UNIVERSITY OF LIÈGE

INFORMATION AND CODING THEORY

Project 1

MASTER 1 IN DATA SCIENCE & ENGINEERING

Authors:
Tom CRASSET
Antoine LOUIS

Professor :
L. WEHENKEL

Academic year 2018-2019

1 Information measures

1.1 Exercises by hand

The answers to Questions 1 to 6 are shown respectively in Table 1 to 6 and were computed with the probability distributions presented in the Appendix. These probability distributions were themselves computed using the given tables, from which conditional probabilities can be extracted, and the Kolmogorov formula that said that, given two events $\mathcal{X}_i$ and $\mathcal{Y}_j$, the conditional probability of $\mathcal{X}_i$ given $\mathcal{Y}_j$ is defined as :

$$P(\mathcal{X}_i|\mathcal{Y}_j) = \frac{P(\mathcal{X}_i, \mathcal{Y}_j)}{P(\mathcal{Y}_j)}$$

	H(·)
$\mathcal{X}$	2.000
$\mathcal{Y}$	1.750
$\mathcal{W}$	0.896
$\mathcal{Z}$	0.896

Table 1: Entropy

	H(·)
$\mathcal{X}, \mathcal{Y}$	3.375
$\mathcal{X}, \mathcal{W}$	2.703
$\mathcal{Y}, \mathcal{W}$	2.531
$\mathcal{W}, \mathcal{Z}$	0.896

Table 2: Joint entropy

	H(·)
$\mathcal{X}\|\mathcal{Y}$	1.625
$\mathcal{W}\|\mathcal{X}$	0.703
$\mathcal{Z}\|\mathcal{W}$	0.000
$\mathcal{W}\|\mathcal{Z}$	0.000

Table 3: Conditional entropy

	I(·)
$\mathcal{X}; \mathcal{Y}$	0.375
$\mathcal{X}; \mathcal{W}$	0.000
$\mathcal{Y}; \mathcal{Z}$	0.000
$\mathcal{W}; \mathcal{Z}$	0.811

Table 4: Mutual information

	H(·)
$\mathcal{X}, \mathcal{Y}\|\mathcal{W}$	0.963
$\mathcal{W}, \mathcal{Z}\|\mathcal{X}$	0.703

Table 5: Conditional joint entropy

	I(·)
$\mathcal{X}; \mathcal{Y}\|\mathcal{W}$	0.928
$\mathcal{W}; \mathcal{Z}\|\mathcal{X}$	0.703

Table 6: Conditional mutual information

2 Implementation

2.1 Entropy

The implementation of the *entropy* function takes as argument:

- `probability_distribution`, the probability distribution $P(\mathcal{X})$ of the random variable $\mathcal{X}$;

From this argument, it computes the entropy $H(\mathcal{X})$ according to the following formula:

$$H(\mathcal{X}) = -\sum_i P(\mathcal{X}_i)\ log_2 P(\mathcal{X}_i) \tag{1}$$

Intuitively, the entropy of a random variable is a function which attempts to characterize the “unpredictability” of the random variable.

2.2 Joint entropy

The implementation of the *joint_entropy* function takes as argument:

- `joint_distribution`, the joint probability distribution $P(\mathcal{X}, \mathcal{Y})$ between the two random variables $\mathcal{X}$ and $\mathcal{Y}$;

From this argument, it computes their joint entropy $H(\mathcal{X}, \mathcal{Y})$ according to the following formula:

$$H(\mathcal{X}, \mathcal{Y}) = -\sum_i \sum_j P(\mathcal{X}_i, \mathcal{Y}_j) \ log_2 P(\mathcal{X}_i, \mathcal{Y}_j) \tag{2}$$

Basically, the joint entropy is the entropy of a joint probability distribution, or a multi-valued random variable. In other words, joint entropy is really no different than regular entropy. We merely have to compute Equation 1 over all possible pairs of the two random variables.

Intuitively, the joint entropy of two random variables is a function which attempts to characterize the “unpredictability” of these random variables considered together.

2.3 Conditional entropy

The implementation of the *conditional_entropy* function takes as argument:

- `joint_distribution`, the joint probability distribution $P(\mathcal{X}, \mathcal{Y})$ between the two random variables $\mathcal{X}$ and $\mathcal{Y}$;
- `distribution_cond`, the probability distribution $P(\mathcal{Y})$ of the conditional random variable $\mathcal{Y}$;

From these arguments, it computes the conditional entropy $H(\mathcal{X}|\mathcal{Y})$ according to the following formula:

$$H(\mathcal{X}|\mathcal{Y}) = -\sum_i \sum_j P(\mathcal{X}_i, \mathcal{Y}_j) \ log_2 \frac{P(\mathcal{X}_i, \mathcal{Y}_j)}{P(\mathcal{Y}_j)} \tag{3}$$

Intuitively, the conditional entropy characterizes the “unpredictability” of a random variable knowing the value of another random variable.

2.4 Mutual information

The implementation of the *mutual_information* function takes as argument:

- `joint_distribution`, the joint probability distribution $P(\mathcal{X}, \mathcal{Y})$ between the two random variables $\mathcal{X}$ and $\mathcal{Y}$;
- `distribution_var1`, the probability distribution $P(\mathcal{X})$ of the first random variable $\mathcal{X}$;
- `distribution_var2`, the probability distribution $P(\mathcal{Y})$ of the second random variable $\mathcal{Y}$;

From these arguments, it computes the mutual information $I(\mathcal{X};\mathcal{Y})$ according to the following formula:

$$I(\mathcal{X};\mathcal{Y}) = \sum_i \sum_j P(\mathcal{X}_i, \mathcal{Y}_j) \ log_2 \frac{P(\mathcal{X}_i, \mathcal{Y}_j)}{P(\mathcal{X}_i) P(\mathcal{Y}_j)} \tag{4}$$

One could also have used the entropy function along with the conditional entropy function to compute the mutual information using the following formula:

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

Mutual information is a quantity that measures a relationship between two random variables that are sampled simultaneously. In particular, it measures how much information is communicated, on average, in one random variable about another. Intuitively, one might ask, how much does one random variable tell me about another?

An important theorem from information theory says that the mutual information between two variables is 0 if and only if the two variables are statistically independent. On the other hand, a large mutual information between two variables implies that they are strongly correlated.

2.5 Conditional joint entropy

The implementation of the *cond_joint_entropy* function takes as argument:

- `joint_distribution`, the joint probability distribution $P(\mathcal{X},\mathcal{Y},\mathcal{Z})$ between the random variables $\mathcal{X}$, $\mathcal{Y}$ and $\mathcal{Z}$;
- `distribution_cond`, the probability distribution $P(\mathcal{Z})$ of the conditional random variable $\mathcal{Z}$;

From these arguments, it computes the conditional joint entropy $H(\mathcal{X},\mathcal{Y}|\mathcal{Z})$ extending the previously defined *joint_entropy* function, with the following formula :

$$H(\mathcal{X},\mathcal{Y}|\mathcal{Z}) = \sum_k P(\mathcal{Z}_k)\ H(\mathcal{X},\mathcal{Y}|\mathcal{Z}_k) \tag{5}$$

2.6 Conditional mutual information

The implementation of the *cond_mutual_information* function takes as argument:

- `joint_distribution`, the joint probability distribution $P(\mathcal{X},\mathcal{Y},\mathcal{Z})$ between the random variables $\mathcal{X}$, $\mathcal{Y}$ and $\mathcal{Z}$;
- `joint_var1`, the joint probability distribution $P(\mathcal{X},\mathcal{Z})$ between the random variable $\mathcal{X}$ and $\mathcal{Z}$;
- `joint_var2`, the joint probability distribution $P(\mathcal{Y},\mathcal{Z})$ between the random variable $\mathcal{Y}$ and $\mathcal{Z}$;
- `distribution_cond`, the probability distribution $P(\mathcal{Z})$ of the conditional random variable $\mathcal{Z}$;

From these arguments, it computes the conditional mutual information $I(\mathcal{X};\mathcal{Y}|\mathcal{Z})$ using the previously defined *mutual_information*, with the following formula:

$$I(\mathcal{X};\mathcal{Y}|\mathcal{Z}) = \sum_k P(\mathcal{Z}_k)\ I(\mathcal{X};\mathcal{Y}|\mathcal{Z}_k) \tag{6}$$

3 Designing informative experiments

3.1 Entropy of a single square

In one square, one can place the digits 1 to 9. Therefore, each digit has an equal probability of $\frac{1}{9}$ of appearing in this square. By associating a random variable $\mathcal{X}$ to each square of the grid, the entropy of such a square is equal to :

$$H(\mathcal{X}) = -\sum_{i=1}^{9} \frac{1}{9}\ log_2(\frac{1}{9}) = -log_2(\frac{1}{9}) = 3.17$$

3.2 Entropy of subgrid

The subgrid, denoting by the random variable $\mathcal{Y}$, has 3 fields that are already filled with the digits 2, 3 and 8. One need to place the remaining digits 1, 4, 5, 6, 7, 9. By randomly selecting a first square, this square can take the value of one of the 6 digits, resulting in a probability of $\frac{1}{6}$. The next chosen square can now take one of the five remaining values, resulting in a probability of $\frac{1}{5}$, and so on until the subgrid is completely filled. The entropy of the subgrid is then the sum of the entropy of each square :

$$\begin{aligned} H(\mathcal{Y}) &= \sum_{i=1}^{9} H(\mathcal{X}_i) \\ &= -log_2(\frac{1}{6}) - log_2(\frac{1}{5}) - log_2(\frac{1}{4}) - log_2(\frac{1}{3}) - log_2(\frac{1}{2}) - log_2(\frac{1}{1}) \\ &= 9.492 \end{aligned}$$

3.3 Entropy of unsolved sudoku grid

Let $\mathcal{Z}$ be the random variable that denotes the completion of one sudoku grid. The entropy $H(\mathcal{Z})$ of the grid is the uncertainty of the initial unsolved sudoku grid. When we fill a cell denoted by $\mathcal{X}$ with a digit, the amount of information that is added can be measured by the mutual information $I(\mathcal{Z};\mathcal{X})$. Thus the remaining amount of uncertainty left in the grid can be calculated by the conditional entropy $H(\mathcal{Z}|\mathcal{X})$. Furthermore, let N be the number of empty cells in the grid and let $n(i,j)$ be the number of digits that can be added to the empty cell at coordinates (i,j) after having considered the subgrid, the row and the column where the cell belongs to.

The $n(i,j)$ are computed by the following formula :

$$n(i,j) = 9 - |R \cup C \cup S| \tag{7}$$

where R is the set of digits in row i, E is the set of digits in column C, S is the set of digits in the subgrid and $|\cdot|$ is the cardinality operator.

The probability of completing any grid is thus

$$P = \frac{1}{\prod_{i,j} n(i,j)}$$

and finally, the entropy is given by

$$\begin{aligned} H(\mathcal{Z}) &= log_2(\prod_{i,j} n(i,j)) \\ &= \sum_{i,j} log_2\ n(i,j) \end{aligned} \tag{8}$$

After computing this sum, we got that : $H(\mathcal{Z}) = 89.275$

3.4 Solving the grid

From N initial empty cells, if we fill a cell X^* with a digit, it will affect every cell in the corresponding subgrid, row and column. Let us denote by $n^*(i,j)$ the number of digits that can be filled in the remaining cells.

The remaining amount of entropy in the grid is

$$H(\mathcal{Z}|\mathcal{X}^*) = \sum_{i,j} log_2\ n^*(i,j) \tag{9}$$

and the mutual information is

$$\begin{aligned}\mathcal{I}(\mathcal{Z};\mathcal{X}^*) &= H(\mathcal{Z}) - H(\mathcal{Z}|\mathcal{X}^*)\\ &= \sum_{i,j} log_2\ n(i,j) - \sum_{i,j} log_2\ n^*(i,j)\\ &= \sum_{i,j} log_2\ \frac{n(i,j)}{n^*(i,j)}\end{aligned} \tag{10}$$

The number of empty cells is now $N-1$.

The remaining amount of uncertainty after N steps is simply equal to the difference of entropy of the initial grid and the sum of mutual information obtained at each step.

Let us denote the amount of uncertainty that is eliminated at each step by T. It is made up of two parts: the mutual information between the cell that is getting processed and the empty cells, denoted by I, and the decrease in entropy of the processed cell, denoted by ΔH. Moreover, let $m(i,j)$ and $m'(i,j)$ be the number of available digits of the other cells in the column, row and subgrid before and after the placement of the digit in cell $\mathcal{X}^*$. The following formulas details the first term above in formal notation:

$$I = \frac{1}{N}\left(\sum_{i,j} log_2\ m(i,j) - \sum_{i,j} log_2\ m'(i,j)\right) \tag{11}$$

The decrease in entropy of the processed cell is:

$$\Delta H(N) = log_2(N) - log_2(N-1) \tag{12}$$

and thus the total amount of information that is added when filling in a cell is the sum:

$$T(\mathcal{Y}^*) = I + \Delta H(N) \tag{13}$$

When attempting to solve the grid, we use the above framework to select the cell which is the most valuable to fill in. This is the one that will cause the greatest decrease in entropy of the grid.

3.5 Choosing additional clue

Revealing a clue is the same as filling in a cell. The best strategy is to choose the cell which has the highest mutual information with the empty cells. Formally,

$$argmax_{i,j}\ I = \frac{1}{N}\left(\sum_{i,j} log_2\ m(i,j) - \sum_{i,j} log_2\ m'(i,j)\right) \tag{14}$$

3.6 Strategy for sequential clues

This is the same as the question before, the subsequent clue you reveal is the one with the highest mutual information.

3.7 Strategy for concurrent clues

When the clues were revealed one after the other, one clue reveal didn't affect subsequent ones as the maximum of the mutual information was checked only after having revealed the first clue. Now however, when revealing two clues simultaneously, revealing one clue might change the value of the mutual information of the other. Thus, we must make sure that the first reveal has the smallest possible impact on the second (or more) reveals. If we denote by I_1 and I_2 the mutual information of the first and second reveal respectively, we must find the cells to reveal that maximize

$$I_1 + I_2 - \mathcal{I}(I_1; I_2) \tag{15}$$

Appendix

Probability distributions

	P(·)
x_1	1/4
x_2	1/4
x_3	1/4
x_4	1/4

Table 7: $P(\mathcal{X})$

	P(·)
y_1	1/2
y_2	1/4
y_3	1/8
y_4	1/8

Table 8: $P(\mathcal{Y})$

	P(·)
w_0	11/16
w_1	5/16

Table 9: $P(\mathcal{W})$

	P(·)
z_0	5/16
z_1	11/16

Table 10: $P(\mathcal{Z})$

Joint probability distributions

	x_1	x_2	x_3	x_4
w_0	1/8	1/8	3/16	1/4
w_1	1/8	1/8	1/16	0

Table 11: $P(\mathcal{X}, \mathcal{W})$

	y_1	y_2	y_3	y_4
w_0	3/8	1/8	1/16	1/8
w_1	1/8	1/8	1/16	0

Table 12: $P(\mathcal{Y}, \mathcal{W})$

	x_1	x_2	x_3	x_4
z_0	1/16	1/16	1/16	1/16
z_1	3/16	3/16	3/16	3/16

Table 13: $P(\mathcal{X}, \mathcal{Z})$

	y_1	y_2	y_3	y_4
z_0	1/8	1/8	1/16	0
z_1	3/8	1/8	1/16	1/8

Table 14: $P(\mathcal{Y}, \mathcal{Z})$

	x_1	x_2	x_3	x_4
y_1	1/8	1/16	1/16	1/4
y_2	1/16	1/8	1/16	0
y_3	1/32	1/32	1/16	0
y_4	1/32	1/32	1/16	0

Table 15: $P(\mathcal{X}, \mathcal{Y})$

	w_0	w_1
z_0	0	5/16
z_1	11/16	0

Table 16: $P(\mathcal{W}, \mathcal{Z})$

	x_1		x_2		x_3		x_4	
	w_0	w_1	w_0	w_1	w_0	w_1	w_0	w_1
y_1	0	0.125	0.0625	0	0.0625	0	0.25	0
y_2	0.0625	0	0	0.125	0.0625	0	0	0
y_3	0.03125	0	0.03125	0	0	0.0625	0	0
y_4	0.03125	0	0.03125	0	0.0625	0	0	0

Table 17: $P(\mathcal{X}, \mathcal{Y}, \mathcal{W})$

	x_1		x_2		x_3		x_4	
	z_0	z_1	z_0	z_1	z_0	z_1	z_0	z_1
y_1	0.125	0	0	0.0625	0	0.0625	0	0.25
y_2	0	0.0625	0.125	0	0	0.0625	0	0
y_3	0	0.03125	0	0.03125	0.0625	0	0	0
y_4	0	0.03125	0	0.03125	0	0.0625	0	0

Table 18: $P(\mathcal{X}, \mathcal{Y}, \mathcal{Z})$

	x_1		x_2		x_3		x_4	
	w_0	w_1	w_0	w_1	w_0	w_1	w_0	w_1
z_0	0	0.125	0	0.125	0	0.06125	0	0
z_1	0.125	0	0.125	0	0.1875	0	0.25	0

Table 19: $P(\mathcal{W}, \mathcal{Z}, \mathcal{X})$

References

[1] Team 2513, When Sudoku Comes across Information Theory http://act.buaa.edu.cn/jianlei/paper/MCM2008.pdf