Internship: Automatic Machine Learning Methods For Clustering

Machine learning techniques [e.g. Witten, 2016] have become very popular in the last decades of scientific research and their usage increases exponentially in the industry [e.g. Chen, 2015]. These methods aim at automatizing the production of algorithms for tasks where designing them manually would be very difficult (too many inputs, unknown and non-deterministic processes, ...). Most of these algorithms are based on the optimization of cost functions and heavily rely on gradient based optimization techniques.

However, most machine learning algorithms also rely on *hyper parameters* conditioning the optimization to perform in such a way that these parameters can't easily be part of the functions to optimize and no gradient is defined for them. Finding the best values of these hyper parameters is a problem by itself which can not use the same optimization techniques. Very popular ways to deal with hyper parameters often involve: 1) defining an arbitrary set of values for them and compare the different values with each other in term of machine learning performances (a.k.a. *grid search*); 2) sampling hyper parameter values and again compare the results (a.k.a. *random search*). These methods are however very limited.

More recent approaches [Snoek, 2012; Li, 2017] try to browse the hyper parameters space in a smarter way so that they are not to be defined by someone anymore but still converge towards good settings. Their promises towards machine learning without (as much) expertise have led to the term *Automatic Machine Learning* to reference them. Of course, since this new field is very promising and ambitious, many problems are open.

The objective of the internship will be to:

1) dive into the Automatic Machine Learning field to study the recent advances in it;

2) implement and manipulate state of the art algorithms in **Scala** language;

3) compare such algorithms on various *clustering* benchmarks;

4) contribute to the existing open source tools developed in the lab (https://github.com/Clustering4Ever/Clustering4Ever).

Practical Details

Profile: Motivated student wishing to dig into data science / machine learning / AI state of the art problem and algorithms to obtain practical results and participate in the implementation of open source software, with interaction perspectives in industrial fields.

Level: Master 2 or engineer level (with Computer Science, Statistics, or Applied Mathematics backgrounds).

Supervisors: anthony.coutant@lipn.univ-paris13.fr, mustapha.lebbah@lipn.univ-paris13.fr

Location of the internship: LIPN - UMR 7030 - CNRS, Université Paris 13.

Start period: March/April 2019, 6 months long

Grant Amount: 577.50 € / month

To apply, please send a resume and transcripts of your last two years to anthony.coutant@lipn.univ-paris13.fr

References

Chen, L. L., Zhao, Y., Zhang, J., & Zou, J. Z. (2015). Automatic detection of alertness/drowsiness from physiological signals using wavelet-based nonlinear features and machine learning. *Expert Systems with Applications*, *42*(21), 7344-7355.

Li, L., Jamieson, K., DeSalvo, G., Rostamizadeh, A., & Talwalkar, A. (2017). Hyperband: A novel bandit-based approach to hyperparameter optimization. *The Journal of Machine Learning Research*, *18*(1), 6765-6816.

Snoek, J., Larochelle, H., & Adams, R. P. (2012). Practical bayesian optimization of machine learning algorithms. In *Advances in neural information processing systems* (pp. 2951-2959).

Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2016). *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann.