),$$

Conditional Entropy The conditional entropy or conditional uncertainty RV X given the random variable Y is defined as:

$$\begin{aligned} H(X\mid Y) &= -\sum_{x\in\mathcal{X},y\in\mathcal{Y}} P_{X,Y}(x,y)\log_2 P_{X|Y}(x\mid y) \\ &= \mathbb{E}_{P_{X,Y}}\left[-\log_2 P_{X|Y}(X\mid Y)\right]. \end{aligned}$$

Chain Rule for Entropy Let $X_1,...,X_n$ be n discrete RVs with a joint PMF $P_{X_1,...,X_n}$. Then

$$\begin{aligned} H(X_1,X_2,\cdots,X_n) &= H(X_1)+H(X_2\mid X_1)+\cdots+H(X_n\mid X_1,X_2,\cdots,X_{n-1}) \\ &= \sum_{k=1}^{n} H(X_k\mid X_1,X_2,\cdots,X_{k-1}) \end{aligned}$$

The above chain rule follows directly from the chain rule for PMFs:

$$P_{X_1,X_2,\cdots,X_n} = P_{X_1}\,P_{X_2|X_1}\,P_{X_3|X_1,X_2}\cdots P_{X_n|X_1,\ldots,X_{n-1}}.$$

Mutual information, on the other hand, specifies the amount of information between two r.v's. In other words, it indicates how much the knowledge of one r.v. helps in predicting the value of another r.v.

$$I(X,Y) = E_{p(x,y)}\left[\log\frac{p(x,y)}{p(x)p(y)}\right] = \sum_{x,y} p(x,y)\log\frac{p(x,y)}{p(x)p(y)}$$

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

Chain Rule for Mutual information

$$I(X;Y,Z) \quad = \quad I(X;Z) + I(X;Y|Z)$$

Using the definitions in the lecture notes, try to solve the below problem. Pay careful attention to the differences between marginal, conditional and joint versions of entropy and mutual information.

The task

U, V and W are three binary random variables. Their joint probability mass function is depicted as below:

U	V	W	$p_{U,V,W}(u,v,w)$
0	0	0	$\frac{1}{4}$
0	0	1	0
0	1	0	$\frac{1}{4}$
0	1	1	$\frac{1}{8}$
1	0	0	0
1	0	1	$\frac{1}{8}$
1	1	0	0
1	1	1	$\frac{1}{4}$

Calculate the information measures below:

1. $H(U)$, $H(V)$ and $H(W)$
2. $H(U|V)$, $H(V|U)$ and $H(W|U)$
3. $I(U;V)$, $I(U;W)$ and $I(U;V;W)$
4. $H(U,V,W)$

For each of the above items you could directly use the definition. However, because they are related to each other in many ways, you could calculate some of them and derive the rest by using their relations: chain rules for entropy and mutual information, the Venn diagrams.

2 Source coding

Source coding tries to express the output of a probabilistic source of information as compact as possible. In fact this problem is very much connected to the problem of data compression.

The idea is to capture the redundancies in a source of information based on the probabilities of the occurrences of different outcomes. Intuitively speaking, source coding tries to assign shorter codes for outcomes that are more probable and longer code words for less probable events. As a result, the overall code-length assigned to the whole sequence will be ideally as small as possible. In fact, this smallest possible value is closely related to the entropy.

Imagine a very simple case where we have a source X without memory that generates a sequence $X_1, X_2, \cdots, X_n$ of random variables. The probability distribution of all these X_i's are identical and independent from each other (i.i.d.). The source is binary and produces '1's with probability $\theta = 0.1$ and '0's with probability $1 - \theta = 0.9$.

Let's consider that the alphabet is constituted of all possible sequences $X_1, X_2, \cdots, X_n$ of length 5 (the size of this alphabet is then $2^5 = 32$).

- Generate a binary sequence with the length n = 10, 000 where probability of 1 is $p = 0.1$. Now you have a text of 2000 symbols where each symbol has the length 5.
- One very simple source coder is the Huffman code. Install the Python package `pip install dahuffman` and apply it to encode the sequence. Use `HuffmanCodec.from_frequencies` method (the input to this method are counts, not frequencies!).
- Calculate the entropy of this sequence analytically and compare it with the length of the output of the above encoder. Can you comment on this? Do they correspond? You may divide by the length of the sequence to give an interpretation.

Submission

Please archive your report and codes in "Prénom Nom.zip" (replace "Prénom" and "Nom" with your real name), and upload to "`Upload TP7: Entropy and Detection Theory.`" on `https://moodle.unige.ch` before **Monday, December 18 2023, 21:59 PM**. Note, that the assessment is mainly based on your report, which should include your answers to all questions and the experimental results. *Importance is given on the mathematical explanations of your works and your codes should be commented*
Please use this overleaf template.