ELEN-0060: Information and Coding Theory

Project 1 - Information Measures

Maxime Goffart
180521

Olivier Joris
182113

Academic year 2021 - 2022

1 Implementation

1.1 Function `entropy` (question 1)

We are using this mathematical formula:

$$\mathcal{H}(\mathcal{X}) = -\sum_{\mathcal{X}_i} P(\mathcal{X}_i) \log_2 P(\mathcal{X}_i)$$

Our implementation is first filtering the negative $P(\mathcal{X}_i)$ because the $\log(x)$ function is only defined for $x \in]0, +\infty[$. Then, it is performing the sum over all $\mathcal{X}_i$ using the `sum` method of the numpy library and return the result according to the above mathematical formula.

Intuitively, the entropy is measuring the average amount of information provided by a random variable. It is based on the fact that less probable event have less chance to occur but when they occur, they provide a lot of information. In the opposite direction, more probable event provide much less information when they occur.

1.2 Function `joint_entropy` (question 2)

We are using this mathematical formula:

$$\mathcal{H}(\mathcal{X}, \mathcal{Y}) = -\sum_{\mathcal{X}_i, \mathcal{Y}_j} P(\mathcal{X}_i \cap \mathcal{Y}_j) \log_2 P(\mathcal{X}_i \cap \mathcal{Y}_j)$$

We notice that this joint entropy is equal to the previously defined entropy taking as argument the joint distribution of the two random variables instead of the marginal distribution of the single random variable.

It is why we are implementing this function by simply reshaping the joint probability distribution to a one dimensional array. In this way, we can just call the `entropy` function with this reshaped array as argument.

1.3 Function `conditional_entropy` (question 3)

We are using this mathematical formula which has been demonstrated in the theoretical course:

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

Our implementation consists thus in computing the marginal probability distribution $P_{\mathcal{Y}}$ by marginalizing the joint probability distribution $P_{\mathcal{X}\mathcal{Y}}$ over $\mathcal{X}$. This is done using the `sum` function of the numpy library using `axis = 0` as argument. Then, we just compute the result using the above mathematical formula and the previous defined functions.

An equivalent way to compute this property is to directly use the mathematical formula of the entropy:

$$\mathcal{H}(\mathcal{X}|\mathcal{Y}) = -\sum_{\mathcal{X}_i, \mathcal{Y}_j} P(\mathcal{X}_i \cap \mathcal{Y}_j) \log_2 P(\mathcal{X}_i|\mathcal{Y}_j)$$

1.4 Function mutual_information (question 4)

We are using this mathematical formula which has been demonstrated in the theoretical course:

$$\mathcal{I}(\mathcal{X};\mathcal{Y}) = \mathcal{H}(\mathcal{X}) - \mathcal{H}(\mathcal{X}|\mathcal{Y})$$

Our implementation consists thus in computing the marginal probability distribution $P_{\mathcal{X}}$ by marginalizing the joint probability distribution P_{xy} over $\mathcal{Y}$. This is done using the `sum` function of the numpy library using `axis = 1` as argument. Then, we just compute the result using the above mathematical formula and the previous defined functions.

The mutual information between two discrete random variables measures the level of dependence between these random variables. Especially, if $\mathcal{I}(\mathcal{X};\mathcal{Y}) = 0$, $\mathcal{X}$ and $\mathcal{Y}$ are independent.

1.5 Functions cond_joint_entropy and cond_mutual_information (question 5)

We are using these mathematical formulas which have been demonstrated in the theoretical course:

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

$$\mathcal{I}(\mathcal{X};\mathcal{Y}|\mathcal{Z}) = \mathcal{H}(\mathcal{X}|\mathcal{Z}) + \mathcal{H}(\mathcal{Y},\mathcal{Z}) - \mathcal{H}(\mathcal{X},\mathcal{Y},\mathcal{Z})$$

Our implementations consist thus in computing the right marginal probability distributions by marginalizing the joint probability distribution P_{xyz}. This is done as in the previous sections using the `axis` argument of the `sum` function of the numpy library. Then, we just compute the result using the above mathematical formulas and the previous defined functions.

2 Weather forecasting

2.1 Entropy and cardinality of each variable (question 6)

The corresponding entropy and cardinality of each variable can be observed in the table 1.

Random variable $\mathcal{X}$	**Entropy $\mathcal{H}(\mathcal{X})$**	**Cardinality**
temperature	1.5113935187	4
air_pressure	0.9999971146	2
same_day_rain	1.4754687972	3
next_day_rain	1.5686562064	3
relative_humidity	0.9997963973	2
wind_direction	1.9995507337	4
wind_speed	1.5848180055	3
cloud_height	1.5846220676	3
cloud_density	1.5844638107	3
month	3.5834131971	12
day	2.8063989677	7
daylight	0.9986283124	2
lightning	0.3249678888	3
air_quality	0.5358803476	3

Table 1: Entropy $\mathcal{H}(\mathcal{X})$ and cardinality of each random variable $\mathcal{X}$.

We see in this table that the higher the cardinality, the higher the entropy. This can be theoretically justified by the fact that the entropy of a instrument corresponds to the amount of information gained when its value is known. It is why with a larger cardinality the entropy is larger. Indeed, the instrument can take more values than with lower cardinalities which implies that the probability of getting a specific value for a variable is smaller thus the entropy is higher.

2.2 Conditional entropy of `next_day_rain` given each of the other variables (question 7)

The corresponding conditional entropy of each variable with `next_day_rain` can be observed in the table 2.

Random variable $\mathcal{Y}$	**Conditional entropy $\mathcal{H}($`next_day_rain`$\|\mathcal{Y})$**
temperature	1.5681010090
air_pressure	0.9399751579
same_day_rain	1.3894855511
relative_humidity	1.3010552471
wind_direction	1.5678153355
wind_speed	1.5677670878
cloud_height	1.5667630290
cloud_density	1.5665898847
month	1.5648797492
day	1.5671568099
daylight	1.5682591877
lightning	1.5682325749
air_quality	1.5678811342

Table 2: Conditional entropy $\mathcal{H}($`next_day_rain`$|\mathcal{Y})$ with each random variable $\mathcal{Y}$.

(a) When the conditioning variable is `wind_direction`, the conditional entropy is nearly equal to the entropy of the `next_day_rain` variable without any conditioning. It can be explained by the fact that knowing the `wind_direction` variable does not really provide any additionnal information on the `next_day_rain` variable which seems intuitively logical.

(b) When the conditioning variable is `same_day_rain`, the conditional entropy is lower than the entropy of the `next_day_rain` variable without any conditioning. It can be explained by the fact that knowing the `same_day_rain` variable provides additionnal information on the `next_day_rain` which seems intuitively logical. The provided amount of information corresponds here to 1.5686562064 - 1.3894855511 = 0.1791706553 bits.

2.3 Mutual information between two variables (question 8)

We can deduce that the variables `relative_humidity` and `wind_speed` are likely to be independent because their mutual information is nearly equal to 0 (it is equal to 0.0001243960).

We can deduce that the variables `month` and `temperature` are dependent because their mutual information is > 0 (here it is equal to 0.5753467937).

2.4 Choose a single instrument (question 9)

Based on the mutual information, the variable kept would be air_pressure because the amount of mutual information between this variable and next_day_rain is the highest compared to all the other variables (it is equal to 0.6286810485).

Based on the conditional entropy, the variable kept would be air_pressure because the conditional entropy of next_day_rain given this variable is the lowest compared to all the other variables (it is equal to 0.9399751579). The variable kept does not change from the one chosen using the mutual information.

2.5 Deletion of the dry sample of the next_day_rain variable from the dataset (question 10)

Based on the mutual information, the variable kept would be relative_humidity because the amount of mutual information between this variable and next_day_rain is the highest compared to all the other variables (it is equal to 0.4391920975).

Based on the conditional entropy, the variable kept would be relative_humidity because the conditional entropy of next_day_rain given this variable is the lowest compared to all the other variables (it is equal to 0.5601193454). The variable kept does not change from the one chosen using the mutual information.

If we delete the dry sample of the next_day_rain variable, the variable kept becomes the relative_humidity.

2.6 Instrument kept if we have a thermometer for free (question 11)

Based on the conditional mutual information, the variable kept would be air_pressure because the conditional mutual information between this variable, next_day_rain and this variable knowing temperature is the highest compared to all the other variables (it is equal to 0.6294687900).

Based on the conditional joint entropy, the variable kept would be air_pressure because the conditional joint entropy of next_day_rain knowing temperature and this variable is the lowest compared to all the other variables (it is equal to 0.9386322190). This has been computed using this formula:

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

This does not change the kept variable according to the question 9.

3 Playing with information theory-based strategy

3.1 Entropy of one field and one word (question 12)

One field can be in 26 states corresponding to the 26 letter of the alphabet, we have $\mathcal{X} \in \{\texttt{a}, \texttt{b}, \ldots, \texttt{z}\}$. We have that $P(\mathcal{X} = \texttt{a}) = P(\mathcal{X} = \texttt{b}) = \ldots = P(\mathcal{X} = \texttt{z}) = \frac{1}{26}$ because the letters are equiprobable in this simplified version of the game.

Now that we have the probability values, we can compute the entropy of each field:

$$\mathcal{H}(\mathcal{X}) = \sum_{i=1}^{26} \frac{1}{26} \log_2{(26)} = \log_2{(26)} = 4.70043971814 \; bits$$

For the following questions, let $\mathcal{X}_i$ be a random variable representing the state of of the i^{th} field of the word. The entropy of the whole game (the 5 letters combined) is equal to:

$$\mathcal{H}(\mathcal{X}_1, \mathcal{X}_2, \mathcal{X}_3, \mathcal{X}_4, \mathcal{X}_5) = \sum_{i=1}^{26^5} \frac{\log_2(26^5)}{26^5} = 23.5021985907 \; bits$$

Thus, the entropy of the word is equal to the product of the entropy of one field and the world length because $\mathcal{X}_1, \mathcal{X}_2, \mathcal{X}_3, \mathcal{X}_4$, and $\mathcal{X}_5$ are independent random variables and the entropy is additive. Indeed, here we have that $\mathcal{H}(\mathcal{X}_1, \mathcal{X}_2, \mathcal{X}_3, \mathcal{X}_4, \mathcal{X}_5) = 5 * \mathcal{H}(\mathcal{X}_i)$

3.2 Entropy and information brought by a first guess using the word TABLE (question 13)

We have that the entropy of a random variable $\mathcal{X}$ representing a gray field of the word TABLE is equal to:

$$\mathcal{H}(\mathcal{X}) = \sum_{i=1}^{22} \frac{1}{22} \log_2{(22)} = \log_2{(22)} = 4.45943161864 \; bits$$

Because we know that 4 letters cannot be part of the word[1].

The entropy of the green field of the word TABLE is equal to 0 because there is no more uncertainty about it.

The entropy of the game is equal to:

$$\mathcal{H}(\mathcal{X}_1, \mathcal{X}_2, \mathcal{X}_3, \mathcal{X}_4, \mathcal{X}_5) = \sum_{i=1}^{22^4} \frac{\log_2(22^4)}{22^4} = \mathcal{H}(\mathcal{X}_1) + \mathcal{H}(\mathcal{X}_3) + \mathcal{H}(\mathcal{X}_4) + \mathcal{H}(\mathcal{X}_5) = 17.8377264745 \; bits$$

The information that this guess has brought is equal to:

$$I(\texttt{TABLE}) = \log_2{(26)} + 4\,(\log_2{(26)} - \log_2(22)) = 5.66447211616 \; bits$$

Because we know that there is an A in the word and we also know that four letters are not part of the word. It also corresponds to the difference between the initial entropy and the entropy after the first guess of the game.

[1] T, B, L, and E.

3.3 Entropy after a second guess using the word ROUGH (question 14)

We have that the entropy of a random variable $\mathcal{X}$ representing a gray field[2] of the word ROUGH at this stage of the game is equal to:

$$\mathcal{H}(\mathcal{X}) = \sum_{i=1}^{18} \frac{1}{18} \log_2(18) = 4.16992500144 \; bits$$

Because we know that there are 8 letters that can not be part of the word according to the previous and actual guesses.

We have that the entropy of a random variable $\mathcal{Y}$ representing the orange field[3] of the word ROUGH at this stage of the game is equal to:

$$\mathcal{H}(\mathcal{Y}) = \sum_{i=1}^{17} \frac{1}{17} \log_2(17) = 4.08746284125 \; bits$$

Because we know that there are 9 letters that can not be part of the word according to the previous and actual guesses.

The entropy of the game at this stage is equal to:

$$\mathcal{H}(\mathcal{X}_1, \mathcal{X}_2, \mathcal{X}_3, \mathcal{X}_4, \mathcal{X}_5) = \frac{3}{4} \sum_{i=1}^{18^4} \frac{1}{18^4} \log_2(18^4) + \frac{1}{4} \log_2(1) + \sum_{i=1}^{17} \frac{1}{17} \log_2(17) = 16.5972378456 \; bits$$

Because we have $\frac{1}{4}$ that the G is the right letter of the field and if it is, there is no more uncertainty about the cell. If G is not the right letter of the field[4], we have $\frac{1}{18}$ chance to have the right letter taking into account the gray fields of the actual and previous guesses.

The entropy of the game and each field are linked by this equation:

$$\mathcal{H}(\mathcal{X}_1, \mathcal{X}_2, \mathcal{X}_3, \mathcal{X}_4, \mathcal{X}_5) < 4 * H(\mathcal{X}) + \mathcal{H}(\mathcal{Y})$$

It is because $\mathcal{X}_1, \mathcal{X}_2, \mathcal{X}_3$, and $\mathcal{X}_5$, are dependent of $\mathcal{X}_4$. Indeed, because the fourth letter is orange, we have additionnal information about the probability of the other variables.

3.4 Entropy of the real game compared to the simplified version (question 15)

The entropy of the real game will be lower than this simplified version because the set of possible words is only composed of 2000 words while there were 26^5 words in the simplified version. Thus, the probability that a given word is the right one will be higher and the entropy lower than in the simplified version. The probability of each letter are not yet equiprobable because they follow the English distribution thus a not frequent letter like Z will give a lot of Information about the word if present unlike vowels like A and E.

[2] A field containing a R, O, U, or H.

[3] The field containing the letter G.

[4] It has a probability of $1 - \frac{1}{4} = \frac{3}{4}$

However, in the simplified version we can make guesses that are not existing words which allows to have a maximal entropy reduction at each guess unlike in the simplified version of the game.

Finally, in the real game, if we repeat multiple time a letter, we have different result than in this simplified version. For example, if we propose the word `ABCDD` and the word that we gave to guess is `ABCDE`, the real game will color the last letter in gray while the simplified version will color it in orange according to its rules. This difference makes that entropy might not represents the true uncertainty about the game in the simplified version while it is more accurated in the real game.

3.5 Approach based on information theory to solve the real game in a minimum number of guesses (question 16)

To solve the game at each step, we would consider the entropies of all the possible word that we can propose in order to make a guess that maximize the entropy reduction for the next guess. This can be done by computing a tree and follow the path maximizing this entropy reduction.