ELEN060-2 - Information and coding theory

Project 1 - Information measures

February 2023

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 15 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 a football game

Let's assume that the coach of a football team has kept track of previous match data to improve his team's performance through statistical analysis. This database is composed of 13 discrete variables described in Table 1. Note that these variables have different cardinalities (i.e., the number of possible values they can take). Using the database provided with this assignment, where each sample corresponds to a set of 13 values related to a previous game, answer the following questions. 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) *wind_speed* and (b) *previous_outcome*?
8. Compute the mutual information between the variables *month* and *capacity*. What can you deduce about the relationship between these two variables? What about the variables *day* and *time*?
9. Let's assume that you have decided to place a bid on the outcome of the match, but the data is now only available through a paid service. With limited funds, you must choose a single variable to invest in. Based on the mutual information, which variable would you keep? Would you make another choice if it was based on the conditional entropy?
10. With the outcome of previous matches between the same opponent now being available for free, would you change your answer? What can you say about the amount of information provided by this variable? Compare this value with previous results.
11. Using information theory, discover the particularity of the stadium of the home team, in particular using the *stadium_state* and *weather* variables. Justify with computations.

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. 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, brown, and black (which will be represented by numbers from 1 to 6).

The player who creates the secret code chooses a combination of 4 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 in the form of black and white pegs. A black peg is placed for each correctly placed colored peg in the guess, and a white peg is placed for each correctly colored peg that is in the wrong position.

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.

12. Given a set of 6 possible colors for the pegs and 4 slots in the Mastermind game, what is the entropy of each of the 4 slots ? Also, what is the entropy of the whole game (the 4 letters combined) ? How are these two quantities linked? Justify.
13. 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

1 correctly placed peg

14. 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

1 correctly placed peg 1 incorrectly placed peg

15. Given a certain number of possible colors ($\boldsymbol{C}$) and a certain number of slots ($\boldsymbol{S}$) 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?

16. 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
0	*outcome*	{win, loss, tie}
1	*previous_outcome*	{win, loss, tie}
2	*day*	{monday, tuesday, wednesday, thursday, friday, saturday, sunday}
3	*time*	{morning, afternoon, evening}
4	*month*	{january, february, march, april, may, june, july, august, september, october, november, december}
5	*wind_speed*	{no_wind, low, high}
6	*weather*	{sunny, cloudy, rainy, snowy}
7	*location*	{away, home}
8	*capacity*	{small, medium, large, elite}
9	*stadium_state*	{wet, dry}
10	*injury*	{yes, no}
11	*match_type*	{competitive, friendly}
12	*opponent_strength*	{weak, average, strong}

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