Implicitly Constrained Semi-Supervised Linear Discriminant Analysis

Jesse H. Krijthe[1,2]
Marco Loog[1,3]

1 Pattern Recognition Laboratory, Delft University of Technology
2 Department of Molecular Epidemiology, Leiden University Medical Center
3 The Image Group, Department of Computer Science, University of Copenhagen

Problem Setting

Semi-supervised learning refers to learning methods that attempt to leverage unlabeled data to improve supervised learners. In natural language processing, image processing and many other fields, unlabeled data is easier to obtain than high-quality labeled data. Effectively using this type of data would allow for increased prediction performance, without the need to obtain more labeled data.

For classification using linear discriminant analysis (LDA) specifically, several semi-supervised variants have been proposed. Using any one of these methods, however, is not guaranteed to outperform the supervised classifier which does not take the additional unlabeled data into account. Consider the situation in Figure 1, for instance. The well-known EM algorithm for missing data leads to worse results than supervised learner that discards the unlabeled data.

The goal of this work is to construct a semi-supervised version of LDA that is robust to deterioration in performance. We propose implicitly constrained LDA and study its performance, both in terms of error rates as well as in terms of log likelihood on the test set. We argue that the latter is a more sensible approach to evaluating a semi-supervised adaptation of LDA.

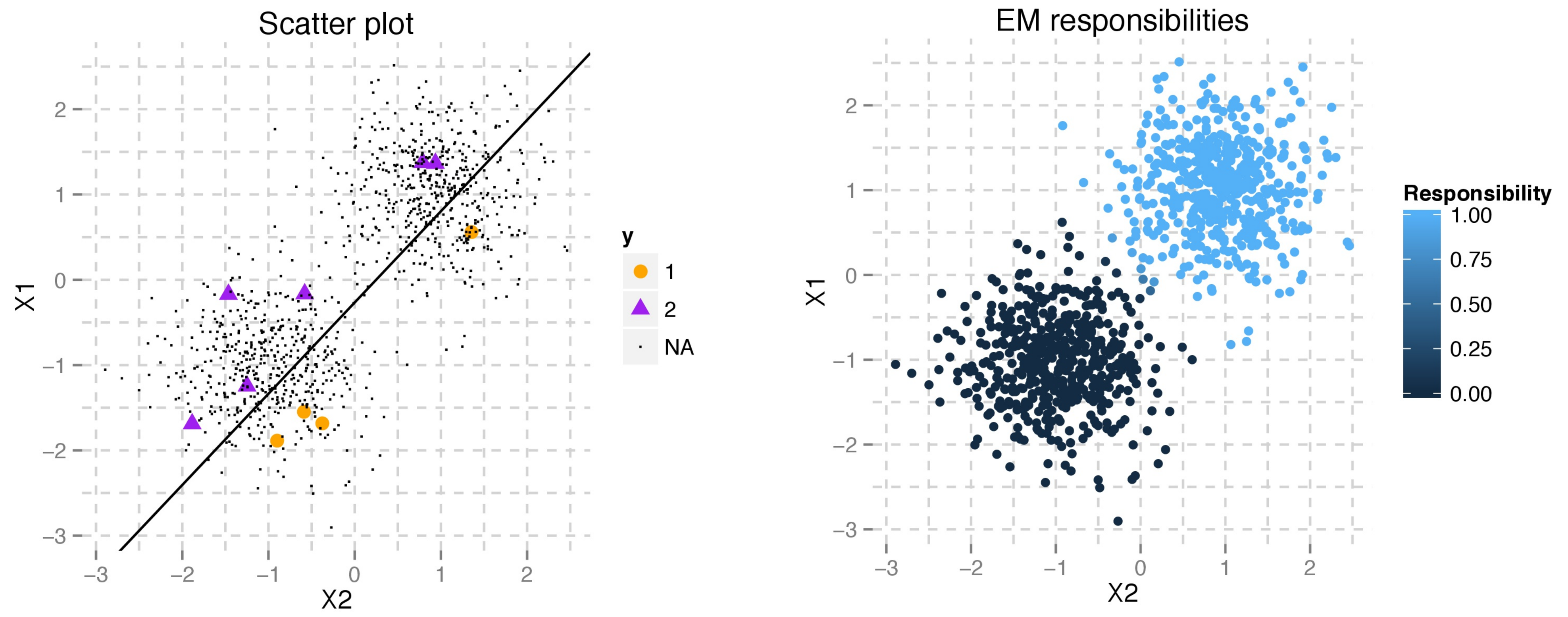

Figure 1: (left) Scatterplot of a 2 class semi-supervised learning problem. Black dots indicate unlabeled objects while the solid black line indicates the decision boundary of supervised Linear Discriminant Analysis trained using only the labeled objects. **(right)** A plot of the class responsibilities assigned to the unlabeled objects by a run of the Expectation Maximization algorithm applied to linear discriminant analysis. Note the inferred LDA solution is much worse than that of supervised LDA.

Implicitly Constrained LDA

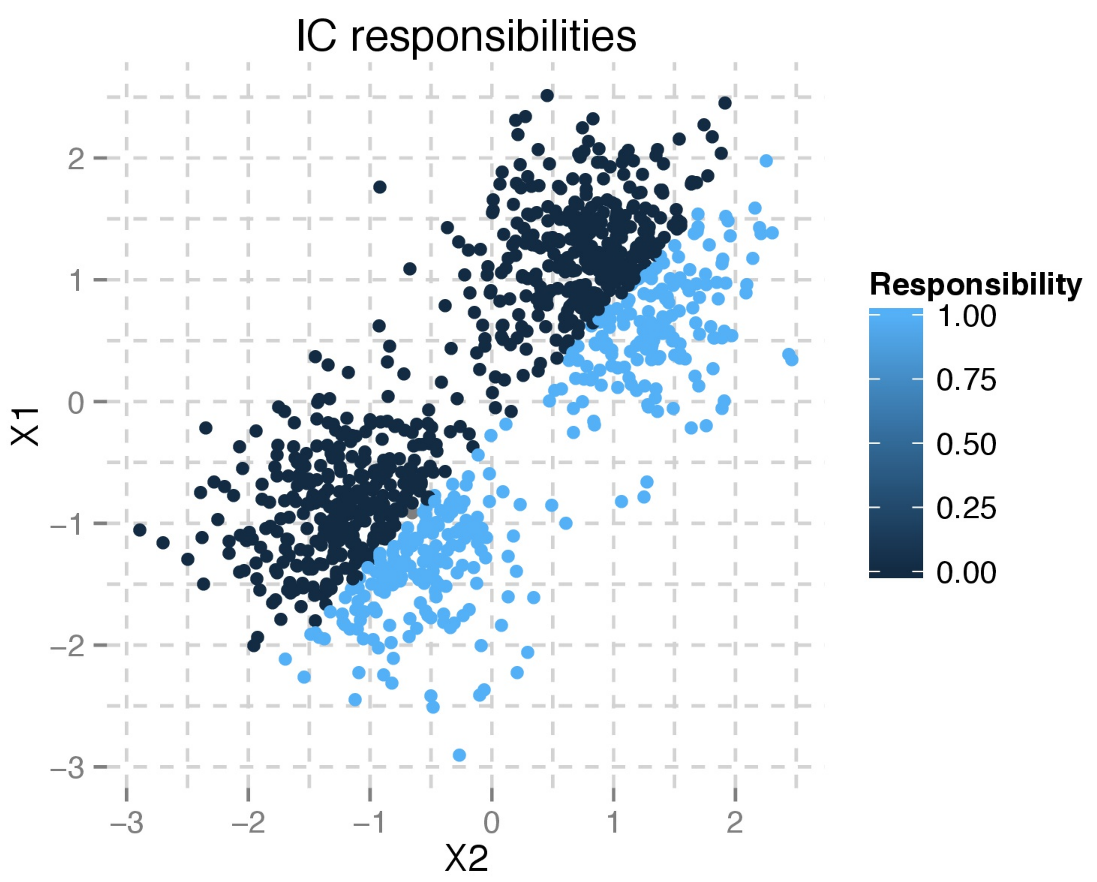

Figure 2: Continuing the example of Figure 1, the implicitly constrained linear discriminant analysis inference does not lead to a worse classifier as compared to the supervised LDA solution.

Most semi-supervised methods adapt the supervised objective function of a classifier by adding terms encoding additional assumptions that dependent on the unlabeled data. In EMLDA, for instance, the unlabeled data are added to the objective function by integrating out their unknown labels. The EM algorithm effectively converges to a local maximum of this solution by imputing the unknown responsibilities (partial labelings) in the E step and using these to estimate the parameters in the M step. As Figure 1 showed, this may lead to worse performance than supervised LDA.

Implicitly constrained LDA does not change the objective function. Instead, we constrain the possible solutions to the maximization of the objective function to be solutions that could be obtained by a possible labeling of the data. This way, we make sure we are still performing well on the labeled data, while also also taking into account that some solutions are no longer possible based on the unlabeled data.

While ICLDA is a conservative estimator, results on benchmark datasets show it still improves performance for many datasets.

LDA negative loss

$$L(\theta|\mathbf{X},\mathbf{y}) = \sum_{i=1}^{N_l} y_i \log(\pi_1 \mathcal{N}(\mathbf{x}_i|\boldsymbol{\mu}_1,\Sigma)) + (1-y_i)\log(\pi_2 \mathcal{N}(\mathbf{x}_i|\boldsymbol{\mu}_2,\Sigma))$$

EMLDA

$$L(\theta|\mathbf{X},\mathbf{y},\mathbf{X}_u) = \prod_{i=1}^{N_l} \left(\pi_1 \mathcal{N}(\mathbf{x}_i|\boldsymbol{\mu}_1,\Sigma)\right)^{y_i} \left(\pi_2 \mathcal{N}(\mathbf{x}_i|\boldsymbol{\mu}_2,\Sigma)\right)^{1-y_i} \times \prod_{i=1}^{N_u} \sum_{c=1}^{2} \pi_c \mathcal{N}(\mathbf{x}_i|\boldsymbol{\mu}_c,\Sigma)$$

ICLDA

$$\underset{(\pi_1,\pi_2,\boldsymbol{\mu}_1,\boldsymbol{\mu}_2,\boldsymbol{\Sigma})\in\mathcal{C}_\theta}{\arg\max} L(\pi_1,\pi_2,\boldsymbol{\mu}_1,\boldsymbol{\mu}_2,\boldsymbol{\Sigma}|\mathbf{X},\mathbf{y})$$

$$\mathcal{C}_{\boldsymbol{\theta}} = \left\{\arg\max L(\pi_1,\pi_2,\boldsymbol{\mu}_1,\boldsymbol{\mu}_2,\Sigma|\mathbf{X}_e,\mathbf{y}_e) : \mathbf{y}_u \in [0,1]^{N_u}\right\}$$

Discussion

Results on several benchmark datasets show that ICLDA provides the safest option among the semi-supervised approaches to LDA that we considered. At least in terms of the log-likelihood, it provides the best performance by far.

While it is the safest version, ICLDA may be too safe, in that it does not attain performance improvement in terms of classification error in many cases where MCLDA or the EM approaches do offer improvements. In terms of the loss on the test set, however, ICLDA is the best performing method. Since this is the objective minimized by supervised LDA as well, perhaps this is the best we could hope for in a true semi-supervised adaptation of LDA.

Apart from a new approach to semi-supervised LDA, the application of implicitly constrained learning to LDA shows that this framework is applicable to other classifiers beyond the least squares classifier studied in earlier work.

Table 1: Average 10-fold cross-validation error and its standard deviation over 20 repeats. Indicated in **bold** is whether a semi-supervised classifier significantly outperform the supervised LDA classifier, as measured using a t-test with a 0.05 significance level. Underlined indicates whether a semi-supervised classifier is (significantly) best among the four semi-supervised classifiers considered.

Dataset	LDA	LDAoracle	MCLDA	EMLDA	SLLDA	ICLDA
Haberman	0.37 ± 0.04	0.25 ± 0.00	0.36 ± 0.03	0.47 ± 0.08	0.36 ± 0.04	0.37 ± 0.04
Ionosphere	0.21 ± 0.02	0.15 ± 0.01	**0.18 ± 0.02**	0.57 ± 0.04	0.20 ± 0.02	**0.18 ± 0.01**
Parkinsons	0.27 ± 0.03	0.15 ± 0.01	**0.22 ± 0.03**	0.41 ± 0.05	0.26 ± 0.03	**0.23 ± 0.03**
Pima	0.34 ± 0.03	0.23 ± 0.00	**0.32 ± 0.02**	0.37 ± 0.03	0.35 ± 0.02	**0.31 ± 0.02**
Sonar	0.29 ± 0.02	0.26 ± 0.02	0.28 ± 0.02	0.35 ± 0.02	0.29 ± 0.02	0.28 ± 0.02
SPECT	0.31 ± 0.03	0.18 ± 0.01	**0.25 ± 0.02**	0.62 ± 0.03	0.33 ± 0.03	0.30 ± 0.03
SPECTF	0.32 ± 0.03	0.24 ± 0.01	**0.28 ± 0.03**	**0.28 ± 0.05**	0.34 ± 0.03	0.33 ± 0.03
Transfusion	0.34 ± 0.03	0.23 ± 0.00	**0.32 ± 0.03**	0.52 ± 0.09	0.37 ± 0.05	0.33 ± 0.03
WDBC	0.11 ± 0.01	0.04 ± 0.00	**0.09 ± 0.01**	0.38 ± 0.05	**0.09 ± 0.01**	**0.08 ± 0.01**
BCI	0.21 ± 0.01	0.16 ± 0.01	**0.20 ± 0.01**	0.21 ± 0.02	0.21 ± 0.02	**0.20 ± 0.01**

Table 2: Average 10-fold cross-validation negative log-likelihood (loss) and its standard deviation over 20 repeats. Indicated in **bold** is whether a semi-supervised classifier significantly outperform the supervised LDA classifier, as measured using a t-test with a 0.05 significance level. Underlined indicates whether a semi-supervised classifier is (significantly) best among the four semi-supervised classifiers considered.

Dataset	LDA	LDAoracle	MCLDA	EMLDA	SLLDA	ICLDA
Haberman	15.88 ± 4.37	10.37 ± 0.02	**11.66 ± 2.45**	**12.02 ± 0.35**	**12.08 ± 0.20**	**10.89 ± 0.16**
Ionosphere	199.58 ± 29.66	21.38 ± 0.34	**25.93 ± 1.44**	**22.55 ± 0.40**	**22.80 ± 0.40**	**22.22 ± 0.33**
Parkinsons	-40.76 ± 11.11	-71.87 ± 0.32	**-71.05 ± 0.40**	**-71.12 ± 0.40**	**-71.03 ± 0.38**	**-71.44 ± 0.31**
Pima	41.98 ± 2.99	29.88 ± 0.02	**31.74 ± 0.99**	**31.95 ± 0.35**	**32.07 ± 0.36**	**30.50 ± 0.13**
Sonar	-59.86 ± 1.08	-83.05 ± 0.59	**-82.23 ± 0.57**	**-82.85 ± 0.55**	**-82.20 ± 0.60**	**-82.58 ± 0.57**
SPECT	27.65 ± 1.89	10.74 ± 0.09	**11.30 ± 0.17**	**12.63 ± 0.18**	**11.84 ± 0.20**	**11.19 ± 0.13**
SPECTF	178.42 ± 2.48	148.13 ± 0.68	**148.78 ± 0.69**	**148.44 ± 0.69**	**149.18 ± 0.72**	**148.67 ± 0.71**
Transfusion	17.00 ± 2.61	11.48 ± 0.02	**12.23 ± 0.54**	16.27 ± 0.53	**14.21 ± 0.47**	**11.88 ± 0.17**
WDBC	33.15 ± 15.14	-28.06 ± 1.29	**-26.73 ± 1.23**	**-26.67 ± 1.32**	**-27.78 ± 1.28**	**-27.86 ± 1.28**
BCI	6.99 ± 1.04	-21.04 ± 0.41	**-20.38 ± 0.40**	**-20.39 ± 0.46**	**-20.44 ± 0.45**	**-20.74 ± 0.41**

ICLDA is a principled and robust adaptation of linear discriminant analysis to the semi-supervised setting. In terms of error rates, it may be overly conservative. When measured in terms of the loss on the test set, however, it outperforms other semi-supervised methods. The usefulness of implicit constraints to LDA indicate that this idea is applicable to a range of classifiers.

Leiden University Medical Center
Delft University of Technology
COMMIT/