Implicitly Constrained Semi-Supervised Linear Discriminant Analysis

Jesse H. Krijthe
Pattern Recognition Laboratory, Delft University of Technology
Department of Molecular Epidemiology, Leiden University Medical Center
jkrijthe@gmail.com

Marco Loog
Pattern Recognition Laboratory, Delft University of Technology
The Image Group, University of Copenhagen
m.loog@tudelft.nl

In many machine learning tasks, apart from a set of labeled data, a large amount of unlabeled observations is often available. The goal of semi-supervised learning is to use this unlabeled data to improve the supervised classification or regression model that was learned based on the labeled data alone. For classification using linear discriminant analysis (LDA) specifically, several semi-supervised variants have been proposed. Using any one of these methods is, however, not guaranteed to outperform the supervised classifier which does not take the additional unlabeled data into account [1]. They may, in fact, reduce performance.

To counter this problem, [4] introduced moment constrained LDA, which offers a more robust type of semi-supervised LDA. This approach required the identification of specific constraints that link parameter estimates that rely on the labeled data to parameters that do not rely on the labels. Ideally, we would like these constraints to emerge implicitly from the choice of the supervised learning model and a given set of unlabeled objects.

Implicitly constrained semi-supervised learning, introduced in [3] attempts to do just that. The underlying intuition is that if we could enumerate all possible labelings of the unlabeled data, and train the corresponding classifiers, the classifier based on the true but unknown labels is in this set. This classifier would generally outperform the supervised classifier. In practice, however, we can not enumerate over all possible labelings, nor do we know which one corresponds to the true labeling. One way to know how well any of these classifiers is going to perform is to estimate its performance using the supervised objective function evaluated on labeled objects alone. Based on this objective, it turns out one can efficiently find the optimal classifier in this set of possible classifiers by allowing for soft label assignments to the unlabeled objects. This all leads to a convex optimization problem that can be solved using a simple bounded gradient descent procedure.

We compare this novel constraint based approach to the moment constrained approach of [4] and to other semi-supervised methods, in particular, expectation maximization and self-learning. We also consider the question if and in what sense we can expect improvement in performance over the supervised procedure. The main conclusion from these analyses is that the constraint based approaches are more robust to misspecification of the original supervised model, and may outperform alternatives that make more assumptions on the data, in particular when performance is measured in terms of the log-likelihood of unseen objects.

This work was presented in [2], while the idea of implicitly constrained learning, applied to the least squares classifier, is described in [3].

Acknowledgments. This work was partly funded by project P23 of the Dutch public/private research network COMMIT.

References

[1] F. G. Cozman, I. Cohen, and M. C. Cirelo. Semi-Supervised Learning of Mixture Models. In *Proceedings of the Twentieth International Conference on Machine Learning*, 2003.

[2] J. Krijthe and M. Loog. Implicitly constrained semi-supervised linear discriminant analysis. In *Pattern Recognition (ICPR), 2014 22nd International Conference on*, pages 3762–3767, Aug 2014.

[3] J. H. Krijthe and M. Loog. Implicitly Constrained Semi-Supervised Least Squares Classification. Technical report, 2013.

[4] M. Loog. Semi-supervised linear discriminant analysis through moment-constraint parameter estimation. *Pattern Recognition Letters*, 37:24–31, Mar. 2014.