ELEN060-2 - Information and coding theory

Project 1 - Information measures

February 2021

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 16 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.

Weather forecasting

Let us assume that you are in charge of a weather station that aims at forecasting the volume of rain for the next day based on the current day's measures. Over the past year, you have collected several samples from your instruments in a database. The database is composed of 14 discrete variables (described in Table 1), 13 of them refer to discretized measures from the instruments (temperature, wind speed, relative humidity,...) for a particular day, and the remaining one corresponds to the discretized volume of rain that fell the next day. Note that these variables have different cardinalities (i.e., the number of possible different values). Using the database provided with this assignment (where each sample corresponds to a set of 14 values), 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 (weather_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 *next_day_rain* given each of the other variables. Considering the variable descriptions, what do you notice when the conditioning variable is (a) *wind direction* and (b) *same_day_rain*?
8. Compute the mutual information between the variables *relative_humidity* and *wind_speed*. What can you deduce about the relationship between these two variables? What about the variables *month* and *temperature*?
9. Let us assume that you need to set up a new weather station in a new location to forecast the *next_day_rain*. Due to budget cuts, you are asked to make the forecast based on a single instrument, *i.e.* knowing only the value of one variable. Based on the mutual information, which variable would you keep? Would you make another choice if it was based on the conditional entropy?
10. Would you change your answer if you should consider only the samples with the volume of rain the next day corresponding to either *deluge* or *drizzle* ? Justify.
11. Let's assume you are given a thermometer for free and that you may still buy an extra instrument. Would you change your answer? Justify.

Playing with information theory-based strategy

Wordle (https://www.powerlanguage.co.uk/wordle/) is a recent word game where players have six attempts to guess a five-letter word (26 letters, from a to z). After each guess, the program indicates which letters are in the correct spot (green), are in the word but at the wrong spot (orange) or are not in the word (gray). Let us use information theory to play this game efficiently. In what follows, for sake of simplicity, we may consider simplistic assumptions that do not exactly correspond to the real game but will be necessary, for example, to compute information measures without actually solving the game (which requires a database of words).

Let us first play a simple version of the game where the word to guess is chosen at random with a uniform distribution for each letter (ranging from *"aaaaa"*, *"aaaab"*,..., *"zzzzz"*). Note that in this simplistic version, your guesses don't have to match an existing English word unlike the real game.

12. In this simpler version of the game, what is the entropy of each of the 5 fields ? Also, what is the entropy of the whole game (the 5 letters combined) ? How are these two quantities linked? Justify.
13. In this simpler version of the game, 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)?

14. In this simpler version of the game, let us now assume that your second 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.

Let us now consider the real game as hosted on the website (not the simplified version described previously), where the objective word is randomly chosen among a predefined dictionary of 2,000 words and the possible guesses you can make have to match a second dictionary of 12,000 words (you can not for instance guess "*aeiou*" since it is not a valid word). The objective words are all equally probable among the 2,000 word dictionary, however each letter (probably) follows the English distribution.

15. What can you expect from the entropy of the real game compared to the simplified version? Justify.
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, which would require a database). In particular, explain how you would choose your next guess based on the information you have.

Instrument	variable name	Possible values
Thermometer	*temperature*	{freezing,cold, medium, high}
barometer	*air_pressure*	{increasing, decreasing}
Rain gauge	*same_day_rain*	{dry, drizzle, deluge}
Rain gauge	*next_day_rain*	{dry, drizzle, deluge}
Sling Psychometrer	*relative_humidity*	{low, high}
Wind vane	*wind_direction*	{north, south, east, west}
Anemometer	*wind_speed*	{no_wind, low, high}
Ceilometer type 1	*cloud_height*	{no_cloud, low, high}
Ceilometer type 2	*cloud_density*	{no_cloud, low, high}
Monthly calendar	*month*	{january, february, march, april, may, june, july, august, september, october, november, december}
Daily calendar	*day*	{monday, tuesday, wednesday, thursday, friday, saturday, sunday}
Campbell Stokes Recorder	*daylight*	{sunny,cloudy}
Lightning Detector	*lightning*	{no_lightning, low, high}
Nephelometer	*air_quality*	{bad, medium, good}

Table 1: List of instruments, their associated variables and their discretized possible values.