LEARNING TO TRANSFER WITH TRIPLY ADVERSARIAL NETS

By Gilles Louppe

New York University

In classification, transfer learning (or its variants known as covariate shift or domain adaptation) arises whenever test instances are governed by a distribution that may be arbitrarily different from the distribution of the training instances. This problem has traditionally been solved by reweighting training instances or by learning robust representations over domains. In this work, we propose a new paradigm based on the assumption that the covariate shift is caused by a change in the representation of the same underlying objects. Accordingly, we propose to learn how to transform training instances into test instances, possibly across input spaces of distinct dimensions, structures or supports. For this purpose, we extend the generative adversarial networks framework of Goodfellow et al. (2014) to a triply adversarial process: a transformer network T for generating test instances from training instances, a discriminative network D for separating transformed training instances from test instances, and a classifier network $C \circ T$ for classifying training instances in the projected space. This 3-player game results in a network T capable of transforming training instances into test instances, while preserving separation between classes as enabled by C in the adversarial setup. Experimental results demonstrate the potential of this novative approach for transfer learning. More fundamentally, this paradigm also raises new theoretical issues, since transforms T may not be unique nor all grounded, yet we seek only the one making most sense.

References.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative adversarial nets. In *Advances in Neural Information Processing Systems*, pages 2672–2680.