New projects may be added during the next couple of days. Feel free to propose your own project ideas in the Blackboard forum at least 48h before the deadline for submitting your project preferences.

ML Research Assignments 2018/19   
Project Descriptions

Version 1.2, 5. Oct. 2018  
Joeran Beel and Douglas Leith

Simple Projects

# How consistent are the implementations of machine learning algorithms in different ML libraries?

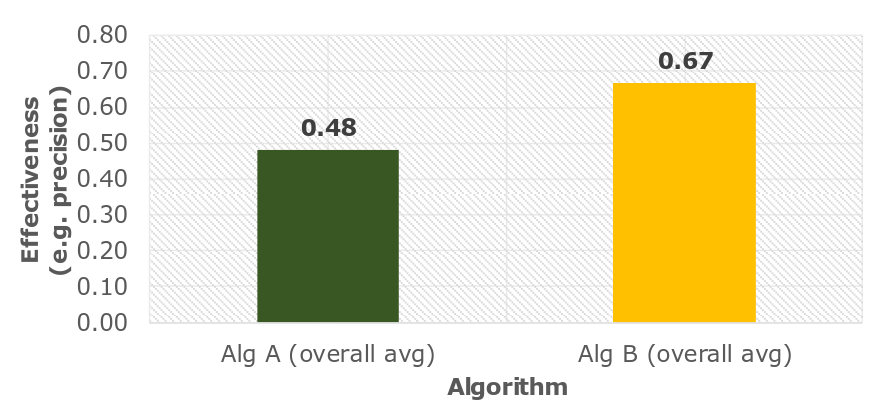
Many machine learning algorithms such as linear regression or neural networks are implemented by various machine learning frameworks such as scikit-learn, TensorFlow and Weka. However, minor changes in implementations (e.g. different default hyper parameters) may sometimes lead to significant changes in the performance of machine learning algorithms. Hence, your task is to answer the question whether implementations of the same algorithms in different frameworks producing the same results?

# Dataset Pruning: What is the effect on Machine Learning Performance?

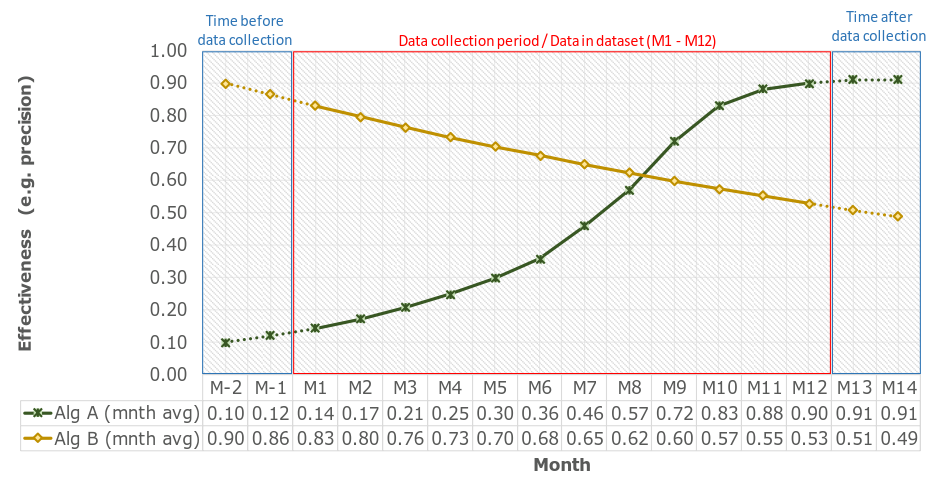
If you prune a dataset, i.e. if you remove certain data that you consider as suboptimal (e.g. all users with less than x ratings), how does that affect the evaluation of machine learning algorithms? For instance, how much “better” is an algorithm becoming when the data is pruned? Or, would an unpruned dataset show that algorithm B is better than algorithm A, but using a pruned dataset would show that algorithm A is better than algorithm B? If you read a research article and the authors report that x% of the original data was removed, how meaningful is the evaluation?

# Does ML algorithm performance change over time?

Typically, when evaluating machine learning algorithms, researchers report the overall e.g. accuracy or RMSE. However, would results appear different when looking at time intervals? For instance, could it be that the overall results look like this:



But when looking at time, results would look like this?



# How much data do you need to train machine learning algorithms effectively?

Find out how the performance of machine learning algorithms changes when the amount of data for training changes.

# Noise in Datasets: How strong is the impact on ML performance?

Obviously, noisy data is not good for machine learning performance. However, how strong exactly is the impact of noise on machine learning performance?

# Hyper-Parameter Tuning: How strong is the impact on ML performance?

Machine learning algorithms have hyper parameters than cannot be learned and that need to be optimized (e.g. “k” in k-nearest neighbour). How strong is the impact of optimizing hyper parameter? In other words, how much better do optimized algorithms perform compared to algorithms for which hyperparameters were not optimized? Or, how effective are different optimization strategies?

# How well can the gender of Twitter users be predicted?

There are different Twitter datasets. They contain information about twitter users (e.g. date of registration, tweet count) and often the users´ gender. Based on the user´s features, how well can their gender be predicted, and which features are particularly predictive?

Advanced Projects

Only pick these projects if you are very interested in them, have already some advanced machine learning knowledge, and if you are willing to spend some more time than usually expected for a 5 ECTS module.

# How suitable are the RARD II and Docear dataset for Machine Learning Tasks?

Our research group has released two datasets related to our own recommender systems:

* RARD II <https://www.scss.tcd.ie/joeran.beel/blog/2018/07/18/rard-ii-the-2nd-related-article-recommendation-dataset-preprint/>
* Docear: <https://www.docear.org/labs/> and <http://docear.org/papers/The%20Architecture%20and%20Datasets%20of%20Docear's%20Research%20Paper%20Recommender%20System.pdf>

We have not yet used these datasets for any machine learning tasks. Your goal would be to explore how suitable the datasets are for machine learning. For instance, how well can you predict whether a recommendation will be clicked? Or can you predict the age or gender of a user based on the documents the user has clicked? ...? Be creative :-).

# Language Detection with Machine Learning: How effective is it?

In machine learning, you often work with text (e.g. emails, web pages, ...). For some applications, it is beneficial if you know the language of a text. To detect languages there are many frameworks, some based on machine learning, others based on other methods. The question is, how good are these machine learning methods/frameworks for language detection?

# How does the effectiveness of ensembles change, based on the number of algorithms in the ensemble? [Student Suggestion]

We want to identify the number of models which brings this saturation of accuracy. This may vary on basis of algorithms, parameters and datasets as well. In short, we want to identify the maximum number of models one can add in an ensemble model to achieve highest accuracy. Or if there is any such maximum number that can be calculated.

Challenging Projects

These projects are only suitable for students who already have solid knowledge in machine learning and who are willing to spend more time for the assignment than would be expected for a 5 ECTS module.

# How stable are Neural Turing Machine implementations?

One of my Bachelor students recently implemented the apparently first stable version of a Neural Turing Machine for TensorFlow. The results and source code were published https://www.scss.tcd.ie/joeran.beel/blog/2018/08/01/a-stable-neural-turing-machine-ntm-implementation-source-code-and-pre-print/ . Can you reproduce the results of this implementation and/or of other Neural Turing Machine implementations or similar algorithms? In other words, how (un) stable are the different implementations?

# Curriculum Learning: Does it work?

“Curriculum Learning” describes the idea that machine learning is more effective and faster if the training data is sorted by the level of difficulty. This means, the model would be trained first with simple data and then with more complex data. However, there is contradicting information on whether this approach works effectively. It would be your task to experiment with a few datasets and algorithms to find out if and when curriculum learning works.

# Is “Stability” suitable as novel evaluation metric for machine learning algorithms?

Typically, when evaluating machine learning algorithms, researchers report the overall e.g. accuracy, precision or RMSE. However, the overall result only provides a superficial picture.

Imagine the following experiment. Three algorithms X, Y and Z are trained and evaluated on the same or a similar dataset, and there are four variants of each algorithm (A´, A´´, A´´´, A´´´´, …, C´´´´). These variants may be variations of the algorithms (e.g. different hyper parameters), different folds during the evaluation or slightly different datasets. All three algorithms have the same overall precision *p,* i.e. *10/20=0.5* (see illustration). However, when looking at the results in more detail, there are notable differences. Algorithm X always has the data points A, B (and C) in its four variations as positive results, and (N), G and H as negative results. This means, all four variations are quite stable and produce, more or less, the same output each time. In contrast, algorithm Y is unstable. Although precision is the same for each variation, the actual output differs in each variation of Y. For instance, the output of Y´ is A, B, C, G, H while for Y´´ the output is D, E, J, L, M. Also, algorithm Z is unstable, though in a different way. Looking at Z, each variation has a very different precision ranging from 0 for Z´´ to 1 for Z´ and only the overall precision is 0.5.



If you had to decide, which of the three algorithms to use in your system, you would consider all three algorithms as equally good based on precision. However, if you consider that the data in your system would be slightly different from the test scenario, you probably would want an algorithm that is more stable. Eventually, you could not be sure how algorithm Z will perform in a slightly different scenario because it behaves unstable.

However, the given example is made up. The question is, do such differences in stability occur in real machine learning experiments, and would it make sense to report the “stability” as an evaluation metric?