Robust Semi-Supervised Learning through Implicit Constraints

Jesse H. Krijthe JKRIJTHE@GMAIL.COM
Pattern Recognition Laboratory, Delft University of Technology
Department of Molecular Epidemiology, Leiden University Medical Center

Marco Loog M.LOOG@TUDELFT.NL
Pattern Recognition Laboratory, Delft University of Technology
The Image Group, University of Copenhagen

Keywords: Semi-supervised learning, Classification, Least Squares

In many learning tasks, labeled data is relatively expensive to obtain, compared to readily available amounts of unlabeled data. Semi-supervised learning (SSL) considers the question of how to improve the generalization performance of supervised classification or regression procedures using this unlabeled data. Previous work on SSL has brought forth a variety of methods that reach this goal by leveraging assumptions on the relationship between the marginal distribution of the labeled and unlabeled input vectors $P(X)$ and the conditional distribution of interest $P(Y|X)$. For instance, one might assume the decision boundary resides in a region of low density, a popular approach considered, for instance, by the Transductive SVM (Joachims, 1999). These methods are effective when this assumption is true. When the assumption does not hold, however, these procedures may actually degrade the performance of a supervised procedure that ignores the unlabeled data.

Inspired by earlier work in (Loog, 2014), we consider whether it is possible to construct a procedure that is robust against this deterioration in performance. In Implicitly Constrained Semi-Supervised Learning (ICSSL) (Krijthe & Loog, 2014) this is done by only considering the set of solutions that could be obtained by the supervised learner, assuming every possible labeling of the unlabeled objects. To select a solution from the resulting set, we minimize the *supervised* objective function over this set of solutions. Additionally, we consider adaptations of this objective function with interesting theoretical performance guarantees.

The ICSSL framework has been applied to both a discriminative linear classifier using a quadratic loss function and the generative linear discriminant analysis. The novelty of this work is three fold:

First, in the discriminative case, theoretical results show that this framework leads to a semi-supervised classifier that *always* outperforms its supervised alternative in the transductive setting, when evaluated in terms of the loss function that the supervised learner is minimizing.

Secondly, unlike other approaches to semi-supervised learning, the objective can be written as a (quasi-)convex optimization problem where we can easily find a global optimum. Moreover, there is no additional hyperparameter involved to control the influence of the unlabeled data that needs to be set by the user.

Thirdly, empirical results indicate that the theoretical results also hold in practice, and the robustness does not only hold in terms of the surrogate loss, but also in terms of classification error.

Acknowledgments. This work was partly funded by project P24 of the Dutch public/private research network COMMIT.

References

Joachims, T. (1999). Transductive inference for text classification using support vector machines. *Proceedings of the 16th International Conference on Machine Learning* (pp. 200–209).

Krijthe, J. H., & Loog, M. (2014). Implicitly Constrained Semi-Supervised Linear Discriminant Analysis. *International Conference on Pattern Recognition*. Forthcoming.

Loog, M. (2014). Semi-supervised linear discriminant analysis through moment-constraint parameter estimation. *Pattern Recognition Letters*, *37*, 24–31.