Master Internship Offer - Spring 2022
Random Fourier Features for PAC-Bayesian Domain Adaptation

Level: Master 2

Information

Advisor(s): Guillaume Metzler, Emilie Morvant, and Paul Viallard
Mail: guillaume.metzler@univ-lyon2.fr emilie.morvant@univ-st-etienne.fr, paul.viallard@univ-st-etienne.fr

Location *(at the candidate's choice)*: Laboratoire ERIC (Lyon) or Laboratoire Hubert Curien (Saint-Etienne)

Duration: 5-6 months, between February and August 2022
Compensation: 575 euros/month

Keywords: Machine Learning, Transfer Learning, Representation Learning

General Overview: The main idea of this internship is to take advantage of the PAC-Bayesian theory and recent works that have been published in this domain in order to develop or build generalization bounds for Domain Adaptation.

Expected profile: Master or engineering degree in Computer Science or Applied Mathematics related to machine/statistical learning. The candidate must show some interest in the theoretical aspects of Machine Learning as well as programming skills. Furthermore, she/he must be fluent in reading and writing English.

How to apply? Send to **guillaume.metzler@univ-lyon2.fr**, **emilie.morvant@univ-st-etienne.fr**, and **paul.viallard@univ-st-etienne.fr**

- a CV
- a motivation letter
- your Master grades

Summary

In Machine Learning, the issue of data representation is crucial: without a data representation that captures the relevant information for the learning task at hand, it will not be possible to solve the task in an optimal way. Nowadays, one of the most popular methods is based on deep neural networks. However, there are still few theoretical results explaining their success. This internship stands in another framework, that is better understood theoretically, where the objective is to learn function that "compares" two data points. We generally talk about metric, similarity or kernel function learning, (see, for example, [1, 2, 3, 7]). We propose to study the learning of such functions from **Random Fourier Features** (RFF [6]), a method that approximates a kernel function based on a combination of random attributes (combination defined by a probability distribution on the attributes).

In this context, we have obtained theoretical guarantees based on the PAC-Bayesian theory [5] known to bring guarantees on models expressed as a weighted combination.This work brought a new point of view on the RFF which allowed us to develop a gradient boosting algorithm to learn a representation based on so-called landmark points [2]. These results are within the classical framework of supervised learning where the training and test data are derived from the same underlying data distribution.

During this internship, we propose then to take advantage of the specificities of the PAC-Bayesian theory for majority votes. First, when constructing a combination, a key point is the notion of diversity of functions involved in the final combination [4]. In PAC-Bayesian theory, it has been shown that the risk associated to a combination of voters strongly depends on the trade-off between the disagreement and the individual risk. Therefore, the objective is to consider this kind of measure to improve a RFF-based approach in both supervised classification and domain adaptation from a practical point of view. Second, another important aspect of the PAC-Bayesian theory is its stochastic nature. Recent works have been published on this topic with several theoretical results based on weighted combinations. In particular, we have recently obtained more accurate theoretical results in the framework of weighted combinations. This so-called stochastic PAC-Bayesian theory, allowing to minimize directly the 0-1 loss and to obtain better results than with the classical PAC-Bayesian approaches based on the notion of diversity [8]. The question of specializing this result to domain adaptation with RFF is an interesting complementary perspective that the candidate will have to explore.

Expected results

- Literature review: Literature review on Domain Adaptation and PAC Bayesian Learning and the preliminary work based on Random Fourier Features
- Theoretical: A possible new PAC-Bayesian analysis of the task
- Practical: Implementation and evaluation of the proposed algorithm(s)

If the candidate want to know more about the context of this research intership, he may consider reading the following articles:

- PAC Bayesian Framework: Pascal Germain et al. *Risk Bounds for the Majority Vote: From a PAC-Bayesian Analysis to a Learning Algorithm*, JMLR, 16(26):787-860, 2015[1]
- Domain Adaptation: Pascal Germain et al. *PAC-Bayes and Domain Adaptation*, Neurocomputing, 379:379-397, 2020[2]

[1]This article can be found following this link : https://jmlr.org/papers/v16/germain15a.html.
[2]This article can be found following this link : http://arxiv.org/abs/1707.05712.

References

[1] Aurélien Bellet, Amaury Habrard, and Marc Sebban. Metric learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, 9(1):1–151, 2015.

[2] Léo Gautheron, Pascal Germain, Amaury Habrard, Guillaume Metzler, Emilie Morvant, Marc Sebban, and Valentina Zantedeschi. Landmark-based ensemble learning with random fourier features and gradient boosting. In *ECML-PKDD*, 2020.

[3] Leo Gautheron, Amaury Habrard, Emilie Morvant, and Marc Sebban. Metric learning from imbalanced data with generalization guarantees. *Pattern Recognition Letters*, 133:298–304, 2020.

[4] Ludmila Kuncheva. *Combining pattern classifiers: methods and algorithms*. John Wiley & Sons, 2014.

[5] Gaël Letarte, Emilie Morvant, and Pascal Germain. Pseudo-bayesian learning with kernel fourier transform as prior. In *AISTATS*, pages 768–776. PMLR, 2019.

[6] Ali Rahimi, Benjamin Recht, et al. Random features for large-scale kernel machines. In *NeurIPS*, 2007.

[7] Rémi Viola, Rémi Emonet, Amaury Habrard, Guillaume Metzler, and Marc Sebban. Learning from few positives: a provably accurate metric learning algorithm to deal with imbalanced data. In *IJCAI-PRICAI2020, the 29th International Joint Conference on Artificial Intelligence and the 17th Pacific Rim International Conference on Artificial Intelligence*, 2020.

[8] Valentina Zantedeschi, Paul Viallard, Emilie Morvant, Rémi Emonet, Amaury Habrard, Pascal Germain, and Benjamin Guedj. Learning stochastic majority votes by minimizing a pac-bayes generalization bound. In *NeurIPS*, 2021.