Modern Statistics and Big Data Analysis - Exam - 15 January 2024

The exam consists of four questions. The first question is the literature question below, which has to be done and submitted *before* the exam, see rules below.

The other three questions will be given out on the exam day and should be done then. As in earlier exams, there will be one question where you analyse data, one theoretical question, and one question where you are asked to interpret data analysis output. The literature question will have a weight of 25%.

Literature question

On Moodle you can find the article "Regularized k-means clustering of high-dimensional data and its asymptotic consistency" by Wei Sun and Junhui Wang, Electronic Journal of Statistics 6 (2012), 148–167.

Keep in mind that it is quite normal to not understand everything in a reseach paper. There are details in this paper that you cannot understand based on just what you have learnt in courses, but you will realise that these are not important to understand for answering the questions. The ability to get from the paper what you need without being affected by the things you do not understand is a key competence.

(a, 10%) Explain in your own words how regularized k-means and regularized model-based clustering work (you don't need to explain details of the algorithm), and for what reasons the differences to standard k-means, model-based clustering (Gaussian mixtures), respectively, have been introduced.

(b, 4%) The authors state that the X-variables should be "centralized"[1]. Why is this important? Do you think it would also be useful to standardize them to unit variance? Why, or why not, or under what circumstances?

(c, 6%) As opposed to the use of the Lasso in regression, the tuning constant λ here cannot be chosen by optimizing a prediction error estimated by cross-validation. Why not? What do the authors propose instead to choose the λ? Do you think this could also be done involving the Rand or adjusted Rand index? Why or why not?

(d, 5%) What are advantages and good properties of the method according to the authors? Do you think this is convincing? Can you think of any potential disadvantages of the method?

Submission rules for the literature question

Your answer to the literature question must be a pdf-file. It should be at most three typed pages long (with reasonable letter size and printing area, all four parts included). It should be sent by email to `christian.hennig@unibo.it` by Monday 15 January, 11:00 (beginning of the exam). Late submissions will not normally be accepted, so in order to make sure that this arrives in time, do not leave it to the last moment.

Obviously I cannot control whether you collaborate before the exam on the literature question. However you are required to use your own words. Solutions and wordings that give strong evidence that they were prepared in collaboration with fellow students will lead to a failure of the exam with zero marks (I may deviate from this if the evidence is too weak). The same holds for the use of chatbots. I will run the question through a chatbot myself several times to see how such answers look like. In other words, you can discuss the article in advance, however it is forbidden to collaborate with fellow students or chatbots or anyone else when it comes to the actual writing.

[1] They mean "centered"; the standard of English language writing in this paper is not good.