Take Home Exam for Statistical Information Theory MT7037 VT19

Joanna Tyrcha, Chun-Biu Li Mar 9, 2019

*****Please read carefully the following instructions before starting**.

- Deadline: **6pm on March 24, 2019**.
- Your report as a single ZIP file should contain the **signed confirmation sheet**, **your answers to the questions** (in pdf format) and **the source code for exercise 4** (in pdf format). The ZIP file should be sent to **cbli@math.su.se** before the deadline. Submission of the report after the deadline will not be accepted.
- Obtainable grades are A (≥ 90 points), B (≥ 80 points), C (≥ 70 points), D (≥ 60 points), E (≥ 50 points), Fx and F. Note that we reserve the right to rescale the points for grades by a factor of c $(0.9 < c < 1)$ stated here depending on the outcome of the exam.
- We will only grade the exam of LADOK registered students (exception are Ph.D. students) - hence make sure that you are registered in LADOK before the submission deadline!
- The credit points will be given based on clear logical explanation and steps leading to the final solution, so do not simply state the final answer.
- The report must be completed independently. Plagiarism or other forms of cheating is a serious act - to underline this your report must as cover page contain the signed "confirmation sheet" that your work is made in accordance with the rules for written exams at Stockholm University.

Exercise 1 (28 points)

(a) (6 points) Let X and Y be discrete random variables. Show that $H\left((X+Y)\middle|X\right) = H(Y|X)$. Argue that if X and Y are independent then $H(X+Y) \geq H(Y)$.

(b) (10 points) Show that among all N-valued random variables (i.e. with values $k = 1,2,\ldots$) with expected value μ, the geometric distributed random variable with expected value μ has the maximum value of Shannon entropy.
Reminder: The probability function in a geometric distribution is

$$p(k) = p(1-p)^{k-1}, \quad k = 1,2,\ldots$$

(c) (6 points) Show that

i) (3 points) $H(X_1,X_2,X_3) \le \frac{1}{2}\left[H(X_1,X_2)+H(X_2,X_3)+H(X_1,X_3)\right]$

ii) (3 points)

$$H(X_1,X_2,X_3) \ge \frac{1}{2}\left[H(X_1,X_2|X_3)+H(X_2,X_3|X_1)+H(X_1,X_3|X_2)\right]$$

(d) (6 points) Suppose that (X,Y,Z) are jointly normal distributed and that $X \to Y \to Z$ forms a Markov chain. Let X and Y have the correlation coefficient ρ_1 and let Y and Z have the correlation coefficient ρ_2. Find mutual information $I(X;Z)$.

Exercise 2 (22 points)

(a) (5 points) Define the transfer entropy, starting from Schreiber's definition in *Phys. Rev. Lett.,* 85, 461, 2000, in terms of mutual information and Shannon entropy.

(b) (2 points) Why is mutual information not a good measure for an information transfer?

(c) (2 points) Does Schreiber's definition of transfer entropy coincide with the definition of mutual information?

(d) (1 point) What is the meaning of local transfer entropy?

(e) (3 points) Is the Granger causality concept the same as transfer entropy? Explain possible similarities and differences.

(f) (9 points) To complete this part you need to read the paper by Lizier and Prokopenko in *Eur. Phys. J. B* 73, 605-615, 2010.

i) (3 points) Does information transfer have the same meaning as information flow in the opinion of the authors of the paper? Explain possible similarities and differences.

ii) (1 point) What is the difference between interventional and standard conditional probabilities?

iii) (2 points) Explain the difference between local transfer entropy and local information flow.

iv) (3 points) What are the advantages of the definition of local information flow according to the authors of the paper?

Exercise 3 (15 points)

To understand better the concepts of multi-information and its decomposition, this exercise requires you to work out some details in the paper by Schneidman et al., *Phys. Rev. Lett.*, 91, 238701, 2003 discussed in the class.

(a) (7 points) For the case of three binary variables (i.e., x_1, x_2, x_3 equal to either 0 or 1) with all pairwise marginals known, derive with clear steps the maximum entropy distribution in terms of the Lagrange multipliers. Hint: You do not need to solve for the Lagrange multipliers and your answer should look like Eq. 1, namely, the Ising model, in Schneidman et al., *Nature*, 440, 1007, 2006.

(b) (4 points) Can you derive Eq. (8) of the paper and correct the typos? Hint: Use the Venn diagram. This exercise shows that 1) although the area-information correspondence of the Venn diagram does not hold in the 3-variable case, the Venn diagram is still useful when deriving relations between different information quantities; 2) mistakes exist in published papers, so always be critical in reading them.

(c) (4 points) Show that the connected information of order k in Eq. (6) of the paper can be written as a relative entropy. When does it equal to zero?

Exercise 4 (35 points)

This exercise allows you to experience the performance of the rate distortion theory in clustering problem. By completing this exercise, you will have your own code of nonparametric information-based clustering. To start with, you need to first download the data file (Data_Exercise4_2019) from the moodle course page that contains 300 data points to be clustered. In the file, the first and second columns are the x- and y-coordinates of the data points, respectively. Moreover, the first 100 and the next 200 points are independently sampled from the normal distributions, $N\left(\mu = (0 \quad 1), \Lambda = \begin{pmatrix} 0.25^2 & 0 \\ 0 & 0.25^2 \end{pmatrix}\right)$ and $N\left(\mu = (1 \quad 0), \Lambda = \begin{pmatrix} 0.4^2 & 0 \\ 0 & 0.4^2 \end{pmatrix}\right)$, respectively.

(a) (2 points) Plot the data points on the x-y plane to see how they look like and prepare the element-to-element distance matrix $d(\vec{x}_i, \vec{x}_j) = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}$ $(i, j = 1,2, \cdots, 300)$ that will be the input for the clustering algorithm. Also assume each data point carries the same weight, i.e., $p(\vec{x}_i) = 1/300$ for all i.

(b) (20 points) Write a code (no restriction on the program language) to implement the Blahut-Arimoto algorithm discussed in the class to evaluate the clustering membership probability, $p(\widetilde{\boldsymbol{x}}|\boldsymbol{x})$, with fixed number of clusters, N_c, and compression-distortion tradeoff parameter, β. Your code should implement a multiple run each starting with random initial conditions. **Note:** Your source code should include clear comments/documentations to describe what is evaluating. We may later randomly ask a few students, especially those without clear documentations, to demonstrate how their code works.

(c) (7 points) Run your program to construct the information curves for N_c = 2, 3 and 4. Hint: choose different values of β in between 1 to 50.

(d) (6 points) As we have already known that the correct number of clusters is 2, propose a reasonable way using the quantities evaluated from your code (e.g.

$I(\tilde{\boldsymbol{x}}, \boldsymbol{x})$, $\langle d(\tilde{\boldsymbol{x}}, \boldsymbol{x})\rangle_{p(\tilde{x},x)}$, the Lagrange function, etc.) and the information curves to correctly identify the number of clusters. You should clearly explain your rationale and show explicitly which quantities, graphs and/or curves are used in the identification.

~ Good Luck ~