

Learning and Domain Adaptation

Yishay Mansour

Blavatnik School of Computer Science,
Tel Aviv University
Tel Aviv, Israel
`mansour@tau.ac.il`

Abstract. Domain adaptation is a fundamental learning problem where one wishes to use labeled data from one or several source domains to learn a hypothesis performing well on a different, yet related, domain for which no labeled data is available. This generalization across domains is a very significant challenge for many machine learning applications and arises in a variety of natural settings, including NLP tasks (document classification, sentiment analysis, etc.), speech recognition (speakers and noise or environment adaptation) and face recognition (different lighting conditions, different population composition).

The learning theory community has only recently started to analyze domain adaptation problems. In the talk, I will overview some recent theoretical models and results regarding domain adaptation.

This talk is based on joint works with Mehryar Mohri and Afshin Rostamizadeh.

1 Introduction

It is almost standard in machine learning to assume that the training and test instances are drawn from the same distribution. This assumption is explicit in the standard PAC model [19] and other theoretical models of learning, and it is a natural assumption since when the training and test distributions substantially differ there can be no hope for generalization. However, in practice, there are several crucial scenarios where the two distributions are similar but not identical, and therefore effective learning is potentially possible. This is the motivation for *domain adaptation*.

The problem of domain adaptation arises in a variety of applications in natural language processing [6, 3, 9, 4, 5], speech processing [11, 7, 16, 18, 8, 17], computer vision [15], and many other areas. Quite often, little or no labeled data is available from the *target domain*, but labeled data from a *source domain* somewhat similar to the target as well as large amounts of unlabeled data from the target domain are at one's disposal. The domain adaptation problem then consists of leveraging the source labeled and target unlabeled data to derive a hypothesis performing well on the target domain.

The first theoretical analysis of the domain adaptation problem was presented by [1], who gave VC-dimension-based generalization bounds for adaptation in

classification tasks. Perhaps, the most significant contribution of that work was the definition and application of a distance between distributions, the d_A distance, that is particularly relevant to the problem of domain adaptation and which can be estimated from finite samples for a finite VC dimension, as previously shown by [10]. This work was later extended by [2] who also gave a bound on the error rate of a hypothesis derived from a weighted combination of the source data sets for the specific case of empirical risk minimization. More refined generalization bounds which apply to more general tasks, including regression and general loss functions appears in [12]. From an algorithmic perspective, it is natural to re-weight the empirical distribution to better reflect the target distribution; efficient algorithms for this re-weighting task were given in [12].

A more complex variant of this problem arises in sentiment analysis and other text classification tasks where the learner receives information from *several* domain sources that he can combine to make predictions about a target domain. As an example, often appraisal information about a relatively small number of domains such as *movies*, *books*, *restaurants*, or *music* may be available, but little or none is accessible for more difficult domains such as *travel*. This is known as the *multiple source adaptation problem*. Instances of this problem can be found in a variety of other natural language and image processing tasks.

The problem of adaptation with multiple sources was introduced and analyzed [13, 14]. The problem is formalized as follows. For each source domain $i \in [1, k]$, the learner receives the distribution of the input points Q_i , as well as a hypothesis h_i with loss at most ϵ on that source. The task consists of combining the k hypotheses h_i , $i \in [1, k]$, to derive a hypothesis h with a loss as small as possible with respect to the target distribution P . Unfortunately, a simple convex combination of the k source hypotheses h_i can perform very poorly; for example, there are cases where *any* such convex combination would incur a classification error of half, even when each source hypothesis h_i makes no error on its domain Q_i (see [13]). In contrast, *distribution weighted combinations* of the source hypotheses, which are combinations of source hypotheses weighted by the source distributions, perform very well. In [13] it was shown that, remarkably, for any fixed target function, there exists a distribution weighted combination of the source hypotheses whose loss is at most ϵ with respect to *any* mixture P of the k source distributions Q_i . For the case that the target distribution P is arbitrary, generalization bounds, based on *Rényi divergence* between the sources and the target distributions, were derived in [14].

References

1. S. Ben-David, J. Blitzer, K. Crammer, and F. Pereira. Analysis of representations for domain adaptation. *Proceedings of NIPS 2006*.
2. J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, and J. Wortman. Learning bounds for domain adaptation. *Proceedings of NIPS 2007*.
3. John Blitzer, Mark Dredze, and Fernando Pereira. Biographies, Bollywood, Boomboxes and Blenders: Domain Adaptation for Sentiment Classification. In *ACL 2007*.

4. Ciprian Chelba and Alex Acero. Adaptation of maximum entropy capitalizer: Little data can help a lot. *Computer Speech & Language*, 20(4):382–399, 2006.
5. Hal Daumé III and Daniel Marcu. Domain adaptation for statistical classifiers. *Journal of Artificial Intelligence Research*, 26:101–126, 2006.
6. Mark Dredze, John Blitzer, Pratha Pratim Talukdar, Kuzman Ganchev, Joao Graca, and Fernando Pereira. Frustratingly Hard Domain Adaptation for Parsing. In *CoNLL 2007*.
7. Jean-Luc Gauvain and Chin-Hui. Maximum a posteriori estimation for multivariate gaussian mixture observations of markov chains. *IEEE Transactions on Speech and Audio Processing*, 2(2):291–298, 1994.
8. Frederick Jelinek. *Statistical Methods for Speech Recognition*. The MIT Press, 1998.
9. Jing Jiang and ChengXiang Zhai. Instance Weighting for Domain Adaptation in NLP. In *Proceedings of ACL 2007*.
10. D. Kifer, S. Ben-David, and J. Gehrke. Detecting change in data streams. *Proceedings of the 30th International Conference on Very Large Data Bases*, 2004.
11. C. J. Legetter and Phil C. Woodland. Maximum likelihood linear regression for speaker adaptation of continuous density hidden markov models. *Computer Speech and Language*, pages 171–185, 1995.
12. Yishay Mansour, Mehryar Mohri, and Afshin Rostamizadeh. Domain adaptation: Learning bounds and algorithms. In *COLT*, 2009.
13. Yishay Mansour, Mehryar Mohri, and Afshin Rostamizadeh. Domain adaptation with multiple sources. *Proceedings of NIPS 2008*.
14. Yishay Mansour, Mehryar Mohri, and Afshin Rostamizadeh. Multiple source adaptation and the Rényi divergence. In *Uncertainty in Artificial Intelligence (UAI)*, 2009.
15. Aleix M. Martínez. Recognizing imprecisely localized, partially occluded, and expression variant faces from a single sample per class. *IEEE Trans. Pattern Anal. Mach. Intell.*, 24(6):748–763, 2002.
16. S. Della Pietra, V. Della Pietra, R. L. Mercer, and S. Roukos. Adaptive language modeling using minimum discriminant estimation. In *HLT '91: Proceedings of the workshop on Speech and Natural Language*, pages 103–106, 1992.
17. Brian Roark and Michiel Bacchiani. Supervised and unsupervised PCFG adaptation to novel domains. In *Proceedings of HLT-NAACL*, 2003.
18. Roni Rosenfeld. A Maximum Entropy Approach to Adaptive Statistical Language Modeling. *Computer Speech and Language*, 10:187–228, 1996.
19. Leslie G. Valiant. *A theory of the learnable*. Communication of the ACM 27(11):1134–1142, 1984.