LIÈGE université
Sciences Appliquées

ELEN0060 - INFORMATION AND CODING THEORY

Project 1 - report

Information measures

Authors:
Antoine DECKERS (s170999)
Antoine DEBOR (s173215)

Lecturer:
Pr. L. WEHENKEL
T.A. :
A. SUTERA

Saturday 13th March, 2021

The *Python* script implementing the following functions and providing the presented results can be found in the file `s173215s170999.py` of the submitted archive.

Implementation

1. The function `entropy` computes the entropy $H(\mathcal{X})$ of a random variable $\mathcal{X}$ from its probability distribution $P_{\mathcal{X}} = (p_1, p_2, \ldots, p_n)$. It uses the mathematical formula of equation eq. (1).

$$H(\mathcal{X}) = -\sum_{i=1}^{n} p_i \log_2(p_i) \tag{1}$$

   The implementation is quite simple: it is composed of a loop over the probability distribution (given as an array) and uses the `log` function of the *math* package. The only key part of the implementation comes from the nature of the logarithm function, which is not defined at the origin. Therefore, if, for any i, $p_i = 0$, $\log(p_i)$ is not defined and this eventually leads to an error on the computer. To avoid that, the theory says that the entropy should not take the zero probabilities into account. This is implemented through a condition in the code, checking that each product inside the sum of eq. (1) is computed only if $p_i \neq 0$.

   Intuitively, the entropy of a random variable, considered as a source, measures the amount of information carried by this variable.

2. The function `joint_entropy` computes the joint entropy $H(\mathcal{X}, \mathcal{Y})$ of two discrete variables $\mathcal{X}$ and $\mathcal{Y}$ from the joint probability distribution $P_{\mathcal{X},\mathcal{Y}}$, whose elements are denoted $P_{\mathcal{X},\mathcal{Y}}(X_i, Y_j) = p_{i,j}$. It uses the mathematical formula of eq. (2).

$$H(\mathcal{X}, \mathcal{Y}) = -\sum_{i=1}^{n}\sum_{j=1}^{m} p_{i,j} \log_2(p_{i,j}) \tag{2}$$

   Again, the implementation is quite simple: it considers the joint distribution as a 2D array, whose rows correspond to the values Y_j and whose columns correspond to the values X_i, and computes the joint entropy using a double loop over the whole distribution 2D array. Again, the key part is to take special care concerning zero-valued joint probabilities, which is done in the exact same way as for the `entropy` function.

   The functions `entropy` and `joint_entropy` are very similar. Indeed, except that the latter implements a double loop over all elements of the joint probability distribution given as a 2D array, they share the same structure. Actually, this double loop could have been avoided using some *Numpy* embedded functions which directly works with 2D arrays for the sum. In this case, the two functions would have been identical, which is logical regarding the two theoretical formulas used, eq. (2) being the same as eq. (1) if the latter considers $\mathcal{Z} = (\mathcal{X}, \mathcal{Y})$ as the random variable whose entropy has to be computed. However, for the sake of clarity, it has been decided to keep a simple structure sticking to the theoretical formula and thus implementing a double sum "by-hand".

3. The function `conditional_entropy` computes the conditional entropy $H(\mathcal{X}|\mathcal{Y})$ of a discrete random variable $\mathcal{X}$ given another discrete random variable $\mathcal{Y}$ from the conditional probability distribution $P_{\mathcal{X}|\mathcal{Y}}$ and the marginal probability distribution $P_{\mathcal{Y}}$. It uses the formula of eq. (3), which uses results from eq. (4) [1].

$$H(\mathcal{X}|\mathcal{Y}) = \sum_{j=1}^{m} P_{\mathcal{Y}}(Y_j) H(\mathcal{X}|\mathcal{Y} = Y_j) \tag{3}$$

[1] Actually, the quantity computed by eq. (3) is the mean conditional entropy, which corresponds to the mean of the conditional entropies obtained using eq. (4).

$$H(\mathcal{X}|\mathcal{Y}=Y_j) = -\sum_{i=1}^{n} P_{\mathcal{X}|\mathcal{Y}}(X_i|Y_j)\log_2(P_{\mathcal{X}|\mathcal{Y}}(X_i|Y_j)) \tag{4}$$

Again, the implementation is quite simple: it first computes the values $H(\mathcal{X}|\mathcal{Y}=Y_j)$ for each value Y_j, using eq. (4) and the given conditional probability distribution. As before, one needs to pay attention to zero-valued probabilities, which is handled in the same way as it has been previously depicted. Then, the total conditional entropy is computed using eq. (3), plugging in the different values $H(\mathcal{X}|\mathcal{Y}=Y_j)$ that have just been computed and using the marginal probability distribution $P_{\mathcal{Y}}$. Both steps of the implementation are performed inside loops, in a similar fashion as for the previous functions.

An equivalent way to compute this quantity is to use the formula of eq. (5).

$$H(\mathcal{X}|\mathcal{Y}) = -\sum_{i=1}^{n}\sum_{j=1}^{m} P_{\mathcal{X},\mathcal{Y}}(X_i,Y_j)\log_2\left(\frac{P_{\mathcal{X},\mathcal{Y}}(X_i,Y_j)}{P_{\mathcal{Y}}(Y_j)}\right) \tag{5}$$

This requires to have the joint probability distribution $P_{\mathcal{X},\mathcal{Y}}$ at disposal, rather than the conditional one required when using the previous procedure to compute the conditional entropy. This formula naturally follows from plugging eq. (4) into eq. (3) and using the relation of eq. (6) for the conditional probability.

$$P_{\mathcal{X}|\mathcal{Y}}(X_i|Y_j) = \frac{P_{\mathcal{X},\mathcal{Y}}(X_i,Y_j)}{P_{\mathcal{Y}}(Y_j)} \tag{6}$$

Yet another way to compute the conditional entropy is to use eq. (7), which can be computed using the functions `entropy` and `joint_entropy`.

$$H(\mathcal{X}|\mathcal{Y}) = H(\mathcal{X},\mathcal{Y}) - H(\mathcal{Y}) \tag{7}$$

4. The function `mutual_information` computes the mutual information $I(\mathcal{X};\mathcal{Y})$ between two discrete random variables $\mathcal{X}$ and $\mathcal{Y}$ from the marginal probability distributions $P_{\mathcal{X}}$ and $P_{\mathcal{Y}}$ and joint probability distribution $P_{\mathcal{X},\mathcal{Y}}$. It uses the formula of eq. (8).

$$I(\mathcal{X};\mathcal{Y}) = \sum_{i=1}^{n}\sum_{j=1}^{m} P_{\mathcal{X},\mathcal{Y}}\left(X_i,Y_j\right)\log_2\left(\frac{P_{\mathcal{X},\mathcal{Y}}\left(X_i,Y_j\right)}{P_{\mathcal{X}}\left(X_i\right)P_{\mathcal{Y}}\left(Y_j\right)}\right) \tag{8}$$

Again, the implementation is quite simple: it iterates over all the elements of the joint probability distribution in a double loop, applies the formula using the given probability distributions, while taking care to discard indices i,j such that $P_{\mathcal{X},\mathcal{Y}}\left(X_i,Y_j\right)=0$ or $P_{\mathcal{X}}\left(X_i\right)=0$ or $P_{\mathcal{Y}}\left(Y_i\right)=0$, for which the logarithm function is not defined.

The mutual information $I(\mathcal{X};\mathcal{Y})$ might allow to deduce some clues about the relationship between $\mathcal{X}$ and $\mathcal{Y}$. Indeed, from the definitions, one has the relation of eq. (9).

$$I(\mathcal{X};\mathcal{Y}) = H(\mathcal{X}) - H(\mathcal{X}|\mathcal{Y}) = H(\mathcal{Y}) - H(\mathcal{Y}|\mathcal{X}) = H(\mathcal{X}) + H(\mathcal{Y}) - H(\mathcal{X},\mathcal{Y}) \tag{9}$$

Therefore, using this relation, one has the following particular cases:

(a) If $I(\mathcal{X};\mathcal{Y}) = H(\mathcal{X})$, therefore $H(\mathcal{Y}) = H(\mathcal{X},\mathcal{Y})$ and $\mathcal{X}$ is a deterministic function of $\mathcal{Y}$. Likewise, if $I(\mathcal{X};\mathcal{Y}) = H(\mathcal{Y})$, therefore $H(\mathcal{X}) = H(\mathcal{X},\mathcal{Y})$ and $\mathcal{Y}$ is a deterministic function of $\mathcal{X}$. In particular, if $I(\mathcal{X};\mathcal{Y}) = H(\mathcal{X}) = H(\mathcal{Y})$, $\mathcal{X}$ is a one-to-one function of $\mathcal{Y}$.

(b) If $I(\mathcal{X};\mathcal{Y}) = 0$, therefore $H(\mathcal{X}) = H(\mathcal{X}|\mathcal{Y})$ and $H(\mathcal{Y}) = H(\mathcal{Y}|\mathcal{X})$, hence $\mathcal{X}$ and $\mathcal{Y}$ are independent.

Otherwise, $0 < I(\mathcal{X};\mathcal{Y}) < \min(H(\mathcal{X}), H(\mathcal{Y}))$ and the only conclusion that can be drawn is that $\mathcal{X}$ and $\mathcal{Y}$ are correlated.

5. The functions `cond_joint_entropy` and `cond_mutual_information` respectively compute $H(\mathcal{X}, \mathcal{Y}|\mathcal{Z})$ and $I(\mathcal{X}; \mathcal{Y}|\mathcal{Z})$, given that $\mathcal{X}$, $\mathcal{Y}$ and $\mathcal{Z}$ are three discrete random variables.

   The first quantity is computed using the relation of eq. (10), which naturally follows from the general property of eq. (11) with $\mathcal{W} = (\mathcal{X}, \mathcal{Y})$.

$$H(\mathcal{X}, \mathcal{Y}|\mathcal{Z}) = H(\mathcal{X}, \mathcal{Y}, \mathcal{Z}) - H(\mathcal{Z}) \tag{10}$$

$$H(\mathcal{W}, \mathcal{Z}) = H(\mathcal{Z}) + H(\mathcal{W}|\mathcal{Z}) \tag{11}$$

   Again considering $\mathcal{W} = (\mathcal{X}, \mathcal{Y})$, this conditional joint entropy can be easily implemented by calling both the `entropy` and `joint_entropy` functions previously defined, and by making the difference between the return values of these two functions. One thus need to use the probability distribution $P_{\mathcal{Z}}$ as input of the first function and $P_{\mathcal{W},\mathcal{Z}} = P_{\mathcal{X},\mathcal{Y},\mathcal{Z}}$ as input of the second one. These two distributions are taken by the function `cond_joint_entropy` as inputs.

   The second quantity is computed using the relation of eq. (12).

$$I(\mathcal{X}; \mathcal{Y}|\mathcal{Z}) = H(\mathcal{X}|\mathcal{Z}) - H(\mathcal{X}|\mathcal{Y}, \mathcal{Z}) \overset{eq.\ (9)}{=} -I(\mathcal{X}; \mathcal{Z}) + H(\mathcal{X}) - H(\mathcal{X}|\mathcal{Y}, \mathcal{Z}) \tag{12}$$

   The sum of the two first terms of this conditional mutual information can be easily implemented by calling both the `entropy` and `mutual_information` functions previously defined. One thus need to use the probability distribution $P_{\mathcal{X}}$ as input of the first function, and the three distributions $P_{\mathcal{X}}$, $P_{\mathcal{Z}}$ and $P_{\mathcal{X},\mathcal{Z}}$ as input of the second one.
   The third term is more tricky to derive. It uses an auxiliary function called `joint_cond_entropy`, which computes $H(\mathcal{B}|\mathcal{A}, \mathcal{C})$[2] given $P_{\mathcal{A},\mathcal{B},\mathcal{C}}$ and $P_{\mathcal{A},\mathcal{C}}$. This function relies on the same structure as the `conditional_entropy` function, as it considers the variable $\mathcal{W} = (\mathcal{A}, \mathcal{C})$ as the conditioning one. Note that, to compute the third term of eq. (12), one considers $\mathcal{W} = (\mathcal{Y}, \mathcal{Z})$. Hence, this auxiliary function is called with $P_{\mathcal{Y},\mathcal{X},\mathcal{Z}}$ (which corresponds to the 3D array depicting $P_{\mathcal{X},\mathcal{Y},\mathcal{Z}}$ transposed) and $P_{\mathcal{Y},\mathcal{Z}}$ as inputs.
   Therefore, the function `cond_mutual_information` takes the following probability distributions as inputs: $P_{\mathcal{X}}$, $P_{\mathcal{Z}}$, $P_{\mathcal{X},\mathcal{Z}}$, $P_{\mathcal{Y},\mathcal{Z}}$ and $P_{\mathcal{X},\mathcal{Y},\mathcal{Z}}$.

   N.B. 1: The last term of eq. (12) could have been computed using the functions `cond_joint_entropy` and `conditional_entropy`. Indeed, one can write the following general additivity property

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

   which, after inverting $\mathcal{X}$ with $\mathcal{Y}$, yields

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

   N.B. 2: The second quantity could also be computed using eq. (13).

$$I(\mathcal{X}; \mathcal{Y} \mid \mathcal{Z}) = \sum_{i,j,k} P(X_i, Y_j, Z_k) \log \frac{P(X_i, Y_j \mid Z_k)}{P(X_i \mid Z_k)\, P(Y_j \mid Z_k)} \tag{13}$$

Medical diagnosis

6. The entropy of each medical variable considered in the data set can be seen in table 1, along with the corresponding cardinality.

[2]Please note that the order of the variables matters. See the code for further insights.

Variable $\mathcal{X}$	Entropy $H(\mathcal{X})$	Cardinality
age	0.9998	2
sex	0.9992	2
obesity	1.2868	3
ALC	0.9703	2
iron	1.9014	4
DIS	0.8969	3
fatigue	0.6531	2
TRI	0.6234	2
ALT	0.5615	2
AST	0.8385	2
GGTP	0.8601	2
CHL	1.4774	3
AMA	0.5749	2
MSC	0.9886	2
BIL	0.4529	2
ITC	0.7414	2
JAU	0.4674	2

Table 1: Entropy (in Shannon) and cardinality of each medical variable

One can notice that, while not being a strict relationship, there is a tendency linking the cardinality of a variable with its entropy. Indeed, one can notice that, most of the time, the larger the cardinality, the larger the entropy. Except for the variable DIS (i.e. the disease itself), each variable whose cardinality is greater than 2 corresponds to an entropy greater than 1, the one with a cardinality equal to 4 giving the maximal entropy value. This can be intuitively explained by the fact that the entropy of a random variable, assuming that it corresponds to a source, corresponds to the quantity of information that the source must provide to know the value of the random variable without ambiguity. Therefore, if a random variable can take a large amount of different values, one could expect this quantity of information to be larger than for a variable that can take a smaller amount of values, *e.g.* 2.
This tendency can theoretically be motivated by the theorem of eq. (14).

$$H_n\left(p_1, p_2, \ldots, p_n\right) \leq \log n, \text{ with equality } \Leftrightarrow \forall i : p_i = \frac{1}{n} \tag{14}$$

Indeed, this theorem states that the entropy of a random variable is upper bounded by the logarithm of its cardinality. This means that, since the logarithm function is a strictly increasing function on its domain, the larger the cardinality, the larger the upper bound.

7. The conditional entropy of the disease given each of the other variables can be seen in table 2.

Conditioning variable $\mathcal{X}$	Conditional entropy $H(DIS\|\mathcal{X})$
age	0.8435
sex	0.8129
obesity	0.8842
ALC	0.8880
iron	0.8940
fatigue	0.6038
TRI	0.7471
ALT	0.7073
AST	0.6414
GGTP	0.7944
CHL	0.8738
AMA	0.45100
MSC	0.8538
BIL	0.6827
ITC	0.8838
JAU	0.7633

Table 2: Conditional entropy (in Shannon) of the disease given each of the other medical variables

The conditional entropy has a larger value when the conditioning variable is *jaundice* than when it is *bilirubin*. This thus corresponds through eq. (9) to a larger mutual information $I(\text{DIS}; \textit{bilirubin})$ than $I(\text{DIS}; \textit{jaundice})$, since $H(\text{DIS})$ is constant. Moreover, since, from a physiological point of view, the jaundice results from an accumulation of bilirubin, one could consider the jaundice as a function of the concentration in bilirubin, which could be written as $\text{JAU} = g(\text{BIL})$. This observation hence illustrates the principle of non creation of information, which states that, whatever the function $g(.)$, one has

$$I(\mathcal{X}; \mathcal{Y}) \geq I(\mathcal{X}; g(\mathcal{Y})).$$

Intuitively, this principle implies that, whatever the treatments performed on $\mathcal{Y}$, the information provided on $\mathcal{X}$ by the result of these treatments can not increase.
In the considered case, one indeed observes

$$I(\text{DIS}; \text{BIL}) \geq I(\text{DIS}; \text{JAU}) = I(\text{DIS}; g(\text{BIL})),$$

which corroborates the above mentioned principle.

8. The mutual information between the variables *obesity* and *age* is equal to

$$I(\text{obesity; age}) = 0.0003833.$$

This value is very small, and could be considered as nearly equal to zero. Therefore, taking into account the fact that the computed results derive from an empirical process (the data set does not correspond to the entire population), and are thus not a hundred percent accurate, one can reasonably conclude that the variables *obesity* and *age* are independent. This conclusion can be justified by the discussion about the mutual information held in the implementation part of this report.

9. Assuming that the medical diagnosis has to be made knowing only *one* variable (except the variable DIS, obviously), the choice of this very variable can be made based on the mutual information. Indeed, one should choose the variable $\mathcal{X}$ corresponding to the largest mutual information $I(\text{DIS}; \mathcal{X})$, because this quantity measures the statistical dependence of the disease and $\mathcal{X}$. As it has been previously explained, a mutual information equal to zero corresponds to two independent variables and, on the contrary, a large value for this quantity expresses a strong dependence between the two considered variables. Therefore, in order to make a proper diagnosis, one should seek for the variable having the strongest dependence with the disease.

Candidate variable $\mathcal{X}$	Mutual information $I(\text{DIS};\mathcal{X})$
age	0.0533
sex	0.0839
obesity	0.0126
ALC	0.0088
iron	0.0028
fatigue	0.2930
TRI	0.1497
ALT	0.1895
AST	0.2554
GGTP	0.1024
CHL	0.0230
AMA	0.4458
MSC	0.0430
BIL	0.2141
ITC	0.0130
JAU	0.1335

Table 3: Mutual information between the disease and each of the other variables

From table 3, one can thus see that the mutual information is maximized for the variable AMA, which should hence be chosen as the measured variable.

If the choice had to be made based on the conditional entropy, one should choose the variable $\mathcal{X}$ providing the smallest conditional entropy $H(\text{DIS}|\mathcal{X})$. Indeed, this quantity represents the amount of information needed to know the behaviour of the variable DIS, knowing $\mathcal{X}$. In other words, it expresses the residual uncertainty on DIS when knowing $\mathcal{X}$. Therefore, one should seek for a variable which, once known, brings the largest amount of information about the behaviour of the disease, hence corresponding to a small conditional entropy. From table 2, one can see that the best choice according to this criterion would still be the variable AMA. This is not surprising because, according to eq. (9), one has $I(\text{DIS}|\mathcal{X}) = H(\text{DIS}) - H(\text{DIS}|\mathcal{X})$, which is the difference between a constant term and $H(\text{DIS}|\mathcal{X})$. Therefore, if the variable AMA corresponds to the largest value of $H(\text{DIS}|\mathcal{X})$, it obviously corresponds to the smallest value of $I(\text{DIS}|\mathcal{X})$ also.

10. Considering only the samples with the disease value corresponding to *healthy* or *steatose*, one can perform the same process as in the previous section. Please note that, however, the different probability distributions have been re-computed considering the new data set (without the occurrences of the third value of the variable DIS).

Candidate variable $\mathcal{X}$	Mutual information $I(\text{DIS};\mathcal{X})$
age	0.00005
sex	0.0004
obesity	0.0138
ALC	0.0101
iron	0.0032
fatigue	0.0119
TRI	0.1715
ALT	0.0057
AST	0.0086
GGTP	0.012
CHL	0.0015
AMA	0.0031
MSC	0.0002
BIL	0.0067
ITC	0.0002
JAU	0.0044

Table 4: Mutual information between the disease and each of the other variables, without the samples for which DIS = *PBC*

From table 4, one can thus see that, in this case, the mutual information is maximized for the variable TRI, which should hence be chosen as the measured variable for the same reason as explained before. Therefore, the choice of the variable of interest is changed from the one identified in the previous section.

As in the previous section, if the choice had to be made based on the conditional entropy, one should choose the variable $\mathcal{X}$ providing the smallest conditional entropy $H(\text{DIS}|\mathcal{X})$ for the same reason as previously explained. Therefore, this quantity is computed considering again only the samples with the disease value corresponding to *healthy* or *steatose*. The results can be seen in table 5.

Candidate variable $\mathcal{X}$	Conditional entropy $H(\text{DIS}|\mathcal{X})$
age	0.3821
sex	0.3817
obesity	0.3684
ALC	0.372
iron	0.3789
fatigue	0.3702
TRI	0.2106
ALT	0.3764
AST	0.3735
GGTP	0.3701
CHL	0.3806
AMA	0.3789
MSC	0.3819
BIL	0.3754
ITC	0.3819
JAU	0.3777

Table 5: Conditional entropy (in Shannon) of the disease given each of the other medical variables, without the samples for which DIS = *PBC*

Based on table 5, one can observe that the smallest value of the entropy is again obtained for the variable TRI. This is again not surprising invoking the reasoning presented in the previous section

using eq. (9). Therefore, based on the mutual information or on the conditional entropy, one identifies the same variable of interest which is different from the one identified considering the whole data set.

11. Based on the age of the patient, one can follow the same process as in the two previous sections to determine the variable of interest for making the best diagnosis. This time, however, the relevant information measure is the conditional mutual information with the variable *age* being the conditioning one. One thus seek the variable $\mathcal{X}$ leading to the largest value of $I(\text{DIS}; \mathcal{X}|age)$, again for the same reason as in the two previous sections.

Candidate variable $\mathcal{X}$	Conditional mutual information $I(\text{DIS}; \mathcal{X}\|age)$
sex	0.0965
obesity	0.0152
ALC	0.0092
iron	0.0088
fatigue	0.2681
TRI	0.1509
ALT	0.1733
AST	0.2384
GGTP	0.1014
CHL	0.0218
AMA	0.4039
MSC	0.0402
BIL	0.2038
ITC	0.0141
JAU	0.1226

Table 6: Conditional mutual information between the disease and each of the other variables, given the variable *age*

From table 6, one can see that, despite the fact that the age is in this case known, the choice of the variable of interest is the same as initially, *i.e.* the variable AMA, since it corresponds to the largest conditional mutual information.

If the choice had to be made based on the conditional entropy, one should choose the variable $\mathcal{X}$ providing the smallest conditional entropy $H(\text{DIS}|\mathcal{X}, age)$, again for the same reason as previously explained. This conditional entropy can be computed using results that have already been computed before, following eq. (15).

$$I(\text{DIS}; \mathcal{X}|\, age) = H(\text{DIS}|\, age) - H(\text{DIS}|\, \mathcal{X}, age) \Leftrightarrow H(\text{DIS}|\, \mathcal{X}, age) = H(\text{DIS}|\, age) - I(\text{DIS}; \mathcal{X}|\, age) \quad (15)$$

This formula indeed uses results computed in section 7 and in this one. The results are written in table 7.

Candidate variable $\mathcal{X}$	Conditional entropy $H(\text{DIS}\|\mathcal{X}, age)$
sex	0.7470
obesity	0.8283
ALC	0.8343
iron	0.8347
fatigue	0.5755
TRI	0.6926
ALT	0.6703
AST	0.6051
GGTP	0.7421
CHL	0.8217
AMA	0.4396
MSC	0.8034
BIL	0.6398
ITC	0.8294
JAU	0.7209

Table 7: Conditional entropy of the disease, given each of the other variables (except *age*) and given the variable *age*

From these results, one can conclude that, if the choice had to be made using conditional entropy and based on the variable *age*, the variable to measure would again be AMA. This choice is the same as the one derived using the mutual information, which is not surprising since eq. (15) computes the difference between a constant term ($H(\text{DIS}|\, age)$) and the conditional mutual information computed before. Therefore, using the same reasoning as before, if the variable AMA corresponds to the largest value of the conditional mutual information, it obviously corresponds to the smallest value of $H(\text{DIS}|\, age)$ also.

Playing with information theory-based strategy

12. As $X_{i,j}$ is a binary random variable, one can write that its entropy is equal to

$$H(\mathcal{X}) = -p \log p - (1-p) \log(1-p) = H_2(p) \tag{16}$$

where p represent the probability of one of the two values of X and 1-p the other one. In this scenario, one can write P($X_{i,j} = 1$) = p (there is a mine on the field (i,j)) and P($X_{i,j} = 0$) = 1-p (no mine on field (i,j)).

The only information one has is that there is M mines amongst RC fields on the board. Therefore, for each field, one could state that $p = \frac{M}{RC}$. Finally, the entropy of each field can be computed as follows :

$$H_2(p) = -\frac{M}{RC} \log \frac{M}{RC} - (1 - \frac{M}{RC}) \log(1 - \frac{M}{RC}) \tag{17}$$

13. For simplicity sake, one could denote the fields using lower case letters as illustrated on table 8.

a	b	c	d	e
f	g	h	i	j
k	l	m	n	o

Table 8: notation

0	1	2	3	2
1	2			
1				

Table 9: case 1

In order to compute the entropy of each field adjacent to a clue, one could be tempted to compute the probability of presence of a mine based on the adjacent clues distribution. However, the data derived from one clue is not independent from the other clues, so we cannot base our reasoning on the clues probability distribution. One could then use a logical reasoning on the clues constraints in order to compute the entropy of the unrevealed field. In the subsequent reasoning, one can use the following notations :

- clue effective value = clue value - number of assumed mines in adjacent unrevealed fields
- * represent a field where a mine is assumed to be
- / represent a field where a clue is assumed to be

First of all, the clue of field k and b requires that there must be a mine on field l and h respectively, regardless of other clues adjacent to these fields. As a consequence, the effective value of the clue on field g is now equal to $2 - 2 = 0$. Therefore, there must be a clue on field m.

0	1	2	3	2
1	2	*		
1	*	/		

Table 10: case 1 update 1

Moreover, the effective value of the clue on field c is now 1 (since we marked field h). Therefore, the only adjacent field, i must be marked with the presence of a mine.

0	1	2	3	2
1	2	*	*	
1	*	/		

Table 11: case 1 update 2

Finally, the effective value of the clue on field d is now equal to $3 - 2 = 1$, imposing that the only unrevealed adjacent field, j, must be marked with the presence of a mine.

0	1	2	3	2
1	2	*	*	*
1	*	/		

Table 12: case 1 update 3

By a logical reasoning, one could deduce the presence or absence of mines on every unrevealed field adjacent to a clue. The results can be summarized as follow :

- $\mathrm{P}(X_{i,j} = 1) = 1$ for $(i, j) = \{(3, 2); (2, 3); (2, 4); (2, 5)\}$
- $\mathrm{P}(X_{i,j} = 0) = 1$ for $(i, j) = (3, 3)$

The entropy of the unrevealed fields can be immediately deduce, using H_2^l to denote the entropy of the field l considered as a binary source :

- $H_2^l(p) = H_2^h(p) = H_2^i(p) = H_2^j(p) = 0$, with $p = 1$
- $H_2^m(p) = 0$, with $p = 0$

One can notice that in the game situation represented on table 9, the entropy of the unrevealed fields adjacent to a clue are all equal to 0, meaning that they do not carry any information as we can estimate the presence of a mine by logical deduction. In other words, revealing any of these fields does not bring any information as there is no uncertainty about their value $X_{i,j}$ (0 or 1).

14. Our assumption is that if a field is a non-zero clue, it should bring information regarding the state of its adjacent unrevealed fields and decrease their uncertainty. Moreover, fields adjacent to a clue will bring little information on this field state as there is no uncertainty. Therefore, our strategy is to choose the field that influence, in the meaning of information measure, the most its adjacent unrevealed fields but is the least influenced by those neighbors.

    Based on this hypothesis, on the oracle knowledge and the current state of the game illustrated on table 9, one could choose to reveal a field (i, j), adjacent to at least one clue, by using one of the following three proposed constraints :

    - The conditional entropy $H(\mathcal{X}_{i,j}|adjacent\_clues)$ is maximised. Therefore, revealing such a field will bring the largest amount of information about neighboring fields, as it will reduce the most their uncertainty. This would also mean that knowing the value of adjacent clues would not bring any information on the state of the field (i, j), *i.e.* this field is the one that is least influenced by its neighbors.
    - The mutual information $I(\mathcal{X}_{i,j}; adjacent\_unrevealed\_fields)$ is maximised. This would mean that one choose to reveal the field that bring the greatest amount of information about its unrevealed neighbors.
    - The conditional entropy $H(adjacent\_unrevealed\_fields|\mathcal{X}_{i,j})$ is minimised. Hence, once known, this cell will bring the smallest uncertainty about its unrevealed neighbors.

    In the next sections, let assume that one consider the board settings to be similar as the ones for question 12, i.e one consider a board of R rows, C columns, M hidden mines, that there is no revealed fields at the start of a new game and that the mines and uniformly distributed on the board. Therefore, at the start, the fields entropy are all the same, equal to the value derived in equation 17. As discuss with the teaching assistant, one should assume that we can ask the oracle information given a current state but not before even starting the game. Nonetheless, one could discuss what would change in his strategy if the oracle could give information measure without even starting the game.

15. One could define a strategy as follows :

    (a) Reveal a first field randomly picked between all the cells. If this cell is a mine, one has lost and the game restarts (one performs (a) again). Otherwise, one go on to play by executing the following steps.

    (b) One should choose one of the constraints stated in the previous section, based on which one would then reveal the next fields.

    (c) From this choice, one would go on playing the game and, at each step, reveal a field, adjacent to at least one clue, that optimizes the constraint chosen.

    (d) Whenever a revealed field is a mine, we start over the game by executing (a).

    (e) The game ends when $RC - M$ clues have been revealed. At that point, we know that the M remaining fields are all mines.

16. One could define a strategy as follows :

    (a) Reveal k fields randomly picked between all the cells. If one or several amongst those cells are a mine, one has lost and the game restarts (one performs (a) again). Otherwise, one go on to play by executing the following steps.

(b) One should choose one of the information measures stated in question 14, based on which one would then reveal the next fields.

(c) From this choice, one would go on playing the game. At each step, one would reveal the k fields, each adjacent to at least one clue, that optimize the best the constraint chosen.

(d) Whenever one of the revealed field is a mine, we start over the game by executing (a).

(e) The game ends when $RC - M$ clues have been revealed. At that point, we know that the M remaining fields are all mine.

In these settings, in order to have a solvable game, we must state the initial condition that $RC - M \ mod \ k \ = \ 0$.

One could discuss the following points regarding the two previous sections :

- If the oracle could give us information measure before even starting the game, we no longer have to execute step (a) and would directly start our strategy by step (b), with the fact that the first revealed field(s) do not have to be adjacent to any clue.
- One could forecast that our strategy would lead to revealing clues in the decreasing order of their value. Indeed, the ones with a highest clue value would lead to a bigger amount of information released and would be revealed firstly compared to the lowest clue values that would release a fewer amount of information.
- It is finally noteworthy that this strategy seems less safe than the one described in the previous section. Indeed, revealing each time k cells increases the risk of revealing a mine. However, one can expect that, once k clues have been revealed at once, using them and applying the chosen criterion to reveal the next ones will increase the speed at which the game is played (and maybe won). Though, even if this is obvious, if at any point the number of unrevealed free fields is lower than k, one will loose for sure when revealing k fields.

N.B.: From a theoretical perspective, one can observe that

$$H(\mathcal{X}, \mathcal{Y}) \geq \max(H(\mathcal{X}), H(\mathcal{Y})),$$

which could justify the fact of revealing several fields at a time instead of one, since it intuitively brings more information. Furthermore, the theory also states that

$$H(\mathcal{X}, \mathcal{Y}) \leq H(\mathcal{X}) + H(\mathcal{Y}),$$

which could encourage to apply the entropy-related criteria on distinct fields and to select the k 'best' ones, instead of looking at the joint entropy of k fields.

17. In this section, an approach that would let one use information theory to play the minesweeper without solving the game is discussed. Unlike the previous sections, no oracle is available here.
It has been decided to first describe a general idea, which aims at making the information measures easier to compute in the case of the minesweeper, and then to present several applications of this general intuition under the form of more specific additional assumptions or rules applied to the minesweeper.

In the original game, one should solve the game in order to determine the different probability distributions in presence to derive the different information measures relative to all fields. As this could become hard and as it is not allowed in this section, one has to find a way to ease the computation of the information measures. This can be done by making the computation of the probability distributions easier or trivial, for instance. This is the basic motivation underlying the following and, to achieve that, the main intuition is to constraint the game in order to decrease the uncertainty about the random variables in presence. The general idea is thus to add one or more rules or assumptions to the game in order to restrict what one could name the state space. Naturally, the more constraints, the more restricted state space. Three different applications of this very basic idea are presented here below:

(a) A first approach would be to consider the game with only one bomb on the board. Therefore, the different probabilities would be easier to compute and the measures too.

(b) A second approach would consist in imposing a given spacing between two bombs on the board, larger or equal to two in order for a field to be located next to at most one bomb. In this case, again, the probabilities will often be easier to compute as this constraint brings less uncertainty about the state of the game.

(c) A third approach could be to consider that the different bombs on the board are following a given coherent organisation schema. For example, it could be assumed that a given number of bombs are on the board, organised such that they draw a square, or any other shape, on the board. The same reasoning as for the previous approaches holds.

(d) A fourth and last idea would be to combine the second one and the third one, in such a way that the bombs are organised in given entities, spaced from each other. The board would in this case look like a battleship one, where the probabilities would again be easier to compute.

For each of these approaches, one could use a similar strategy as described before, *i.e.* choosing the field presenting the largest conditional entropy, for instance. In the case of several fields presenting the same uncertainty, one will however have to discriminate arbitrarily. It is worth noting that, just like in the previous sections, while making it easier to play using information theory, this strategy combined with the constraints described here above does not ensure to win the game. Actually, one can expect the winning rate to be quite low since, at first sight, the described strategy does not seem very efficient at winning the game.