Posudek diplomové práce

Matematicko-fyzikální fakulta Univerzity Karlovy

Autor práce	Jakub Beran		
Název práce	Machine learning on small datasets with large number of features		
Rok odevzdání	2020		
Studijní program	Computer Science	**Studijní obor**	Umělá inteligence

Autor posudku	Filip Matzner	**Role**	oponent
Pracoviště	KSVI		

Text posudku:

The topic of the presented thesis is the selection of relevant features from large feature sets with only a few observations. Such datasets are often found in field of studies known as "omics" (such as genomics) where the number of features increases with the available measuring technology progress throughout the years. The author has thoroughly evaluated multiple classification and feature selection methods on both artificial and real-world datasets.

The work is written in a very good English without typographical errors. The text is smoothly guiding the reader through the thesis and the language is easily readable even in the technical parts, such as description of the used methods. The author often mentions useful and less known insights (such as limitations and properties of the lasso method) where appropriate.

One thing I would suggest to improve: often a chapter starts by discussing the outcomes of the experiments and the results and the plots are presented later in the chapter. I believe the text would be smoother by presenting the results first, so that the reader can build a deeper understanding on which could the discussion stand and build later. That is the classical text order in which the information provided so far implies the information to follow, not vice versa.

One more thing that did not feel quite right was that the thesis was presented as a theoretical work, but the text sometimes felt as a software documentation.

For instance, "Outputs are saved only if the program parameter SAVE is set to True." in chapter Experimental Design. I believe the text would be more approachable if it would keep the implementation details and the design of the experiments separated.

The source code is written in Python and consists of approx. 1900 lines of code in total. The code is well documented and easy to read. However, there are a few parts that could be improved, for instance, nested parameter parsing at the end of `mainCVRealWorld.py` would be better suited for e.g., Python's `argparse`. On the other hand, I consider the source code to be more of an experimental tool to produce the results and plots for the text itself.

Overall, the problematics is interesting, the cited literature is recent and the author was able to fulfill the goals of the thesis.

Práci doporučuji k obhajobě.

Práci nenavrhuji na zvláštní ocenění.

V Praze dne 28. 06. 2020

Podpis: