Information and coding theory

Project 1

Corentin Merle s162662
Jad Akkawi s216641

Data Science and Engineering, bloc 1
Liège Université
School of Engineering
Academic year 2021-2022

1 Implementation

Question 1

The entropy $\mathcal{H}(\mathcal{X})$ of a random variable $\mathcal{X}$ form its probability distribution $P_{\mathcal{X}} = (p_1, p_2, .., p_n)$ is given by,

$$\mathcal{H}(\mathcal{X}) = -\sum_{i=1}^{n} P(\mathcal{X}_i) log(P(\mathcal{X}_i)) \tag{1}$$

The implementation of such a function which takes as input Px, the marginal probability distribution of the random variable $\mathcal{X}$ in a numpy array where $Px[i] = P(\mathcal{X} = \mathcal{X}_i)$, can simply be achieved thanks to a for loop with an enumerator which allows to apply the formula. Moreover a condition skips the computation when $Px[i] = 0$ in order to avoid computing the logarithm of a null value.

Intuitively, the entropy measures the level of information (expected value of self-information over all possible realisations) given by a random variable about the event it describes.

Question 2

The joint entropy $\mathcal{H}(\mathcal{X}, \mathcal{Y})$ of two discrete random variables $\mathcal{X}$ and $\mathcal{Y}$ from the joint probability distribution $P_{\mathcal{X},\mathcal{Y}}$ is given by,

$$\mathcal{H}(\mathcal{X}, \mathcal{Y}) = -\sum_{i=1}^{n}\sum_{j=1}^{m} P(\mathcal{X}_i \cap \mathcal{Y}_j) log(P(\mathcal{X}_i \cap \mathcal{Y}_j)) \tag{2}$$

The implementation of such a function which takes as input Pxy, the joint probability distribution of $\mathcal{X}$ and $\mathcal{Y}$ in a 2-D numpy array where $Pxy[i][j] = P(\mathcal{X} = \mathcal{X}_i, \mathcal{Y} = \mathcal{Y}_j)$, can simply be achieved thanks to a for loop with a N dimensions enumerator (`np.ndenumerate()`) which allows to apply the formula. Moreover a condition skips the computation when $Px[i][j] = 0$ in order to avoid computing the logarithm of a null value.

We can notice that equation 1 and 2 are very similar. Indeed they both consist in summing the negative product between an element and the logarithm of this element for each element of the input array, however the first is applied on the probability distribution array of a random variable while the last is applied on the joint probability distribution array of two random variables. One can imagine a third random variable $\mathcal{Z}$ that is the intersection (a tuple) of the $\mathcal{X}$ and $\mathcal{Y}$ and things become clear.

Question 3

The conditional entropy $\mathcal{H}(\mathcal{X}|\mathcal{Y})$ of a discrete random variable $\mathcal{X}$ given another discrete random variable $\mathcal{Y}$ from the joint probability distribution $P_{\mathcal{X},\mathcal{Y}}$ is given by,

$$\begin{aligned}\mathcal{H}(\mathcal{X}|\mathcal{Y}) &= -\sum_{i=1}^{n}\sum_{j=1}^{m} P(\mathcal{X}_i \cap \mathcal{Y}_j) log(P(\mathcal{X}_i|\mathcal{Y}_j)) && (3)\\ &= -\sum_{i=1}^{n}\sum_{j=1}^{m} P(\mathcal{X}_i \cap \mathcal{Y}_j) log\left(\frac{P(\mathcal{X}_i \cap \mathcal{Y}_j)}{P(\mathcal{Y}_j)}\right) && (4)\end{aligned}$$

As we can see in the equation above, the implementation of the conditional entropy function requires the joint probability distribution of $\mathcal{X}$ and $\mathcal{Y}$ and the conditional probability of $\mathcal{X}$ given $\mathcal{Y}$. Thus in order to implement the conditional entropy function which takes as input $P(\mathcal{X}_i \cap \mathcal{Y}_j)$ we have to compute $P(\mathcal{X}_i|\mathcal{Y}_j)$ which can be done thanks to the conditional probability definition given by :

$$P(\mathcal{X}|\mathcal{Y}) = \frac{P(\mathcal{X} \cap \mathcal{Y})}{P(\mathcal{Y})} \qquad \text{with} \qquad P(\mathcal{Y} = \mathcal{Y}_j) = \sum_{i=1}^{n} P(\mathcal{X}_i \cap \mathcal{Y}_j) \tag{5}$$

Regarding implementation, as in the previous point it can simply be achieved thanks to a for loop with a N dimensions enumerator (`np.ndenumerate()`) which allows to apply the formula. Moreover a condition skips the computation when $Px[i][j] = 0$ in order to avoid computing the logarithm of a null value.

An equivalent way of writing equation 4 using entropy and joint entropy is :

$$\mathcal{H}(\mathcal{X}|\mathcal{Y}) = \mathcal{H}(\mathcal{X},\mathcal{Y}) - \mathcal{H}(\mathcal{Y}) \tag{6}$$

Question 4

The mutual information $\mathcal{I}(\mathcal{X};\mathcal{Y})$ between two discrete random variables $\mathcal{X}$ and $\mathcal{Y}$ from the joint probability distribution $P_{\mathcal{X},\mathcal{Y}}$ is given by,

$$\mathcal{I}(\mathcal{X};\mathcal{Y}) = \sum_{i=1}^{n}\sum_{j=1}^{m} P(\mathcal{X}_i \cap \mathcal{Y}_j) log\left(\frac{P(\mathcal{X}_i \cap \mathcal{Y}_j)}{P(\mathcal{X}_i)P(\mathcal{Y}_j)}\right) \tag{7}$$

$$\text{with} \quad P(\mathcal{X} = \mathcal{X}_i) = \sum_{j=1}^{m} P(\mathcal{X}_i \cap \mathcal{Y}_j) \quad \text{and} \quad P(\mathcal{Y} = \mathcal{Y}_j) = \sum_{i=1}^{n} P(\mathcal{X}_i \cap \mathcal{Y}_j)$$

We also have learned that $\mathcal{I}(\mathcal{X};\mathcal{Y}) = \mathcal{H}(\mathcal{X}) - \mathcal{H}(\mathcal{X}|\mathcal{Y}) = \mathcal{H}(\mathcal{Y}) - \mathcal{H}(\mathcal{Y}|\mathcal{X})$ So if $\mathcal{X}$ is independent of $\mathcal{Y}$ we get $\mathcal{I}(\mathcal{X};\mathcal{Y}) = \mathcal{H}(\mathcal{X}) - \mathcal{H}(\mathcal{X}) = 0$. So the mutual information is a measure of how much information do $\mathcal{X}$ and $\mathcal{Y}$ bring about each other. So when the are independent they provide no information about each other.

Question 5

Let $\mathcal{X}$, $\mathcal{Y}$ and $\mathcal{Z}$ be three discrete random variables. The conditional joint entropy $\mathcal{H}(\mathcal{X},\mathcal{Y}|\mathcal{Z})$ is obtained by applying the chain rule and transforming as follows,

$$\begin{gathered} \mathcal{H}(\mathcal{X},\mathcal{Y},\mathcal{Z}) = \mathcal{H}(\mathcal{Z}) + \mathcal{H}(\mathcal{X},\mathcal{Y}|\mathcal{Z}) \\ \iff \\ \mathcal{H}(\mathcal{X},\mathcal{Y}|\mathcal{Z}) = \mathcal{H}(\mathcal{X},\mathcal{Y},\mathcal{Z}) - \mathcal{H}(\mathcal{Z}) \end{gathered}$$

And the conditional mutual information $\mathcal{I}(\mathcal{X};\mathcal{Y}|\mathcal{Z})$ is given by,

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

2 Weather forecasting

Question 6

	temperature	air_pressure	same_day_rain	next_day_rain	relative_humidity
entropy	1.511394	0.999997	1.475469	1.568656	0.999796
cardinality	3	2	3	3	2
log(card)	1.584963	1.000000	1.584963	1.584963	1.000000

	wind_direction	wind_speed	cloud_height	cloud_density	month
entropy	1.999551	1.584818	1.584622	1.584464	3.583413
cardinality	4	3	3	3	12
log(card)	2.000000	1.584963	1.584963	1.584963	3.584963

	day	daylight	lightning	air_quality
entropy	2.806399	0.998628	0.324968	0.535880
cardinality	7	2	3	3
log(card)	2.807355	1.000000	1.584963	1.584963

TABLE 1 – Entropy of each random variable

We can obeserve that the entropy is bounded above by the log of cardinality (entropy of equally probable values of a r.v). And most random variables here are equally probable except for lightning and air quality.(we have less uncertainty because we know that there is most likely no lightning and bad air quality)

Question 7

	temperature	air_pressure	same_day_rain	relative_humidity	wind_direction
conditional entropy	1.568101	0.939975	1.389486	1.301055	1.567815

	wind_speed	cloud_height	cloud_density	month	day
conditional entropy	1.567767	1.566763	1.56659	1.56488	1.567157

	daylight	lightning	air_quality
conditional entropy	1.568259	1.568233	1.567881

TABLE 2 – Conditional entropy of next day rain given each of the other random variables

In general, conditional (a posteriori) entropy is the entropy of a r.v after having acquired knowledge about some other r.v. We have from question 6 that $\mathcal{H}(next\_day\_rain) = 1.56 Shannon$ and since most conditional entropy values found in the table are $1.56 Shannon$ this means that for example $\mathcal{H}(next\_day\_rain|wind\_direction) = \mathcal{H}(next\_day\_rain)$ and hence these events are independent of next day's rain. Except for 3 r.v for example same day rain, because this information reduces next day rain's entropy.

Question 8

The mutual information between the variables wind speed and relative humidity is very small $0.0001243 Shannon$ which means that these 2 variables are independent and don't bring any information about each other.

On the other hand the mutual information between the variables month and temperature is considerable with value $0.5753 Shannon$ and this is an indication of the high dependency or correlation between these 2 events. Naturally, knowing the month we are in gives us information about what is the temperature and vice versa.

Question 9

	temperature	air_pressure	same_day_rain	relative_humidity	wind_direction
Mutual information	0.000555	0.628681	0.179171	0.267601	0.000841

	wind_speed	cloud_height	cloud_density	month	day
Mutual information	0.000889	0.001893	0.002066	0.003776	0.001499

	daylight	lightning	air_quality
Mutual information	0.000397	0.000424	0.000775

TABLE 3 – Mutual information between 'next day rain' and each of the other random variables

We would chose 'air pressure' since it has the highest mutual information with 'next day rain'. And i wouldn't make another choice since from question 7 we saw that the conditional entropy of 'next day rain' given 'air pressure' was the lowest with value 0.93Shannon which means that knowing the 'air pressure' is the most indicative of tomorrow's rain forecast and we should install the barometer.

Care to notice that mutual information and conditional information are complementary and together construct the entropy $\mathcal{I}(X, Y) + \mathcal{H}(X|Y) = \mathcal{H}(X)$.

Question 10

	temperature	air_pressure	same_day_rain	relative_humidity	wind_direction
Mutual information	0.000822	0.008075	0.158054	0.439192	0.000773

	wind_speed	cloud_height	cloud_density	month	day
Mutual information	0.000701	0.000474	0.00026	0.003855	0.000637

	daylight	lightning	air_quality
Mutual information	0.000347	0.000412	0.001078

TABLE 4 – Mutual information between 'next day rain' and each of the other random variables without the case of a 'dry' next day rain

If we consider only the samples with the volume of rain the next day corresponding to either deluge or drizzle we realise that the variable 'relative humidity' becomes our choice of measure and we install the Sling Psychometrer, because now its has the highest mutual information with 'next day rain'.

Question 11

In order to solve this question, we should compute the mutual information between (temperature, r.v) and (next day rain) for each of the rest r.v and hope to find a higher mutual information than the ones found in the above questions. We have : $\mathcal{I}(X, Y; Z) = \mathcal{I}(X; Z) + \mathcal{I}(Y; Z|X)$

We already have created the two functions required for this equation : mutual_information(Pxy) and cond_mutual_information(Pxyz)

But we have faced a challenge to create a joint distribution of 3 random variables in order to feed it to the functions. (pd.crosstab wasn't helpful for more than 2 random variables) So we are sorry for not providing a functioning implementation of this solution.

3 Playing Wordle with information theory-based strategy

Question 12

Considering $\mathcal{X}$ the random variable which corresponds to the letter in one field and knowing that we are considering a uniform distribution for each letter of the alphabet, the probability distribution is given by $P_{\mathcal{X}} = (p_A = \frac{1}{26}, p_B = \frac{1}{26}, ..., p_Z = \frac{1}{26})$. Thereby, by computing the entropy over this probability distribution we get an entropy $\mathcal{H}(\mathcal{X}) = 4.70$.

In order to compute the entropy of the game we can consider a random variable $\mathcal{Y}$ which corresponds to the five letters word. This random can take 5^{26} different values which are uniformly distributed. Thus the probability distribution is given by $P_{\mathcal{Y}} = (p_1 = \frac{1}{5^{26}}, p_2 = \frac{1}{5^{26}}, ..., p_{5^{26}} = \frac{1}{5^{26}})$ and the entropy is given by $\mathcal{H}(\mathcal{Y}) = 23.50$.

The relation between the entropy of one field $\mathcal{H}(\mathcal{X})$ and the entropy of the game $\mathcal{H}(\mathcal{Y})$ is

$$\mathcal{H}(\mathcal{Y}) = 5 \times \mathcal{H}(\mathcal{X})$$

This result can be theoretically justify by the fact that random variables of each field are independent one from an other (no mutual information) thus we have,

$$\begin{aligned}\mathcal{H}(\mathcal{Y}) &= \mathcal{H}(\mathcal{X}_1, \mathcal{X}_2, \mathcal{X}_3, \mathcal{X}_4, \mathcal{X}_5) \\ &= \mathcal{H}(\mathcal{X}_1) + \mathcal{H}(\mathcal{X}_2) + \mathcal{H}(\mathcal{X}_3) + \mathcal{H}(\mathcal{X}_4) + \mathcal{H}(\mathcal{X}_5)\end{aligned}$$

Question 13

The result of the "TABLE" proposition word brings information. Firstly we know that the second field contains the letter A, thus the entropy of this field is null. Secondly we know that letters B, E, L and T are not in the word, thereby we can remove these letters from the set of possible letters. Thus the probability distribution of random variable of field 1, 3, 4 and 5 are given by $P_{\mathcal{X}_i} = (p_A = \frac{1}{22}, p_B = 0, ..., p_Z = \frac{1}{22}), \forall i = 1, 3, 4, 5$ Entropy values of each field are given in TABLE 5 below.

	$i = 1$	$i = 2$	$i = 3$	$i = 4$	$i = 5$
$\mathcal{H}(\mathcal{X}_i)$	4.46	0	4.46	4.46	4.46

TABLE 5

Values of the random variable $\mathcal{Y}$ which corresponds to the five letters word are now restricted to 4^{22} different values because of the discussion above about information brings by the proposition. The entropy of $\mathcal{Y}$ is now equals to $\mathcal{H}(\mathcal{Y}) = 17.83$ which correspond to an information gain of 5.67 $bits$ due to the proposition "TABLE".

Question 14

Knowing the new proposition "ROUGH" the probability distribution of random variable of field 1, 3 and 5 are given by

$$\begin{aligned}P_{\mathcal{X}_i} =(& p_A = \frac{2/3}{17}, p_B = 0, p_C = \frac{2/3}{17}, p_D = \frac{2/3}{17}, p_E = 0, p_F = \frac{2/3}{17}, p_G = 1/3, p_H = 0, p_I = \frac{2/3}{17}, p_J = \frac{2/3}{17}, \\ & p_K = \frac{2/3}{17}, p_L = 0, p_M = \frac{2/3}{17}, p_N = \frac{2/3}{17}, p_O = 0, p_P = \frac{2/3}{17}, p_Q = \frac{2/3}{17}, p_R = 0, p_S = \frac{2/3}{17}, p_T = 0, \\ & p_U = 0, p_V = \frac{2/3}{17}, p_W = \frac{2/3}{17}, p_X = \frac{2/3}{17}, p_Y = \frac{2/3}{17}, p_Z = \frac{1}{17}), \forall i = 1, 3, 5\end{aligned}$$

the probability distribution of random variable of field 4 is given by

$$
\begin{aligned}
P_{\mathcal{X}_4} =& (p_A = \frac{1}{17}, p_B = 0, p_C = \frac{1}{17}, p_D = \frac{1}{17}, p_E = 0, p_F = \frac{1}{17}, p_G = 0, p_H = 0, p_I = \frac{1}{17}, p_J = \frac{1}{17},\\
& p_K = \frac{1}{17}, p_L = 0, p_M = \frac{1}{17}, p_N = \frac{1}{17}, p_O = 0, p_P = \frac{1}{17}, p_Q = \frac{1}{17}, p_R = 0, p_S = \frac{1}{17}, p_T = 0,\\
& p_U = 0, p_V = \frac{1}{17}, p_W = \frac{1}{17}, p_X = \frac{1}{17}, p_Y = \frac{1}{17}, p_Z = \frac{1}{17})
\end{aligned}
$$

Thereby we can compute entropy values of each field which are reported in the TABLE 6 below,

	$i = 1$	$i = 2$	$i = 3$	$i = 4$	$i = 5$
$\mathcal{H}(\mathcal{X}_i)$	3.64	0	3.64	4.08	3.64

TABLE 6

By opposition to the previous question the sum of the entropy of each field is no longer equals to the entropy of the game. This is due to the mutual information add by the orange field containing the letter G.

Question 15

The entropy of the real game is much less than that of the simplified version. Firstly because the number of possible words (cardinality) have decreased from 5^{26} to 2000 words and entropy is bounded by the log of cardinality. Secondly we know that English distribution of letters can lead to a lot of mutual information (dependencies between fields). This means that at each guess we will

Question 16

In order to solve the game in an optimal way we can imagine strategies based on information theory. The strategy will be to take advantage of the information about the English language such as the occurrence of letters, or to go deeper the occurrence of letters patterns and the information that the game provides at each step. The main goal of such a strategy will be to take decisions that minimize the entropy at each step.