KSD DM mandatory

leod

1 Mandatory assignment

The mandatory assignment will take the form of a micro-scale data mining project.

You will use the data set we created during the first lecture, through all of you filling in a questionnaire. You will formulate one or several questions to be answered based on this data, write the code and carry out the analysis necessary to answer them.

You must apply at least two different pre-processing methods, one clustering method and one supervised learning method (Naïve Bayes and ID3 fit in this category). (For example, normalization+missing value replacement+Apriori+k-means+ID3, though many other combinations are possible.) All code for these algorithms must be written by yourself. Source code will only be accepted if it's reasonably well commented.

What to hand in: A two-page report (1 page = "2.400 [...] units per page, including spaces and notes.") as a .pdf-file, in which you describe what questions you tried to answer, which methods you used, the results you reached, any problems you encountered along the way (including the dirtiness of the dataset) and, if applicable, any other observations. Please include numerical results and preferably also graphs. Additionally, you must hand in all the source code you've written for the assignment. If your report physically uses more than two pages (e.g. due to layout or graphs etc.) please include the total unit count (including spaces and notes) in the report.

The assignment is a natural continuation of the work done during the labs. You will be able to reuse the code you wrote during the labs, and even some data set analysis you performed. If you completed all the labs, you will have most of the work on the assignment. Therefore, use the lab sessions effectively, and don't be afraid of asking the TAs and teacher questions!

The assignment will only be marked as a pass/fail. A successful assignment will not affect the final grade in any way. If you fail the assignment (or didn't hand in on time) you will be able to resubmit the assignment after the group project deadline. However, the resubmission will require you to implement more algorithms and the deadline will interfere with your exam preparation. Furthermore, to be able to go to the final exam you must pass the resubmission.

The assignment is individual, which means that you are supposed to formulate the questions, write the code, do the experiments and write the report yourself. Plagiarism will not be tolerated and will result in the plagiarist failing the assignment. On the other hand, minor technical errors will not be a cause for failure, at least as long as they are acknowledged within the report.

Due on the Tuesday of week 44 at 14:00.

Handing in the assignment on time is mandatory.