Learning to Transfer with Triply Adversarial Nets

Gilles Louppe
New York University
g.louppe@nyu.edu

Abstract

In classification, transfer learning (or its variants known as covariate shift or domain adaptation) arises whenever target instances are governed by a distribution that may be arbitrarily different from the distribution of the source instances. This problem has traditionally been solved by reweighting approaches 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 the use of a different representation of the same underlying objects. Accordingly, we propose to learn how to transform source instances into target instances, possibly across input spaces of distinct dimensions, structures or supports. For this purpose, we extend the generative adversarial networks framework of [1] to a triply adversarial process: a transformer network T for generating target instances from source instances, a discriminative network D for separating transformed source instances from actual target instances, and a classifier network $C \circ T$ for classifying source instances in the projected space. This 3-player game results in a network T capable of transforming source into target instances, while preserving separation between classes as enabled by C in the adversarial setup. Experiments demonstrate the potential of this novative approach, with promising results when the construction of C can be bootstrapped in a semi-supervised way from a few labeled instances from the target space.

1 Introduction

2 Method

3 Experiments

4 Related work

5 Conclusions

Acknowledgments

[GL: todo]

References

[1] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, pages 2672–2680, 2014.