Reliable and fair machine learning for risk assessment

Pereira Barata, A.P.

Curriculum Vitae

António Pedro Pereira Barata was born in Lisbon, Portugal, on the 21st of December 1989. He completed his B.Sc. in Biology in 2013, specialising in Evolutionary and Developmental Biology, in the Faculty of Science of the University of Lisbon. Within the same institute, he acquired his M.Sc. in Bioinformatics and Computational Biology in 2017. Thereafter, he was admitted to Leiden University, the Netherlands, to conduct the research presented in this doctoral thesis, under the attentive care and supervision of H. Jaap van den Herik, Cor J. Veenman, and Frank W. Takes. At the time of writing, he is employed by the Ministry of Infrastructure and Water Management of the Netherlands, within the Innovation and Data Lab of the ILT.

Publications

While working towards this thesis, the following contributions were made.

- Pereira Barata, A., Takes, F. W., van den Herik, H. J., and Veenman, C. J. (2022). Fair tree classifier using strong demographic parity. *Machine Learning (under review)*
- Pereira Barata, A., Takes, F. W., van den Herik, H. J., and Veenman, C. J. (2022). Noise-resilient classifier learning. *Pattern Recognition (under review)*
- Pereira Barata, A., Takes, F. W., van den Herik, H. J., and Veenman, C. J. (2021). The eXPose approach to crosslier detection. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 2312–2319. IEEE
- Pereira Barata, A., Takes, F. W., van den Herik, H. J., and Veenman, C. J. (2019). Imputation methods outperform missing-indicator for data missing completely at random. In *2019 International Conference on Data Mining Workshops (ICDMW)*, pages 407–414. IEEE
- Pereira Barata, A., de Bruin, G. J., Takes, F. W., Veenman, C. J., and van den Herik, H. J. (2018b). Finding anomalies in waste transportation data with supervised category models. In *2018 27th Belgian Dutch Conference on Machine Learning (BeNeLearn)*
- Pereira Barata, A., de Bruin, G. J., Takes, F. W., Veenman, C. J., and van den Herik, H. J. (2018a). Data-driven risk assessment in infrastructure networks. In *ICT.open*
- de Bruin, G. J., Pereira Barata, A., van den Herik, H. J., Takes, F. W., and Veenman, C. J. (2022). Fair automated assessment of noncompliance in cargo ship networks. *EPJ Data Science*, 11(1):13
- Angenent, M. N., Pereira Barata, A., and Takes, F. W. (2020). Large-scale machine learning for business sector prediction. In *Proceedings of the 35th Annual ACM Symposium on Applied Computing*, pages 1143–1146

Acknowledgements

A thesis is not something one can muster alone. Throughout this long and arduous journey, I had the privilege of meeting, working with, and learning from some of the most fascinating people I had ever come across.

I would like to thank Cor Veenman, for all his wisdom and patience in teaching me how to do research in machine learning. Many a *fiery* discussion were had between the two of us. I sincerely hope you enjoyed them as much as I did. In extension, I thank the Netherlands Organisation for Applied Scientific Research (TNO).

I thank Frank Takes for his forever good disposition, and for teaching me that I too could be a *glass half-full* kind of person; not only towards myself, but especially towards others. I honestly believe you have made me a better person. In extension, I thank the Leiden Institute of Advanced Computer Science (LIACS).

To Jaap van den Herik, I will forever be grateful: thank you for promoting the best in me, for teaching me to pursue the *exceptional* rather than the *passable*. This thesis would not be as it is without your aid. In extension, I thank the Leiden Centre of Data Science (LCDS).

I would like to thank also Gerrit-Jan de Bruin, for being not only my colleague but also my friend. You were the only person who *really* understood what this project entailed, its ups and downs. We have been on this same PhD boat for years now, and I eagerly look forward to our next trip. In your own boat. You own a boat. Please invite me to your boat.

I would like to express my thanks to the Innovation and Data Lab team. Particularly, I thank Jasper van Vliet for doing his utmost to integrate me in the team and the challenges we tackled, Arjan van der Put for our many adventures in the land of trial-and-error, Corline Koolhaas for her continuous concern for my well-being especially during the last stages of the thesis, and Tony Liebregts for (amongst other things) our delightful —yet short— afternoon in Lisbon. I extend my thanks also to the ILT of the Ministry of Infrastructure and Water Management, especially thanking Jan van den Bos without whom none of this work would have been possible.

With regards to Alexa and Patins, I have no idea how you two were able to put up with me during these PhD years. Without you both, I would have crumbled, and this thesis would not exist. I could not have asked for better people with whom to share my daily struggles. You were (almost) always supportive, (almost) always understanding, (almost) always impeccable; but most importantly, you were *always* there for me.

To all my family, to all my friends, to all the unnamed heroes who shaped me into what I am today: thank you.