Master Internship Offer - Spring 2024
PAC-Bayesian Fair Learning

Level: Master 1 or Master 2

Information

Advisor(s): Guillaume Metzler and Emilie Morvant
Mail: guillaume.metzler@univ-lyon2.fr emilie.morvant@univ-st-etienne.fr
Location ***(at the candidate's choice)*****:** Laboratoire ERIC (Lyon) or Laboratoire Hubert Curien (Saint-Etienne)

Duration: 4 to 6 months, between February and August 2024
Compensation: 575 euros/month

Perspectives (for M2 student): this offer, if the recruited candidate is motivated and shows a great interest for subject by carrying out a quality internship, could lead to a PhD thesis on a close subject and financed by the ANR.

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 .

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 and emilie.morvant@univ-st-etienne.fr

- a CV
- your Master grades

Summary

Machine learning methods are sometimes heavily biased, leading to unfair decisions [1, 6]. This behavior is due to biases in the data used for learning. These biases can be linked to so-called sensitive variables, such as gender, or to under-represented data [3, 5, 10]. In the literature, there are various methods for building the fairest possible models. Many methods aim to learn a model by minimizing a loss function that captures a trade-off between the errors made by the model and its ability to respect a certain notion of fairness [14, 15]. The idea is to be able to maximize the model's performance while constraining it to be as fair as possible.

To the best of our knowledge, very little work in theoretical machine learning studies generalization capabilities in this framework (for example, [9, 8, 2]). The aim of this internship is to initiate work around machine learning theory to better understand how to learn a fair model in the presence of bias. More specifically, we propose to exploit the framework of PAC-Bayesian theory. We hope that this work will improve our understanding of models, leading to the development of new and fair learning methods.

Two lines of research are possible, the candidate will be free to choose the line in which they would like to carry out their internship.

The first possible line of research stands in the PAC-Bayesian theory [11, 7]. In this framework given a training set, a family of models and a so-called "*prior*" distribution on this family, the objective is to learn a so-called "*posterior*" distribution optimizing a certain loss function capturing the quality of the *posterior* distribution. In this framework, the quality of the *posterior* is estimated as a generalization bound expressed on average according to the *posterior* distribution over the family of models, and are therefore by nature suitable for stochastic models (in constrast to the more classical "worst-case" bounds). Oneto *et al.* [9] have demonstrated the interest of this stochastic aspect to ensure a certain fairness during learning. They also demonstrated the potential of PAC-Bayesian models and the ability of the *posterior* distribution to ensure fairness in learning. While this promising work focuses on stochastic models (or algorithms), our objective is to derive results for deterministic models (or algorithms). As a first step, based on our recent work [13] on the "dereandomization" of PAC-Bayesian results, we would like to generalize the results of Oneto *et al.* [9] to deterministic models, while having theoretical guarantees on the generalization capacity of the models learned. These results are set within a framework of traditional loss functions and fairness measures (such as *Demograhic Parity*, *Equality of Opportunity* or a more recent fairness measure, *CVaR (Conditional Value of Risk) [8, 14]*).

The other possible line of research is to focus on the development of new fairness measures, taking advantage not only of the knowledge acquired through our initial theoretical results, but also of the properties specific to PAC-Bayesian theory (such as the diversity of the family of models, and their ability to capture information that is different from, but complementary to, the attributes of the data). A first approach would be to reformulate classical equity measures to define a performance/equity trade-off specific to PAC-Bayesian theory. This has the advantage of enabling us to derive bounds in more precise generalizations that will probably have the property of being directly optimizable to obtain self-certified algorithms [4, 12]. A second avenue is to exploit one of our recent results [12, Chapter 7] which allows such measures to be directly integrated into the bound as an arbitrary "complexity measure" that can further improve theoretical and empirical results.

Main objectives of the proposed Internship The candidate will be required to develop both theoretical and practical skills in the field of learning under fairness constraints. His work will therefore comprise the following two parts, which may be carried out jointly:
(i) derived generalization bounds on the obtained models under fairness guarantees
(ii) define new measures of fairness based on the PAC-Bayesian framework.

Expected results

- Literature review: Literature review PAC Bayesian Learning and Fairness in Machine Learning.
- Theoretical: study how we can derive or improve fairness guarantees using the PAC-Bayesian framework.
- Practical: Implementation and evaluation of the proposed algorithm in terms of both fairness and performances.

If the candidate want to know more about the context of this research internship, 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]
- Possible definitions of fairness presented by Ben Hutchinson *et al.*, *50 Years of Test (Un)Fairness: Lessons for Machine Learning*, FAT*'19, page 49-58, 2019[2]
- Solving a task under fairness constraint by Michele Donini *et al.*, *Empirical Risk Minimization under Fairness constraints*, Advances in Neural Information Processing Systems, 2019. [3]
- A derandomized PAC-Bayesian bound: Paul Viallard *et al. A General Framework for the Practical Disintegration of PAC-Bayesian Bounds*, Machine Learning, 2023[4]

[1] https://jmlr.org/papers/v16/germain15a.html.

[2] https://arxiv.org/pdf/1811.10104.pdf.

[3] https://papers.nips.cc/paper_files/paper/2018/hash/83cdcec08fbf90370fcf53bdd56604ff-Abstract.html.

[4] https://arxiv.org/abs/2102.08649.

References

[1] Solon Barocas, Moritz Hardt, and Arvind Narayanan. *Fairness and Machine Learning: Limitations and Opportunities*. fairmlbook.org, 2019.

[2] Cyrus Cousins. Revisiting fair-pac learning and the axioms of cardinal welfare. In *International Conference on Artificial Intelligence and Statistics*, volume 206, 2023.

[3] Michele Donini, Luca Oneto, Shai Ben-David, John S Shawe-Taylor, and Massimiliano Pontil. Empirical risk minimization under fairness constraints. *Advances in neural information processing systems*, 31, 2018.

[4] Yoav Freund. Self bounding learning algorithms. In *COLT*, 1998.

[5] Sara Hooker, Aaron Courville, Gregory Clark, Yann Dauphin, and Andrea Frome. What do compressed deep neural networks forget? *arXiv preprint arXiv:1911.05248*, 2019.

[6] Ben Hutchinson and Margaret Mitchell. 50 Years of Test (Un)Fairness: Lessons for Machine Learning. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, page 49–58, 2019.

[7] David McAllester. Some PAC-Bayesian Theorems. In *COLT*, 1998.

[8] Zakaria Mhammedi, Benjamin Guedj, and Robert C Williamson. Pac-bayesian bound for the conditional value at risk. In *Advances in Neural Information Processing Systems*, 2020.

[9] Luca Oneto, Michele Donini, Massimiliano Pontil, and John Shawe-Taylor. Randomized learning and generalization of fair and private classifiers: From pac-bayes to stability and differential privacy. *Neurocomputing*, 416:231–243, 2020.

[10] Irina Proskurina, Guillaume Metzler, and Julien Velcin. The other side of compression: Measuring bias in pruned transformers. In *Advances in Intelligent Data Analysis XXI: 21st International Symposium on Intelligent Data Analysis, IDA 2023, Louvain-la-Neuve, Belgium, April 12–14, 2023, Proceedings*, pages 366–378. Springer, 2023.

[11] John Shawe-Taylor and Robert Williamson. A PAC Analysis of a Bayesian Estimator. In *COLT*, 1997.

[12] Paul Viallard. *PAC-Bayesian Bounds and Beyond: Self-Bounding Algorithms and New Perspectives on Generalization in Machine Learning*. PhD thesis, Université Jean Monnet de Saint-Etienne, 2023.

[13] Paul Viallard, Pascal Germain, Amaury Habrard, and Emilie Morvant. A general framework for the practical disintegration of pac-bayesian bounds. *Machine Learning*, 2023.

[14] Robert Williamson and Aditya Menon. Fairness risk measures. In *International Conference on Machine Learning*, pages 6786–6797. PMLR, 2019.

[15] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez-Rodriguez, and Krishna P. Gummadi. Fairness constraints: A flexible approach for fair classification. *Journal of Machine Learning Research*, 20(75):1–42, 2019.