ELEN0060 - Information and coding theory
Project 1 - Information measures

February 2021

The goal of this first project is to get accustomed to the information and uncertainty measures. We ask you to write a *brief* report (*pdf* format) collecting your answers to the different questions.

The assignment must be carried out by group[1] of *two students* and the report and the scripts should be submitted as a *tar.gz* or *zip* file on Montefiore's Submission platform (`http://submit.montefiore.ulg.ac.be`) before **March 14 23:59**. 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).

From a practical point of view, every student should have registered on the Submission platform and have joined a group **before** the deadline. Group, archive and report should be named by the concatenation of your student ID (sXXXXXX) (*e.g.*, s000007s123456).

Implementation

In this project, you will need to use information measures to answer several questions. Therefore, in this first part, let us start this project by writing several functions (in *python* or *julia*) that implement some of the main measures seen in the first theoretical lectures.

1. Write a function *entropy* that computes the entropy $H(\mathcal{X})$ of a random variable $\mathcal{X}$ from its probability distribution $P_{\mathcal{X}} = (p_1, p_2, \ldots, p_n)$. 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 $H(\mathcal{X}, \mathcal{Y})$ of two discrete random variables $\mathcal{X}$ and $\mathcal{Y}$ from the 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 $H(\mathcal{X}|\mathcal{Y})$ of a discrete random variable $\mathcal{X}$ given another discrete random variable $\mathcal{Y}$ from the conditional probability distribution $P_{\mathcal{X}|\mathcal{Y}}$ and the marginal probability distribution $P_{\mathcal{Y}}$. 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 $I(\mathcal{X}; \mathcal{Y})$ between two discrete random variables $\mathcal{X}$ and $\mathcal{Y}$ from the marginal probability distributions $P_{\mathcal{X}}$ and $P_{\mathcal{Y}}$ and 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 $I(\mathcal{X}; \mathcal{Y})$ 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 functions *cond_joint_entropy* and *cond_mutual_information* that respectively compute $H(\mathcal{X}, \mathcal{Y}|\mathcal{Z})$ and $I(\mathcal{X}; \mathcal{Y}|\mathcal{Z})$. Give the mathematical formulas that you are using and explain the key parts of your implementation (in particular which probability distributions you are using as inputs of your functions).
Suggestion: Observe the mathematical definitions of these quantities and think how you could derive them from the joint entropy and the mutual information.

[1] See instructions on `https://people.montefiore.uliege.be/asutera/ICT.php`.

Medical diagnosis

Let us then consider a biomedical application that aims at identifying the nature of an illness by examination of the symptoms and some medical parameters. In our problem[2], we have 17 discrete variables (described in Table 1), 16 of them refer to several parameters of interest regarding the disease, and the remaining one refers to the disease. Note that these variables have different cardinalities (*i.e.*, the number of possible different values). Using the given set of samples (where each sample corresponds to a set of 17 values), answer the following questions.

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 the disease given each of the other variables. Considering the variable descriptions, what do you notice when the conditioning variable is (a) *jaundice* and (b) *bilirubin*?

8. Compute the mutual information between the variables *obesity* and *age*. What can you deduce about the relationship between these two variables?

9. Let us assume that you need to make the diagnosis 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 disease value corresponding to healthy or steatose? Justify.

11. Would you change your answer based on the age of the patient? Justify.

Playing with information theory-based strategy

The minesweeper is a very famous puzzle video game. The objective is to clear a rectangular board containing hidden *mines*, without detonating (*i.e.*, clicking on it) any of them, with the help from clues about the number of neighboring mines in each field (*i.e.*, square or cell)[3]. The neighboring of a field is an adjacent field (either horizontally, vertically, or in diagonal). A field is either *revealed* (*i.e.*, either a mine or a clue, and unclickable) or *unrevealed* (and clickable). Let us use information measures to play to this game. In what follows, for sake of simplicity, we may consider simplistic assumptions that do not exactly correspond to the real game (or the game you know) but will be necessary, for example, to compute information measures without actually solving the game (which might be very complicated). Please be sure to consider the proper game setting in your answers.

Let us model the game as follows. Let us consider the board as a matrix where each element (i, j) is a field (and where $(0, 0)$ corresponds to the field at the upper left corner). Let us associate a binary random variable $\mathcal{X}$ to each field, and denote $\mathcal{X}_{i,j}$ the random variable of field (i, j). $X_{i,j}$ is equal to 1 if there is a mine on the field (i, j) and 0 otherwise. To each unrevealed field corresponds a probability distribution $P_{i,j} = [p^0_{i,j}, p^1_{i,j}]$ where $p^0_{i,j}$ and $p^1_{i,j}$ are the probability of $X_{i,j}$ being equal to 0 and 1 respectively. To each clue can be associated a uniform distribution where the value of the clue (*i.e.*, the number of mines in adjacent unrevealed fields) is evenly distributed among all adjacent unrevealed fields. The entropy of a clue is therefore the entropy of a uniform distribution where each probability is $p_i = \frac{value\ of\ the\ clue}{N}$ with $i = 1, \ldots, N$ and N being the number of unrevealed adjacent fields.

12. Let us consider a board of R rows and C columns. At the start of a new game and assuming that there is no revealed field and M hidden mines, what is the entropy of each field?

13. What are the entropies of each unrevealed field (adjacent to a clue) in the game situation represented by Figure 1? What do you notice?

[2] Inspired from HEPAR dataset: Onisko, A., Druzdzel, M. J., & Wasyluk, H. (1999, September). A Bayesian network model for diagnosis of liver disorders. In *Proceedings of the Eleventh Conference on Biocybernetics and Biomedical Engineering* (Vol. 2, pp. 842-846).

[3] As described on the Wikipedia page.

0	1	2	3	2
1	2			
1				

Figure 1: Case 1

In what follows, let us assume that you can compute any information measure (*e.g.*, given by an oracle that knows every probability distribution and computes information measures for you) for unrevealed fields that are adjacent to a least one clue.

14. Based on information measures, how would you choose the next field to reveal? In particular, specify which information measure(s) you would use and why.

15. Let us assume that you reveal the fields one by one. Based on your answer to the previous question, design a strategy that would only rely on information measures to play this game from the start to the end.

16. Let us assume that you must reveal k fields simultaneously at each turn. Based on your answer to the previous questions, design a strategy that would only rely on information measures to play this game from the start to the end. In particular, compare it with the one proposed at the previous question.

17. Propose and discuss an approach that would let you use information theory to play the game without actually solving the game (*i.e.*, finding everywhere the mines from the clues or using an oracle). Your proposition can take the form of additional assumptions and/or additional rules. Note that your proposition should follow the same goal as the original game (*i.e.*, revealing the board without detonating a mine).

18. [**BONUS: no extra point**] You can implement the approach proposed in the previous question by modifying the provided code if you wish.

Name	Abbr.	Possible values	Description (if not obvious)
Age	-	$\{\leq 40, > 40\}$ (in years)	
Sex	-	{man, woman}	
Obesity	-	{thin, regular, overweigh}	
Alcoholic antecedents	ALC	{yes, no}	
Iron	-	{low, normal, high, very high}	*The iron rate in the blood.*
Disease	DIS	{healthy, PBC, steatosis }	*The state of the patient: healthy, or affected by hepatic steatosis, or by primary biliary cirrhosis (PBC).*
Fatigue	-	{yes, no}	
Triglycerides	TRI	{abnormal, normal}	*The triglycerides rate in the blood. Triglycerides are the main constituents of fat.*
Alanine transaminase	ALT	{abnormal, normal}	*The ALT rate in the blood. The ALT is an enzyme mainly present in the liver.*
Aspartate transaminase	AST	{abnormal, normal}	The ALT rate in the blood. The ALT is an enzyme mainly present in the liver.
Gamma-glutamyl transpeptidase	GGTP	{abnormal, normal}	*The GGT rate in the blood. The GGTP is an enzyme.*
Cholesterol	CHL	{low, normal, high}	*The cholesterol rate in the blood.*
Anti-mitochondrial antibody	AMA	{yes, no}	
Muscular ache	MSC	{yes, no}	
Bilirubin	BIL	{abnormal, normal}	*The bilirubin rate in the blood. The bilirubin is a yellow pigment.*
Itching	ITC	{yes, no}	
Jaundice	JAU	{yes, no}	

Table 1: List of variables