March 31, 2020

The Essay of Data Mining instead of Passing the Examination

The task is to write an essay, the topic of which is one of those given at the end of this text. The prerequisite is that a student (author in the following instructions) has made at least 30% of the weekly exercises of the Data Mining course in January and February 2020.

The length has to be from 10 to 12 pages. The pages contain page numbers. The text has to be on the basis of articles that the author searches for from internet of collections of Tampere University library. These are scientific computers science or related field articles or books, but not from BSc. or MSc. theses. Nothing is allowed to be copied from the course material of Data Mining or directly anywhere. The author has to read the articles, understand their parts of texts that are relevant and then write, "with one's own words", the methods to be presented in the essay. This means that it is neither allowed to directly copy sentences from the original texts nor to copy and then to slightly vary them. Of course, conventional scientific terms are used freely (using them is not copying). No computational test results given in articles can be used as such, because they are research results of someone else, not those of the one writing the essay. However, they can be characterized briefly. **The purpose is to describe some relevant computational methods or algorithms subject to a given topic.** It is not allowed to directly copy each of such algorithms exactly (in the sense of writing) from the source article. The author uses the same concepts as in the source article, but so that the author uses similar forms for concept names - say 'input vector **x**', or 'matrix **Z**' or 'variable *y*' - throughout all methods due to be presented. This means that the author does not copy the presentation form directly from one article and then differently the form of the same concept from another article for the next method in the essay, but chooses a suitable way at the beginning and then keeps this for all the following methods included in the essay. Naturally, there may appear a new concept in some later method to be described that was not used in the preceding methods.

Copy-paste would be equal to plagiarism and is prohibited – when detected the essay is failed. If plagiarism is suspected in a text, this can be checked and identified by the teacher.

The font due to be used is 11 and line spacing 1.5. The final text has to be printed in the pdf format.

The text has to be written in the form of a scientific article: the title, author's name, abstract less than ½ page, introduction of 1 or 2 pages, main text (methods), summary from ½ page to 1 page and references. If the author draws figures, their maximum number is six. The maximum size of a figure is ½ page. Any figure has to be prepared by the author her/himself – no copy is allowed. The references have to be those articles or books that the author used for preparing the essay, nothing else, but all those have to be included in the reference list the author used as sources. They have to be written according to some normally applied scientific way. For example, the author refers to [1] or Smith, 2010. The references of journal articles, congress articles and book texts are written in some systematic ways that the author can find from scientific articles and that are partially different for journal and congress articles and books.

The evaluation of an essay text is from 12 to 30 scores (similarly to examinations) on the basis of how fluent, clear, systematic and correct it is. It is essential that the author has understood what has written. Do not write such a method that you did not understand well. Scores (from 0 to 5) from the author's weekly exercise solutions are added to the scores of the essay.

The essay can also be written in Finnish. However, the use of terms and concepts of data mining should then mostly obey those corresponding, normally used in Finnish. (There might exist such that do not have established translations.)

TOPICS:

(1) On nearest neighbor searching and classification

(2) Variable selection in data mining

(3) Detection of outliers in data mining

The essay has to be sent by email no later than at 24 o'clock on the **30th April, 2020**, to Martti.Juhola@tuni.fi. Any text sent later is equal to 'failed'. In the case of some *force majeure* one has to take contact early enough before the above date.