«Machine Learning and Data Mining»

Faculty of Computer Science, 1^{st} *year master students*

Group or individual project

Author: Dmitry Ignatov.

TA: Dmitry Egurnov.

The project is to be chosen by 25.05.20 (team, problem description, data source) and the final report is due on the week prior to defence (18.06.2020).

Problem statement

This Group project aims at solving a machine learning and data mining problem or challenge in collaboration with other students or individually.

Before you start the project, it is necessary to:

1. Form teams of at least one (individual project) and at most three students (group project);
2. Find a dataset to be analysed in the project;
3. Provide concise and clear statement of research goals, a brief summary of the dataset, and outline a roadmap for a prospective solution.

It is important to submit the project proposal for approval by the lecturer or a teaching assistant before the first deadline (10.05.19). Only after the project proposal has been approved it advisable to start working on it. Below is a short list of web repositories with freely available datasets (the list is not exhaustive):

UC Irvine Machine Learning Repository
`http://www.kaggle.com/competitions`
`http://www.openml.org/`
`http://www-stat.stanford.edu/~tibs/ElemStatLearn/`
`http://lib.stat.cmu.edu/datasets`
`http://www.statsci.org/datasets.html`
`http://www.amstat.org/publications/jse/jse_data_archive.htm`
`http://www.physionet.org/physiobank/database`
`http://biostat.mc.vanderbilt.edu/twiki/bin/view/Main/DataSets`.

Projects concerning Text Mining and Natural Language Processing are also allowed. To this end one may have a look at these web sites (but not limited to): `http://universaldependencies.org/`, TREC, , and `http://pan.webis.de/data.html`, etc.

The following structure of the report is suggested:

1. Problem statement;
2. Dataset summary with basic statistics and respective plots;
3. Methodology (justify the selected ML/DM approach);
4. Experiment setup and results; error analysis.
5. Discussion (comparison, interpretations, etc.);
6. Conclusion.

Q.: *Are there any restrictions to the size of the dataset, the report size, and/or the set of applied machine learning algorithms?*

A.: *Yes, there are. The dataset must not be too small, at least 50 objects with at least 7-10 features (generated features can be taken into account). The report must be relevant to the proposal, coherent and sufficiently detailed, so that either the examiner or a student of the same class, could be able to understand the work you did, the research steps undertaken and the goals achieved. This entails reasonable and relevant usage of tables, plots, or other visualization tools. In general, the more data analysis approaches are used the better; note that usage of no more than two different methods is penalised. The comparison and interpretation of the obtained results are crucial. For instance, when using clustering, it is important to compare the results of different methods, and analyse the effects of parameter tuning within the same algorithm (at least empirically). Implementation of new ideas, methods or algorithms are strongly encouraged. It is necessary to demonstrate the complete data analysis pipeline: data collection, data preprocessing (scaling, outlier elimination, feature selection, etc.), method application, comparison, analysis and interpretation.*

Q.: *What are the typical machine learning algorithms to apply?*

A.: *They are not limited by the methods studied in the class. You may choose algorithms depending on your task: classification, regression, clustering, ranking, recommendation, pattern mining, etc.*

The defense of the project will take place during the exam week. The highest grade is 10.

The following packages are suggested for the use in the project:

- Scikit-leran `http://scikit-learn.org/stable/`
- Orange `http://orange.biolab.si/`;
- Weka `www.cs.waikato.ac.nz/ml/weka`;
- Matlab or R.

Other advanced libraries like Vowpal Wabbit (`http://hunch.net/~vw/`) or MLlib (`http://spark.apache.org/mllib/`) are welcome as well. The use of frequent itemset mining, for example SPMF package (`http://www.philippe-fournier-viger.com/spmf/`) is encouraged, but not mandatory.

Emails with the project proposals and the final version of the reports are to be sent to these emails: "to" *dmitrii.ignatov@gmail.com*, with a mandatory "cc" to *egurnovdima@gmail.com*.

The topic of the letter must have the following format
[MLDM2020m-Project]-LAST NAME(s)-FIRST NAME(s).