Faculty of Applied Sciences

Information and coding theory

ELEN0060-2

Project 1:
Information measures

Authors :
Delcour Florian - s181063
Makedonsky Aliocha - s171197

Instructors :
Wehenkel Louis

March 2022

1 Implementation

1.1 Entropy

Having the probability distribution of the random variable $\mathcal{X}$, we can compute its entropy $\mathcal{H}(\mathcal{X})$ following the theoretical definition :

$$\mathcal{H}(\mathcal{X}) = -\sum_{i=1}^{n} P(X_i) \cdot \log_2 P(X_i)$$

We use the logarithm in base 2, though it only multiplies the entropy by a constant if we take an other base, but the observations would be the same. The base 2 is used because the Shannon entropy unit is in bits. To implement this formula, we simply computes the matrix-vector multiplication of $P(X_i)$ and $\log_2 P(X_i)$, without considering zero probabilities.
Intuitively, the Shannon entropy measures the quantity of information provided by a random variable. This amount of information is related to the unpredictability of the random variable. Indeed, a random variable for which the outcome is always the same does not tell us lot information and thus the entropy is equal to 0 as this random variable is totally predictable.

Analyzing the theoretical definition, $P(X_i) \in [0\ ;1]$, $\log_2 P(X_i)$ will always be negative and will decrease quicker than $P(X_i)$ when $P(X_i)$ decreases. So smaller is $P(X_i)$, higher is $|P(X_i) * \log_2 P(X_i)|$, but the product is always negative, that's why we take the opposite of the sum of products to compute the entropy, so a big entropy represents a big uncertainty since the $P(X_i)$ are smaller.

1.2 Joint entropy

Having the joint probability distribution $P_{\mathcal{X},\mathcal{Y}}$ of the random variables $\mathcal{X}$ and $\mathcal{Y}$, we can compute the joint entropy $\mathcal{H}(\mathcal{X},\mathcal{Y})$. It was defined in the lectures as :

$$\mathcal{H}(\mathcal{X},\mathcal{Y}) = -\sum_{i=1}^{n}\sum_{j=1}^{m} P(X_i \cap Y_j) \log_2 P(X_i \cap Y_j)$$

Having the property $P(X,Y) = P(X)P(Y|X) = P(Y)P(X|Y) = P(X \cap Y)$, we can rewrite the joint entropy as :

$$\mathcal{H}(\mathcal{X},\mathcal{Y}) = -\sum_{i=1}^{n}\sum_{j=1}^{m} P(X_i, Y_j) \log_2 P(X_i, Y_j)$$

and we have $P(X_i, Y_j)$ as input. To implement this, we loop over the elements of the inputs which are the joint probabilities present in the formula, again without considering zero joint probabilities.
The expression of the entropy and the joint entropy is exactly the same, except that we replace the marginal probability by the joint probability. So the joint probability is the entropy of the random variable $\mathcal{Z} = (\mathcal{X},\mathcal{Y})$. The only difference is in the implementation where the input is 1D-vector for the entropy and 2D-matrix for joint entropy.

1.3 Conditional entropy

The condition entropy $\mathcal{H}(\mathcal{X}|\mathcal{Y})$ was defined in the lectures as

$$\mathcal{H}(\mathcal{X}|\mathcal{Y}) = -\sum_{i=1}^{n}\sum_{j=1}^{m} P(X_i \cap Y_j)\log_2 P(X_i|Y_j)$$

However we don't have access to $P(X|Y)$ but only to $P(X,Y)$, so using the property we talked about in section 1.2, we rewrite $P(X|Y) = \dfrac{P(X,Y)}{P(Y)}$ where $P(Y)$ can be computed as follows : $P(Y_j) = \sum_{i=1}^{n} P(X_i, Y_j) \quad ,\forall j = 1, ..., m$

We get

$$\mathcal{H}(\mathcal{X}|\mathcal{Y}) = -\sum_{i=1}^{n}\sum_{j=1}^{m} P(X_i, Y_j)\log_2 \frac{P(X_i, Y_j)}{P(Y_j)}$$

where each term of this expression is known. To implement this, we loop over the elements of the inputs which are the joint probabilities, without considering zero joint probabilities of $(\mathcal{X}, \mathcal{Y})$ and zero marginal probabilities of the variable $\mathcal{Y}$.

An alternative way to compute this quantity is given by :

$$\begin{aligned}
\mathcal{H}(\mathcal{X}|\mathcal{Y}) &= -\sum_{i=1}^{n}\sum_{j=1}^{m} P\left(X_i, Y_j\right)\log_2 \frac{P\left(X_i, Y_j\right)}{P\left(Y_j\right)} \\
&= -\sum_{i=1}^{n}\sum_{j=1}^{m} P\left(X_i, Y_j\right)\log_2 P\left(X_i, Y_j\right) + \sum_{i=1}^{n}\sum_{j=1}^{m} P\left(X_i, Y_j\right)\log_2 P\left(Y_j\right) \\
&= -\sum_{i=1}^{n}\sum_{j=1}^{m} P\left(X_i, Y_j\right)\log_2 P\left(X_i, Y_j\right) + \sum_{j=1}^{m} P\left(Y_j\right)\log_2 P\left(Y_j\right) \\
&= \mathcal{H}(\mathcal{X}, \mathcal{Y}) - \mathcal{H}(\mathcal{Y})
\end{aligned}$$

where these last expressions were both defined in the previous sections.

1.4 Mutual information

The mutual information was defined in the lectures as

$$\mathcal{I}(\mathcal{X}; \mathcal{Y}) = +\sum_{i=1}^{n}\sum_{j=1}^{m} P(X_i \cap Y_j)\log_2 \frac{P(X_i \cap Y_j)}{P(X_i)P(Y_j)}$$

As we said before, $P(X \cap Y) = P(X, Y)$, and we can compute $P(X_i) = \sum_{j=1}^{m} P(X_i, Y_j)$ and $P(Y_j) = \sum_{i=1}^{n} P(X_i, Y_j)$. Since we have $P(X, Y)$, we have all the terms that we need, and we can compute

$$\mathcal{I}(\mathcal{X}; \mathcal{Y}) = +\sum_{i=1}^{n}\sum_{j=1}^{m} P(X_i, Y_j)\log_2 \frac{P(X_i, Y_j)}{P(X_i)P(Y_j)}$$

To implement that, we loop over the elements of the inputs which are the joint probabilities, without considering zero joint probabilities of $(\mathcal{X}, \mathcal{Y})$ and zero marginal probabilities

of the variables $\mathcal{X}$ and $\mathcal{Y}$.

Mutual information quantifies the amount of information we learned about $\mathcal{Y}$ knowing the value of $\mathcal{X}$. The greater $I(\mathcal{X};\mathcal{Y})$, the more $\mathcal{X}$ and $\mathcal{Y}$ are correlated. If $\mathcal{X}$ and $\mathcal{Y}$ are independent, $P(X_i, Y_j) = P(X_i)P(Y_j)$ and then $I(\mathcal{X};\mathcal{Y}) = 0$.

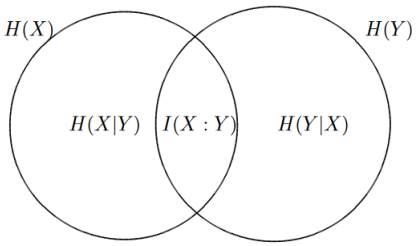

Figure 1: Link between entropy, conditional entropy and mutual information

1.5 Conditional joint entropy and conditional mutual information

We can express the conditional joint entropy $\mathcal{H}(\mathcal{X},\mathcal{Y}|\mathcal{Z})$ as the conditional entropy $\mathcal{H}(\mathcal{W}|\mathcal{Z})$ where $\mathcal{W} = (\mathcal{X},\mathcal{Y})$. Thus we have

$$\mathcal{H}(\mathcal{W}|\mathcal{Z}) = -\sum_{i=1}^{n}\sum_{j=1}^{m} P(W_i, Z_j)\log_2 \frac{P(W_i, Z_j)}{P(Z_j)}$$

and when re-expressing it with $\mathcal{X},\mathcal{Y}$ we have

$$\mathcal{H}(\mathcal{X},\mathcal{Y}|\mathcal{Z}) = -\sum_{i=1}^{n}\sum_{j=1}^{m}\sum_{k=1}^{l} P(X_i, Y_j, Z_k)\log_2 \frac{P(X_i, Y_j, Z_k)}{P(Z_k)}$$

Once again we can compute $P(Z_k)$ as $P(Z_k) = \sum_{i=1}^{n}\sum_{j=1}^{m} P(X_i, Y_j, Z_k)$. To implement that, we loop over the elements of the inputs which are the joint probabilities, without considering zero joint probabilities of $(\mathcal{X},\mathcal{Y},\mathcal{Z})$ and zero marginal probabilities of the variable $\mathcal{Z}$.

About the conditional mutual information, it can be expressed as

$$\begin{aligned}
\mathcal{I}(\mathcal{X};\mathcal{Y}|\mathcal{Z}) &= \sum_{i=1}^{n}\sum_{j=1}^{m}\sum_{k=1}^{l} P(X_i, Y_j, Z_k)\log_2 \frac{P(X_i, Y_j \mid Z_k)}{P(X_i \mid Z_k)P(Y_j \mid Z_k)} \\
&= \sum_{i=1}^{n}\sum_{j=1}^{m}\sum_{k=1}^{l} P\left(X_i, Y_j, Z_k\right)\log_2 \frac{P(X_i, Y_j, Z_k)P(Z_k)P(Z_k)}{P(Z_k)P(X_i, Z_k)P(Y_j, Z_k)} \\
&= -H(\mathcal{X},\mathcal{Y},\mathcal{Z}) + \sum_{i=1}^{n}\sum_{j=1}^{m}\sum_{k=1}^{l} P(X_i, Y_j, Z_k)\cdot\Big(\log_2 P(Z_k) - \log_2 P(X_i, Z_k) - \log_2 P(Y_j, Z_k)\Big) \\
&= -H(\mathcal{X},\mathcal{Y},\mathcal{Z}) - H(\mathcal{Z}) + H(\mathcal{X},\mathcal{Z}) + H(\mathcal{Y},\mathcal{Z})
\end{aligned}$$

In addition, we know from 1.3 that $\mathcal{H}(\mathcal{X},\mathcal{Y}) = \mathcal{H}(\mathcal{X} \mid \mathcal{Y}) + \mathcal{H}(\mathcal{Y})$, thus we have :

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

To implement this, we can use the functions defined previously.

2 Weather Forecasting

2.1 Question 6

Variable	Entropy	Cardinality
Temperature	1.511	3
Air pressure	1.0	2
Same day rain	1.475	3
Next day rain	1.569	3
Relative humidity	1.0	2
Wind direction	2.0	4
Wind speed	1.585	3
Cloud height	1.584	3
Cloud density	1.584	3
Month	3.583	12
Day	2.806	7
Daylight	0.999	2
Lightning	0.325	3
Air quality	0.536	3

Table 1: Entropy and cardinality of each variable

We see that the entropy are globally proportional to their cardinality, higher the cardinality, bigger the entropy. It is logic since if there is more possible values for a variable, the probability of each of these values will be smaller, so the uncertainty over the variable is more important, and since the log dominates the other term in the product of the entropy, we will have a bigger entropy with the sum. The surprise is over the `Lightning` variable and `Air quality` variables which have a low entropy with a cardinality of 3, which may be explained by the fact that there is a low uncertainty on these variables, so some of their values have a high probability, while others have a really low. (for lightning the probability distribution is [0.952 0.0278 0.0202] and [0.9078 0.0478 0.0444] for air quality, we are really far from a uniform distribution, it explains the low entropy, since the uniform distribution has a maximum entropy).
From a theoretically point of view, we know that the entropy of each variable is bounded by its cardinality such that :

$$\mathcal{H}(\mathcal{X}) \leq log_2(|\mathcal{X}|)$$

This theoretical theorem states that the associated entropy of a variable with a uniform distribution is always maximal. Table 2 shows the entropies that would have the variables if they were uniformly distributed. Moreover, knowing

Cardinality	Entropy
2	1.0
3	1.585
4	2.0
7	2.807
12	3.585

Table 2: Maximum entropy, obtained with an uniform distribution

We see that the entropy of the majority of our variables are close to the maximum entropy, which means that we have high uncertainty over our variables.

2.2 Question 7

Variable given	$\mathcal{H}$(next_day_rain| .)
Temperature	1.5681
Air pressure	0.94
Same day rain	**1.3895**
Relative humidity	1.3011
Wind direction	**1.5678**
Wind speed	1.5678
Cloud height	1.5668
Cloud density	1.5666
Month	1.5649
Day	1.5672
Daylight	1.5683
Lightning	1.5682
Air quality	1.5679

Table 3: Conditional entropy of next day rain given other variables

Knowing that the mutual information is given by :

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

we have $\mathcal{I}(\textit{next\_day\_rain; wind\_direction}) = 1.569 - 1.5678 = 0.0012$ and
$\mathcal{I}(\textit{next\_day\_rain; same\_day\_rain}) = 1.569 - 1.3895 = 0.1795$
The mutual information measures how much one random variable tell us about another. This means that having information about same day rain will help us more to predict next day rain than having information on wind speed.

2.3 Question 8

The mutual information between relative humidity and wind speed is equal to 0.00012, while the one between month and temperature is equal to 0.575.

The mutual information $\mathcal{I}(\mathcal{X};\mathcal{Y})$ is equal to 0 when the 2 random variables $\mathcal{X}$ and $\mathcal{Y}$ are totally independent. So here relative humidity and wind speed can be considered as independent, where month and temperature are way more dependent from each other (and we know that indeed the temperature depends partially on the month, so nothing surprising).

2.4 Question 9

Variable	Mutual information
Temperature	0.0006
Air pressure	**0.6287**
Same day rain	0.1792
Relative humidity	0.2676
Wind direction	0.0008
Wind speed	0.0009
Cloud height	0.0019
Cloud density	0.0021
Month	0.0038
Day	0.0015
Daylight	0.0004
Lightning	0.0004
Air quality	0.0008

Table 4: Mutual information between each variable and *next_day_rain*

We would choose the air pressure as the only variable to measure because it has the highest mutual information with next day rain, which indicates that this is the variable that explains the most next day rain. We would also choose air pressure looking to the conditional entropy since it has the lowest one, which means that the uncertainty over next day rain is the smallest when air pressure is given.

2.5 Question 10

Variable	Conditional entropy	Mutual information
Temperature	0.9985	0.0008
Air pressure	0.9912	0.0081
Same day rain	0.8413	0.1581
Relative humidity	0.5601	0.4392
Wind direction	0.9985	0.0008
Wind speed	0.9986	0.0007
Cloud height	0.9988	0.0005
Cloud density	0.9991	0.0003
Month	0.9955	0.0039
Day	0.9987	0.0006
Daylight	0.999	0.0003
Lightning	0.999	0.0004
Air quality	0.9982	0.0011

Table 5: Conditional entropy and mutual information between each variable and *next_day_rain*, filtered possible values (`deluge, drizzle`)

Considering only the values `drizzle` and `deluge` of the *next_day_rain* variable, we would choose the relative humidity variable for the same reason as in question 9. Once again, the choice being made according to the conditional entropy or the mutual information gives us the same result.

2.6 Question 11

Variable $\mathcal{X}$	$\mathcal{H}(next\_day\_\ rain, \mathcal{X} \mid temperature)$	$\mathcal{I}(next\_day\_rain; \mathcal{X} \mid temperature)$
Air pressure	1.9386	0.6295
Same day rain	2.8629	0.1799
Relative humidity	2.300	0.2675
Wind direction	3.5651	0.0017
Wind speed	3.1514	0.0012
Cloud height	3.1501	0.0024
Cloud density	3.1488	0.0032
Month	4.5685	0.0076
Day	4.3684	0.0048
Daylight	2.5653	0.0014
Lightning	1.8207	0.0009
Air quality	2.0437	0.0016

Table 6: Conditional joint entropy and conditional mutual information between *next_day_rain* and each variable, given the temperature

Table 6 reports the conditional joint entropy and mutual information between *next_day_rain* and each variable, given the temperature.
According to the conditional joint entropy $\mathcal{H}(\texttt{next day rain}, \mathcal{X}|\texttt{temperature})$, the variable $\mathcal{X}$ we should take is the lightning since it has the lowest conditional joint entropy.
According to the conditional mutual information $\mathcal{I}(\texttt{next day rain}; \mathcal{X}|\texttt{temperature})$ the variable $\mathcal{X}$ that we should take is air pressure since it has the highest conditional mutual information.

3 Playing with information theory-based strategy (Wordle)

3.1 Question 12

In this initial context, all the letters have the same probability to be in each of the fields, so the probability distribution of each field corresponds to a uniform distribution of 26 items, thus $P(\text{field}) = (p_1, ...p_{26})$ where $p_j = \frac{1}{26}, \ \forall j = 1, ..., 26$. The entropy of a field is then equal to 4.7.
For the whole board, we can make 26^5 combinations of words since we have 26 letters and 5 fields, and all those words have the same probability to appear. The probability distribution is thus $P(\text{board}) = (p_1, ...p_{26^5})$ where $p_j = \frac{1}{26^5}, j = 1, ..., 26^5$. It gives us an entropy of 23.5022.
The two quantities $P(\text{field})$ and $P(\text{board})$ are linked because $P(\text{board}) = 5 \cdot P(\text{field})$ thanks to the uniform distribution.

3.2 Question 13

As we know that the letter in the second field is A, there is no uncertainty left and its entropy is equal to 0. For the other fields, we delete 4 letters from our pool of letters (T, B, L and E). The A can be repeated in the other fields so we can't delete it from our pool. We are left with 22 letters and, as in the 12^{th} question, all the letters have the same probability to appear in each of the field. We have a uniform distribution of 22 values for the probability distribution of each field. It gives us an entropy of 4.4549 for all the fields.
We have 22 letters that can go in any combination in 4 fields, it gives us a probability distribution $P(\text{board}) = (p_1, ...p_{22^4})$ where $p_j = \frac{1}{22^4}, \ \forall j = 1, ..., 22^4$, which yields an entropy of 17.8377.
The A placed in the right field brings us $23.5022 - 17.8377 = 5.6645$ bits of information.

3.3 Question 14

Here we delete 4 more letters from our pool of letters, R, O, U and H, we are left with 18 letters.

For the $1^{st}, 3^{rd}, 5^{th}$ fields, we have 18 possible letters, but we don't have a uniform probability distribution. Indeed G is at least in one of those three fields (since we know it is in the board, but we don't know in which field, but it may be in those 3 fields simultaneously). Since we know that it must appear in the board, it is at least in one of the 3 fields, it has a probability of 1/3 in each of the three fields. The other letters thus share the 2/3 probability uniformly. It leads us to an entropy of 3.738.
For the 2^{nd} field the probability distribution doesn't move. The entropy is still 0.
For the 4^{th} field there can be any letter still in our pool of letters excepted the G, so there is 17 possible letters with all the same probability. The resulting entropy is 4.0875.
To compute the entropy of the game we first need to compute the number of combinations possible to have the probability distribution (since all those combinations have the same probability to appear). To do so, we can compute the number of combinations where there is 1 G, the ones with 2 G and the ones with 3 G, as we know that there is at least 1 G in the board but not in the fields 2 and 4, but they may be until 3 G's. The number of possible combinations with only 1 G and the 17 other letters are : (number of locations of G) * (possible combinations of the 17 letters in the 3 fields where the G is not) = $3 * 17^3$. The number of locations of G is 3 since it can't be neither int the second or the fourth fields.
With 2 G's we get : (number of locations of the pair of G's) * (possible combinations of the 17 letters in the 2 fields where the G is not) = $3 * 17^2$. And we 3 G's : $1 * 17$ (since the G's occupies the 3 possible fields ($1^{st}, 3^{rd}, 5^{th}$, and there is only one field left in which we can put any of the 17 remaining letters.
It leads us to $3 * 17^3 + 3 * 17^2 + 17$ combinations, all with the same probability, which gives us an uniform distribution of $3 * 17^3 + 3 * 17^2 + 17$ values. It yields an entropy of 13.931.
The sum of the entropies of the different fields is bigger than the entropy of the whole board. It's because the fields are dependant from one to another, which leads to different results when considering the whole board or the fields individually.

3.4 Question 15

In the simplified version we have $26^5 = 11881376$ possible words, so much more than the 12 000 words dictionary we have to match. In the simplified version, the probability distribution is $P(\text{board}) = (p_1, ...p_{26^5})$ where $p_j = \frac{1}{26^5}, \ \forall j = 1, ..., 26^5$. In the real game, we cannot compute the the entropy of the whole game because the distribution of letters is unknown. However the maximum entropy is when we have a uniform distribution which would give 1/12000. The entropy, thus the uncertainty, is way lower in the real game than in the simplified one.

3.5 Question 16

One possible approach based on available information would be to select the guess that would maximize the entropy of our guess word. Thus it would maximize the expected amount of information gained with this guess.