ELEN060-2 - Information and coding theory

Project 1 - Information measures

February 2024

The goal of this first project is to get accustomed to information and uncertainty measures. We ask you to write a brief report (pdf format) collecting your answers to the different questions. All codes must be written in Python inside the Jupyter Notebook provided with this assignment, no other code file will be accepted. Note that you can not change the content of locked cells or import any extra Python library than the ones provided.

The assignment must be carried out by groups of two students. The report and the notebook should be submitted on Gradescope (https://www.gradescope.com/) before March 20 23:59 (CET). Note that attention will be paid to how you present your results and your analyses. By submitting the project, each member of a group shares the responsibility for what has been submitted (e.g., in case of plagiarism in the pdf or the code). From a practical point of view, every student should have registered on the platform before the deadline. Group, archive and report should be named by the concatenation of your student ID (sXXXXXX) (e.g., s000007s123456.pdf and s000007s123456.ipynb).

Implementation

In this project, you will need to use information measures to answer several questions. Therefore, in this first part, you are asked to write several functions that implement some of the main measures seen in the first theoretical lectures. Remember that you need to implement the functions in the Jupyter Notebook at the corresponding location, and answer the questions in the pdf file.

1. Write a function *entropy* that computes the entropy $\mathcal{H}(\mathcal{X})$ of a random variable $\mathcal{X}$ from its probability distribution $P_{\mathcal{X}} = (p_1, p_2, ..., p_n)$. Give the mathematical formula that you are using and explain the key parts of your implementation. Intuitively, what is measured by the entropy?
2. Write a function *joint_entropy* that computes the joint entropy $\mathcal{H}(\mathcal{X}, \mathcal{Y})$ of two discrete random variables $\mathcal{X}$ and $\mathcal{Y}$ from their joint probability distribution $P_{\mathcal{X},\mathcal{Y}}$. Give the mathematical formula that you are using and explain the key parts of your implementation. Compare the *entropy* and *joint_entropy* functions (and their corresponding formulas), what do you notice?
3. Write a function *conditional_entropy* that computes the conditional entropy $\mathcal{H}(\mathcal{X}|\mathcal{Y})$ of a discrete random variable $\mathcal{X}$ given another discrete random variable $\mathcal{Y}$ from their joint probability distribution $P_{\mathcal{X},\mathcal{Y}}$. Give the mathematical formula that you are using and explain the key parts of your implementation. Describe an equivalent way of computing that quantity.

4. Write a function *mutual_information* that computes the mutual information $\mathcal{I}(\mathcal{X};\mathcal{Y})$ between two discrete random variables $\mathcal{X}$ and $\mathcal{Y}$ from their joint probability distribution $P_{\mathcal{X},\mathcal{Y}}$. Give the mathematical formula that you are using and explain the key parts of your implementation. What can you deduce from the mutual information $\mathcal{I}(\mathcal{X};\mathcal{Y})$ on the relationship between $\mathcal{X}$ and $\mathcal{Y}$? Discuss.
5. Let $\mathcal{X}$, $\mathcal{Y}$ and $\mathcal{Z}$ be three discrete random variables. Write the functions *cond_joint_entropy* and *cond_mutual_information* that respectively compute $\mathcal{H}(\mathcal{X},\mathcal{Y}|\mathcal{Z})$ and $\mathcal{I}(\mathcal{X};\mathcal{Y}|\mathcal{Z})$ of two discrete random variables $\mathcal{X}$, $\mathcal{Y}$ given another discrete random variable $\mathcal{Z}$ from their joint probability distribution $P_{\mathcal{X},\mathcal{Y},\mathcal{Z}}$. Give the mathematical formulas that you are using and explain the key parts of your implementation. Suggestion: Observe the mathematical definitions of these quantities and think about how you could derive them from the joint entropy and the mutual information.

Predicting the outcome of an election campaign

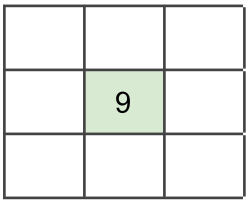
In the context of the approaching election, a campaign manager requires your expertise to leverage information theory to gain a strategic advantage in winning the upcoming election. To achieve this goal, the manager has provided a dataset derived from multiple past elections, encompassing a diverse set of information. This dataset, named "data.csv" (available on the course website) consists of samples collected from previous campaigns, each comprising 11 variables with distinct cardinalities (see Table 1, below). Your task is to analyze and apply information theory principles to uncover valuable insights from the dataset, ultimately contributing to the campaign's success. Include all your codes below the last cell of the Jupyter notebook (you may create several cells for better readability). Note that you have to answer the questions in the pdf report, including the numbers you get in the Notebook! The data is available on the website (data.csv).

6. Compute and report the entropy of each variable, and compare each value with its corresponding variable cardinality. What do you notice? Justify theoretically.
7. Compute and report the conditional entropy of *outcome* given each of the other variables. Considering the variable descriptions, what do you notice when the conditioning variable is (a) *weather* and (b) *previous_outcome*?
8. Compute the mutual information between the variables *target_demographic* and *budget*. What can you deduce about the relationship between these two variables? What about the variables *duration* and *reach*?
9. A student in Computer Science from the University of Liège bets his friends that he can predict the upcoming election by accessing the dataset. However, his hacking skills are still weak. Therefore, he can only access a single variable of the dataset to make its prediction. Using only the mutual information, which variable should he choose to get? Would using conditional entropy lead to another choice?
10. With the outcome of the previous campaign considered as known, would you change your answer from the previous question? What can you say about the amount of information provided by this variable? Compare this value with previous results.

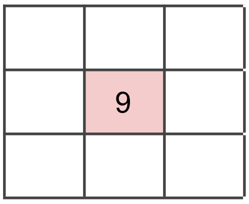
Playing with information theory-based strategy

Let us first consider a fictional game in which players have to guess all numbers ($\mathcal{N}$, from 1 to $\mathcal{N}$) of a $\mathcal{C}$ x $\mathcal{C}$ grid. The same number may appear several times. At each turn, the player chooses to guess a single number in one of the fields (or cells). The game then lets the player know if the number is in the correct spot (green), if it is not in the grid (red), or if it is in the grid but at the wrong spot (orange).

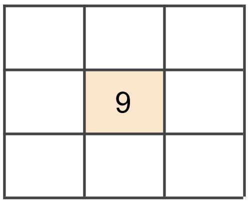
11. In this simple version of the game, what is the entropy of each of the fields for a grid with $\mathcal{C}$=3 and $\mathcal{N}$=9? Also, what is the entropy of the whole game (the 9 fields combined) ? How are these two quantities linked? Justify.
12. In this simple version of the game with $\mathcal{C}$=3 and $\mathcal{N}$=9, let us assume that your first guess is a 9 in the center. What is now the entropy of each field, and the entropy of the game at this stage for each of the three possible outcomes (red, green, orange)? How much information has this guess brought you in each case (in bits)?



(a)

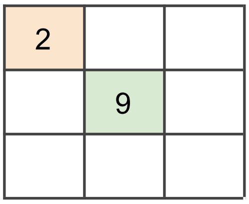


(b)



(c)

13. In this simple version of the game with $\mathcal{C}$=3 and $\mathcal{N}$=9, let us now assume that your second guess is a 2 in the top-left corner. What is now the entropy of each field, and the entropy of the game at this stage for the following outcome? How much information has this second guess brought you (in bits)?



(d)

Let us now consider an advanced game in which the numbers in the grid to guess follows the same constraints as the Sudoku game (i.e., the same number can not appear twice on the same row, the same column, and the same sub 3x3 grid).

14. In this advanced version of the game with $\mathcal{C}$=3 and $\mathcal{N}$=9, let us assume that your first guess is a 9 in the center and the second guess is a 2 in the top-left corner. What is now the entropy of each field, and the entropy of the game at every stage (including the initial stage) for the following outcome? How much information have these guesses brought you in this case (in bits)?

2		
	9	

(e)

15. In this advanced version of the game with $\mathcal{C}$=9 and $\mathcal{N}$=9, what is the entropy of an empty grid? How does it compare to the grid in the simple version?
16. Propose and discuss an approach based on information theory that would let you solve the real game in a minimum number of guesses (without actually solving the game). In particular, explain how you would choose your next guess based on the information you have.

	variable name	Possible values
1	*target_demographic*	{18-30, 31-60, 61+}
2	*budget_allocation*	{50 000, 250 000, 1 000 000}
3	*reach*	{low, medium, high}
4	*conversion_rate*	{low, medium, high}
5	*duration*	{short, medium, long}
6	*time_of_year*	{winter, spring, summer, fall}
7	*content_type*	{video, image, text}
8	*platform_used*	{Instagram, Facebook, Twitter, Youtube, TikTok, others (newspaper,...)}
9	*weather*	{sunny, rainy, snowy}
10	*previous_outcome*	{success, fail}
11	*outcome*	{success, fail}

Table 1: List of the variables and their discretized possible values.