ELEN060-2 - Information and coding theory

Project 1 - Information measures

February 2025

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 19 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 result of the Information and Coding Theory exam

In the context of the future session in June, you would like to know your chances of success for the exam. Therefore, you collected a dataset of various observations from previous years. This dataset contains 11 columns and 5000 rows, including variables that could be useful to predict whether you will succeed in the exam or not (see the table at the end of the assignment). Information theory provides a good framework to analyze the links between the different variables and the chances of success or failure of the exam.

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 *Exam result* given each of the other variables. Considering the variable descriptions, what do you notice when the conditioning variable is (a) *Interest in the course* and (b) *master*?
8. Compute the mutual information between the variables *location* and *Evalens score of the course*. What can you deduce about the relationship between these two variables? What about the variables *Time spent on the project* and *project grade*?
9. A student in Computer Science from the University of Liège bets his friends that he can predict the upcoming exam 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 *interest in the course* 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

Mastermind is a code-breaking game that involves two players, one who creates a secret code and the other who tries to guess the code. Let's consider a simplified version, in which the game is played on a board with a series of slots, and each slot can be filled with a colored peg. The colors used in the game are typically chosen from a limited alphabet, such as red, blue, green, yellow, and brown, (which will be represented by numbers from 1 to 5).

The player who creates the secret code chooses a combination of 5 colors and places them in the slots on the board. The same color can be placed at several spots. The player trying to guess the code then places pegs of different colors in the slots, trying to match the colors of the secret code. After each guess, the player who created the code provides feedback by revealing which pegs are correctly placed in the guess and which pegs are in the code, but at the wrong position (*i.e.*, an incorrectly placed peg).

The goal of the game is for the player trying to guess the code to correctly deduce the secret code in as few guesses as possible. This requires careful analysis of the feedback provided after each guess, as well as strategic thinking and planning. In the following, we consider that the probability distribution of the secret codes is uniform.

11. Given a set of 5 possible colors for the pegs and 5 slots in the Mastermind game, what is the entropy of each of the 5 slots ? Also, what is the entropy of the whole game (the 5 slots combined) ? How are these two quantities linked? Justify.
12. Let us assume that your first guess gives you the following result. What is now the entropy of each field, and the entropy of the game at this stage? How much information has this guess brought you (in bits)?

1	2	4	2	3

1 correctly placed peg at the first position

13. Now let us assume that the same first guess gives you the following result. What is now the entropy of each field, and the entropy of the game at this stage? How are these two quantities linked? Justify. How much information has this guess brought you (in bits)? Finally, compare this gain to the one of the previous question and explain.

1	2	4	2	3

1 correctly placed peg at the first position and 1 incorrectly placed peg at the second position

14. Given a certain number of possible colors ($\mathcal{C}$) and a certain number of slots ($\mathcal{S}$) in the game board, express the formula of the maximum entropy of the system. How does the number of colors and the number of slots affect the maximum entropy?
15. Propose and discuss an approach based on information theory that would let you solve the game in a minimum number of guesses. In particular, explain how you would choose your next guess based on the information you have.

	Variable name	Possible values
1	*Exam result*	Success, Failure
2	*Grade for the probability class*	0-50, 50-80, 80-100
3	*Project grade*	90-100, 70-90, 50-70, 0-50
4	*Time spent on project*	Less than 2h, 2-5h, 5-10h, 10-20h, More than 20h
5	*Time spent studying*	Less than 2h, 2-5h, 5-10h, 10-20h, More than 20h
6	*Interest in the course*	Interested, Neutral, Not interested
7	*Weather the week before the exam*	Sun, Rain
8	*Date*	Beginning of the exam session, Middle of the exam session, End of the exam session
9	*Location*	R3, R7, Math amphi
10	*Master*	Data science, Computer science, Electricity engineering
11	*Evalens score of the course*	Good, Medium, Bad

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