LEARNING TO TRANSFER WITH TRIPLY ADVERSARIAL NEURAL NETWORKS

By Gilles Louppe

New York University

Learning under covariate shift, also known as transfer learning or domain adaptation, arises in supervised learning 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 either by reweighting training instances or by learning feature representations that are robust to domain adaptation. In this work, we propose a new paradigm which consists instead in learning to transform training instances into test instances. We extend the generative adversarial networks framework (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 estimating whether an instance comes either from the training or the test distributions, and a classifier network C for classifying training instances in the projected space. Accordingly, this 3-player game results in a network T capable of transforming training instances into test instances, while preserving the separation between classes. Experimental results demonstrate the potential of this novative approach and illustrate the importance of both regularization and network architecture for controlling the learned transformation.

1. **Introduction.**

2. **Method.**

3. **Experiments.**

4. **Related work.**

5. **Conclusions.**

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.