Implicitly Constrained Semi-Supervised Linear Discriminant Analysis

Jesse H. Krijthe*‡, Marco Loog†‡
jkrijthe@gmail.com, m.loog@tudelft.nl
*Pattern Recognition Laboratory, Delft University of Technology, The Netherlands
†The Image Group, Department of Computer Science, University of Copenhagen, Denmark
‡Department of Molecular Epidemiology, Leiden University Medical Center, The Netherlands

***Abstract*—Semi-supervised learning is an important and active topic of research in pattern recognition. For classification using linear discriminant analysis specifically, several semi-supervised variants have been proposed. Using any one of these methods is not guaranteed to outperform the supervised classifier which does not take the additional unlabeled data into account. In this work we compare traditional Expectation Maximization type approaches for semi-supervised linear discriminant analysis with approaches based on intrinsic constraints and propose a new principled approach for semi-supervised linear discriminant analysis, using so-called implicit constraints. We explore the relationships between these methods and consider the question if and in what sense we can expect improvement in performance over the supervised procedure. The constraint based approaches are more robust to misspecification of the model, and may outperform alternatives that make more assumptions on the data in terms of the log-likelihood of unseen objects.**

I. Introduction

In many real-world pattern recognition tasks, obtaining labeled examples to train classification algorithms is much more expensive than obtaining unlabeled examples. These tasks include document and image classification [1] where unlabeled objects can easily be downloaded from the web, part of speech tagging [2], protein function prediction [3] and many others. Using unlabeled data to improve the training of a classification procedure, however, requires semi-supervised variants of supervised classifiers to make use of this additional unlabeled data. Research into semi-supervised learning has therefore seen an increasing interest in the last decade [4].

Unlike in supervised learning, where adding additional labeled training data improves performance for most classification routines, this does not generally hold for semi-supervised learning [5]. Adding additional unlabeled data may actually deteriorate classification performance. This can happen when the underlying assumptions of the model do not hold. In effect, disregarding the unlabeled data may have been a better solution.

In this work we consider the well-known linear discriminant analysis applied to classification. Several adaptations of this supervised procedure have been proposed. These approaches may suffer from the problem that additional unlabeled data degrade performance. To counter this problem, [6] introduced moment constrained LDA, which offers a more robust type of semi-supervised LDA. The recently introduced idea of implicitly constrained estimation [7], is another method that relies on constraints given by the unlabeled data. We compare these two approaches to other semi-supervised methods, in particular, expectation maximization and self-learning, and empirically study in what sense we can expect improvement by employing any of these semi-supervised methods.

The contributions of this work are the following:

- Introduce a new, principled approach to semi-supervised LDA: implicitly constrained LDA
- Offer a comparison of semi-supervised versions of linear discriminant analysis
- Explore ways in which we can expect these semi-supervised methods to offer improvements over the supervised variant, in particular in terms of the log likelihood

The rest of this paper is organized as follows. After discussing related work, we introduce several approaches to semi-supervised linear discriminant analysis. These methods are then compared on an illustrative toy problem and in an empirical study using several benchmark datasets. We end with a discussion of the results and conclude.

II. Related Work

Some of the earliest work on semi-supervised learning was done by [8], [9] who studied the self-learning approach applied to linear discriminant analysis. Another example of such an approach is Yarowsky's algorithm [10]. This is closely related to Expectation Maximization [11], where, in a generative model, the unknown labels are integrated out of the likelihood function and the resulting marginal likelihood is maximized [12]. More recent work on discriminative semi-supervised learning has focussed on introducing assumptions that relate unlabeled data to the labeled objects [4]. These assumptions usually take the form of either a manifold assumption [13], encoding that labels change smoothly in a low-dimensional manifold, or a low-density class separation assumption used in, for instance, transductive support vector machines [14], [15] and Entropy regularization [16].

Work on semi-supervised LDA has tried to incorporate unlabeled data by leveraging the increase in accuracy of the estimators of some quantities that do not rely on labels. An approach relying on the more accurate estimate of the total covariance matrix of both labeled and unlabeled objects is taken for dimensionality reduction in Normalized LDA, proposed by [17] and similar work by [18]. In addition to this covariance matrix, [6] also include the more accurate estimate of the overal mean of the data and propose two solutions to solve a subsequent optimization problem. Building on these results, in [7] we introduced implicitly constrained least squares classification, a semi-supervised adaptation of least squares classification. Since this procedure proved proved both theoretically and practically successful for a discriminative classifier, here we consider whether the idea of implicitly constrained semi-supervised learning can also be applied to a generative classifier such as LDA.

III. Methods

We will first introduce linear discriminant analyzing as a supervised classification algorithm and discuss alternative semi-supervised procedures. We will consider 2-class classification problems, where we are given an $N_l \times d$ design matrix $\mathbf{X}$, where N_l is the number of labeled objects and d is the number of features. For these observations we are given a label vector $\mathbf{y} = \{0,1\}^{N_l}$. Additionally, in the semi-supervised setting we have an $N_u \times d$ design matrix $\mathbf{X_u}$ without a corresponding $\mathbf{y}_u$ for the unlabeled observations.

A. Supervised LDA

In supervised linear discriminant analysis, we model the 2 classes as having multivariate normal distributions with the same covariance matrix $\mathbf{\Sigma}$ and differing means $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$. To estimate the parameters of the Gaussian, we maximize the likelihood, or, equivalently, the log likelihood function:

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

where $\theta = \{\pi_1, \pi_2, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \mathbf{\Sigma}\}$, $\mathcal{N}(\mathbf{x}_i|\boldsymbol{\mu},\mathbf{\Sigma})$ denotes the density of a multivariate normal distribution with mean $\boldsymbol{\mu}$ and covariance $\mathbf{\Sigma}$ evaluated at $\mathbf{x}_i$ and π_c denotes the prior probability for class c. The closed form solution to this maximization is given by the estimators:

$$\begin{aligned} \hat{\pi}_1 &= \frac{\sum_{i=1}^{N_l} y_i}{N_l}, \qquad \hat{\pi}_2 = \frac{\sum_{i=1}^{N_l}(1-y_i)}{N_l} \\ \hat{\boldsymbol{\mu}}_1 &= \frac{\sum_{i=1}^{N_l} y_i \mathbf{x}_i}{\sum_{i=1}^{N_l} y_i}, \qquad \hat{\boldsymbol{\mu}}_2 = \frac{\sum_{i=1}^{N_l}(1-y_i)\mathbf{x}_i}{\sum_{i=1}^{N_l}(1-y_i)} \\ \hat{\mathbf{\Sigma}} &= \frac{1}{N_l}\sum_{i=1}^{N_l} y_i(\mathbf{x}_i-\boldsymbol{\mu}_1)(\mathbf{x}_i-\boldsymbol{\mu}_1)^T \\ & \quad + (1-y_i)(\mathbf{x}_i-\boldsymbol{\mu}_2)(\mathbf{x}_i-\boldsymbol{\mu}_2)^T \end{aligned} \tag{2}$$

Where the maximum likelihood estimator $\hat{\mathbf{\Sigma}}$ is a biased estimator for the covariance. Given a set of labeled objects $(\mathbf{X},\mathbf{y})$, we can estimate these parameters and find the posterior for a new object $\mathbf{x}$ using:

$$p(c=1|\mathbf{x}) = \frac{\pi_1 \mathcal{N}(\mathbf{x}|\hat{\boldsymbol{\mu}}_1,\hat{\mathbf{\Sigma}})}{\sum_{c=1}^{2}\pi_c \mathcal{N}(\mathbf{x}|\hat{\boldsymbol{\mu}}_c,\hat{\mathbf{\Sigma}})} \tag{3}$$

This posterior distribution can be employed for classification by assigning objects to the class for which its posterior is highest. We now consider several adaptations of this classification procedure to the semi-supervised setting.

B. Self-Learning LDA (SLLDA)

A common and straightforward adaptation of any supervised learning algorithm to the semi-supervised setting is self-learning, also known as bootstrapping or Yarowsky's algorithm [8], [10]. Starting out with a classifier trained on the labeled data only, labels are predicted for the unlabeled objects. These objects, with their imputed labels are then used in the next iteration to retrain the classifier. This new classifier is now used to relabel the unlabeled objects. This is done until the predicted labels on the unlabeled objects converges. [11] study the underlying loss that this procedure minimizes and prove its convergence.

C. Expectation Maximization LDA (EMLDA)

Assuming the mixture model of Equation 1 and treating the unobserved labels as latent variables $\mathbf{y}_u$, a possible adaptation to this objective is to add a term for the unlabeled data to the objective and to integrate out the latent variable to find the marginal likelihood:

$$\begin{aligned} l(\theta|\mathbf{X},\mathbf{y},\mathbf{X}_u) = \prod_{i=1}^{N_l} & (\pi_1 \mathcal{N}(\mathbf{x}_i|\boldsymbol{\mu}_1,\mathbf{\Sigma}))^{y_i} (\pi_2 \mathcal{N}(\mathbf{x}_i|\boldsymbol{\mu}_2,\mathbf{\Sigma}))^{1-y_i} \\ & \times \prod_{i=1}^{N_u}\sum_{c=1}^{2}\pi_c \mathcal{N}(\mathbf{x}_i|\boldsymbol{\mu}_c,\Sigma) \end{aligned} \tag{4}$$

Note the unknown labels $\mathbf{y}_u$ are integrated out. Maximizing this marginal likelihood, or equivalently, the log of this function is harder then the supervised objective in Equation 1, since the expression contains a log over a sum and is no longer convex. However, we can solve this optimization problem using the well known expectation maximization (EM) algorithm [12], [1]. In EM, the log over the sum is bounded from below through Jensen's inequality. In the M step of the algorithm, we maximize this bound by updating the parameters using the imputed labels obtained in the E step. In practice, the M step consists of the same update as in 2, where the sum is no longer over the labeled objects but also the unlabeled objects using the imputed posteriors, or responsibilities, from the E step. In the E step the lower bound is made tight by updating the imputed labels using the posterior under the new parameter estimates. This is done until convergence. In effect this procedure is very similar to self-learning, where except instead of hard labels, a probability over labelings is used. Both self-learning and EM suffer from the problem wrongly imputed labels can reinforce themselves, because the parameters are updated as if they were the true labels.

D. Moment Constrained LDA (MCLDA)

An alternative to the EM-like approaches like EMLDA and SLLDA was proposed by [6] in the form of moment constrained parameter estimation. The main idea is that there are certain constraints that link parameters that are calculated using feature values alone, with parameters which require the labels. In the case of LDA, for instance, the overal mean of the data is linked to the means of the two classes through:

$$\boldsymbol{\mu}_t = \pi_1 \boldsymbol{\mu}_1 + \pi_2 \boldsymbol{\mu}_2 \tag{5}$$

Were $\boldsymbol{\mu}_t$ is the overal mean on all the data and therefore does not depend on the labels. The total covariance matrix $\boldsymbol{\Sigma}_t$ is linked to the within covariance matrix $\boldsymbol{\Sigma}$ and between covariance matrix $\boldsymbol{\Sigma}_b$, the covariance matrix of the means. Only the latter two rely on the labels:

$$\boldsymbol{\Sigma}_t = \boldsymbol{\Sigma} + \boldsymbol{\Sigma}_b \tag{6}$$

Recognizing that the unlabeled data allows us to more accurately estimate the parameters in these constraints that do not rely on the labels, [6] point out that this more accurate estimate will generally violate the constraints, meaning the other label-dependent estimates should be updated accordingly.

An ad hoc way to update the parameters based on these more accurate estimates [6] leads to the following updated moment constraint estimators:

$$\hat{\boldsymbol{\mu}}_c^{MC} = \hat{\boldsymbol{\mu}}_c - \sum_{j=1}^{2} \hat{\pi}_j \hat{\boldsymbol{\mu}}_j - \hat{\boldsymbol{\mu}}_t \tag{7}$$

$$\hat{\boldsymbol{\Sigma}}^{MC} = \hat{\boldsymbol{\Theta}}^{\frac{1}{2}} \hat{\boldsymbol{\Sigma}}_t^{\frac{1}{2}} \hat{\boldsymbol{\Sigma}} \hat{\boldsymbol{\Sigma}}_t^{\frac{1}{2}} \hat{\boldsymbol{\Theta}}^{\frac{1}{2}} \tag{8}$$

where $\hat{\boldsymbol{\mu}}_t$ and $\hat{\boldsymbol{\Theta}}$ are the overal mean and overal covariance estimated on all labeled and unlabeled data, while $\hat{\boldsymbol{\Sigma}}_t$ is the overal covariance estimated on the labeled data alone.

Alternatively and slightly more formally, [19] force the constraint by maximizing the likelihood on the labeled objects under the constraints in Equations (5) and (6). This leads to a non-convex objective function that can be solved numerically. In this work we use the ad hoc constraints.

E. Implicitly Constrained LDA (ICLDA)

The former approach requires the identification of specific constraints. Ideally, we would like these constraints to emerge implicitly from a choice of supervised learner and a given set of unlabeled objects. Implicitly constrained semi-supervised learning attempts to do just that. The underlying intuition is that if we could enumerate all possible 2^{N_u} labelings, and train the corresponding classifiers, the classifier based on the true labels is in this set. This classifier would generally outperform the supervised classifier. Two problems arise:

1) How do we find classifier in this set that is close to the one based on the true but unknown labels?
2) How do we efficiently traverse this enormous set of possible labelings without having to enumerate them all?

As for the first problem: the only way to know how well a solution helps us classify is to estimate its performance using the labeled points. We therefore propose the following objective:

$$\underset{\{\pi_1,\pi_2,\boldsymbol{\mu}_1,\boldsymbol{\mu}_2,\boldsymbol{\Sigma}\}\in\mathcal{C}_\theta}{\arg\max} \quad L(\pi_1, \pi_2, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\Sigma} | \mathbf{X}, \mathbf{y}) \tag{9}$$

where

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

and $\mathbf{X}_e = [\mathbf{X}^T \mathbf{X}_u]^T, \mathbf{y}_e = [\mathbf{y}^T \mathbf{y}_u^T]^T$ are the design matrix and class vector extended with the unlabeled data. This can be interpreted as optimizing the same objective function as supervised LDA, with the additional constraint that the solution has to attainable by a particular assignment of responsibilities (partial assignments to classes) for the unlabeled objects.

As for the second problem: since, for a given imputed labeling, we have a closed form solution for the parameters, the gradient of the supervised loss (9) with respect to the responsibilities $\mathbf{y}_u$ can be found using

$$\frac{\partial L(\theta|\mathbf{X},\mathbf{y})}{\partial \mathbf{y_u}} = \frac{\partial L(\theta|\mathbf{X},\mathbf{y})}{\partial \theta} \frac{\partial \phi(\mathbf{y}_u)}{\partial \mathbf{y}_u} \tag{10}$$

where $\phi(\mathbf{y}_u = \theta$ is the function that has as input a particular labeling of the points, and outputs the parameters $\theta = \{\pi_1, \pi_2, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\Sigma}\}$, similar to Equation 2.

This can be used in a simple gradient ascent procedure, taking into account the $[0,1]$ bounds on the responsibilities.

IV. Experimental setup and results

We present simulations on an illustrative toy dataset and a set of benchmark datasets. Other than the classifiers covered in the previous section, we also include the LDA classifier train using all labels of the unlabeled data as well (LSoracle) as an upper bound on the performance of any semi-supervised procedure. The experiments can be reproduced using code from the authors' website.

A. Toy problems

To illustrate the behaviour of ICLDA when compared to EMLDA we consider two toy datasets. In both cases we have two multivariate normal distributions centered at respectively $\boldsymbol{\mu}_1 = [1,1]^T$ and $\boldsymbol{\mu}_2 = [-1,-1]^T$ and equal covariance $\boldsymbol{\Sigma} = 0.6\mathbb{1}$, with $\mathbb{1}$ the 2×2 identity matrix. An example is given in Figure 1). In the bottom row, these two gaussians correspond to the different classes. In the top row, we consider the case where the decision boundary is actually perpendicular to the boundary in the other setting. This means that the bottom row corresponds exactly to the assumptions of EM, while this is not the case in the top row. Figure 1 illustrates what happens in a particular sample from this problem were we draw 10 labeled and 990 unlabeled objects. When the assumption does not hold, EMLDA forces the decision boundary to fall between the two gaussian clusters leading to a much worse solution then the supervised LDA based on only a few data samples. The ICLDA solution does not deviate from the correct boundary by much. When the assumptions do hold, EMLDA finds the correct boundary, as expected, while ICLDA only does minor adjustments in this case. While, one could claim that ICLDA is more robust, it might also be that it never leads to any improvement. Figure 2 shows the results from resampling from the data distribution in the second example and show that indeed ICLDA does lead to improvement on average, while not making the mistake in the first dataset where the LDA assumptions do not hold.

B. Simulations on Benchmark datasets

We test the behaviour of the considered procedures using datasets from the UCI machine learning repository [20], as well from [4]. Their characteristics can be found in Table I.

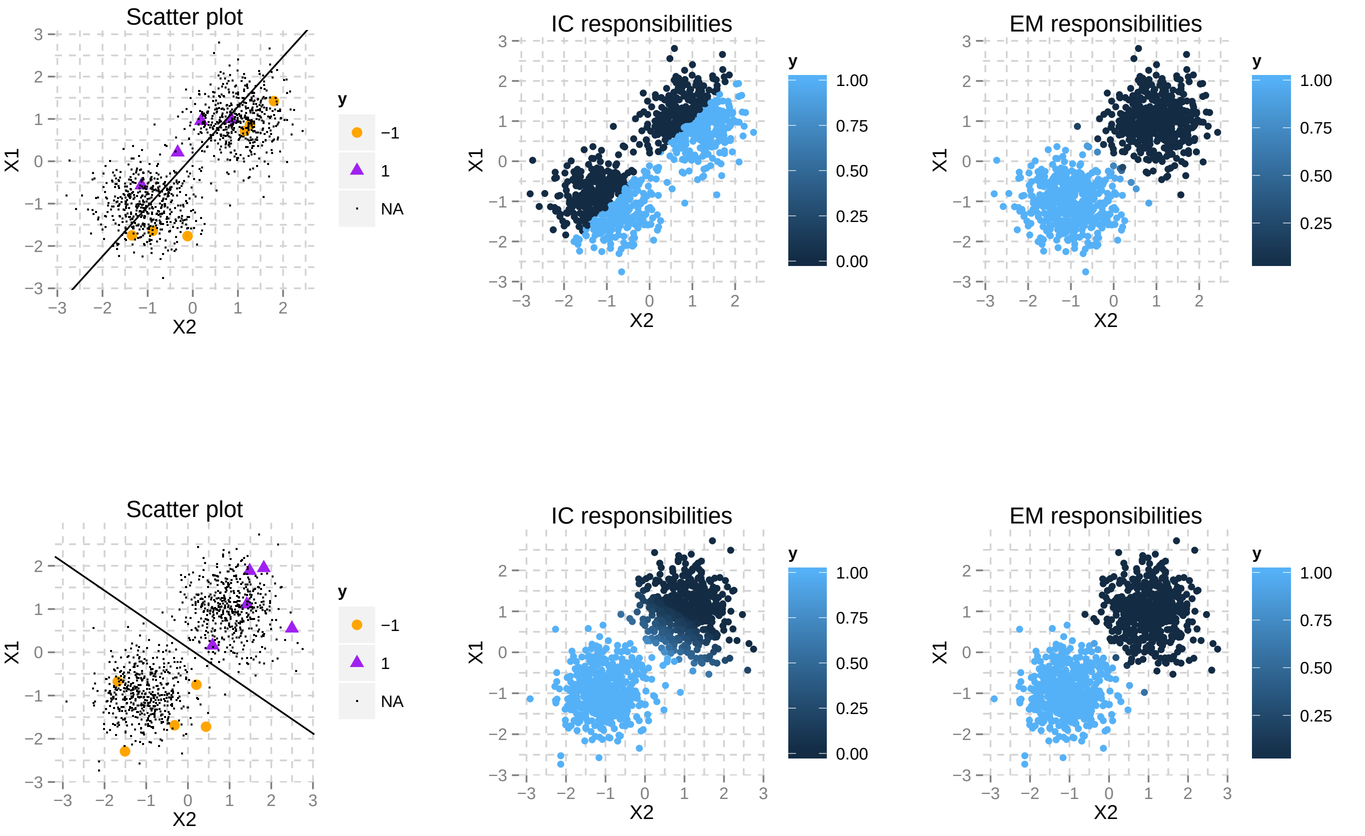

Fig. 1. Behaviour on the two-class two dimensional gaussian datasets, with 10 labeled objects and 990 unlabeled objects. The first row shows the scatterplot and the trained responsibilities for respectively ICLDA and EMLDA on a dataset where the decision boundary does not adhere to the assumptions of EM. The second row shows the results when the decision boundary is in between the two Gaussian classes. The black line boundary indicates the decision boundary of a supervised learner trained only on the labeled data. Note that in the first row, the responsibilities of EM are very different from the true labels, while IC is not as sensitive to this problem.

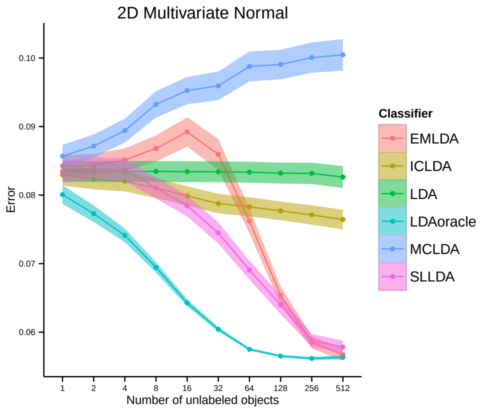

Fig. 2. Semi-supervised learning curve on the Gaussian data set using 500 repeats. The shaded regions indicate one standard error around the mean. Since its assumptions hold exactly, SLLDA and EMLDA work very well. ICLDA also outperforms the supervised LDA.

A cross-validation experiment was carried out as follows. Each of the datasets were split into 10 folds. Every fold was used as a validation set once, while the other nine folds were used for testing. The data in these nine folds was randomly split into a labeled and an unlabeled part, where the labeled part had $\min(2d, 10)$ objects, while the rest was used as unlabeled objects. This procedure was repeated 20 times and the average error and average log likelihood on the test set was determined. The results can be found in Tables II and III.

TABLE I. DESCRIPTION OF THE DATASETS USED IN THE EXPERIMENTS

Name	# Objects	#Features	Source
Haberman	305	4	[20]
Ionosphere	351	33	[20]
Parkinsons	195	20	[20]
Pima	768	9	[20]
Sonar	208	61	[20]
SPECT	265	23	[20]
SPECTF	265	45	[20]
Transfusion	748	4	[20]
WDBC	568	30	[20]
BCI	400	118	[4]

To study the behaviour of these classifiers for differing amount of unlabeled data, we estimated semi-supervised learning curves by randomly drawing $\min(2d, 10)$ labeled objects from the datasets, and using an increasing randomly chosen set as unlabeled data. The remaining objects formed the test set. This procedure was repeated 500 times and the average and standard error of the classification error and negative log likelihood were determined. The learning curves for 3 datasets can be found in Figure 3.

We find that overal in terms of error rates (Table II), MCLDA seems to perform best, being both more robust then the EM approaches as well as effective in using the unlabeled information in improving error rates. While ICLDA is robust

TABLE II. Average 10-fold cross-validation error and its standard deviation over 10 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 three 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	$\mathbf{0.18 \pm 0.02}$	0.57 ± 0.04	0.20 ± 0.02	$\mathbf{0.18 \pm 0.01}$
Parkinsons	0.27 ± 0.03	0.15 ± 0.01	$\mathbf{0.22 \pm 0.03}$	0.41 ± 0.05	0.26 ± 0.03	$\mathbf{0.23 \pm 0.03}$
Pima	0.34 ± 0.03	0.23 ± 0.00	$\mathbf{0.32 \pm 0.02}$	0.37 ± 0.03	0.35 ± 0.02	$\mathbf{0.31 \pm 0.02}$
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	$\mathbf{0.25 \pm 0.02}$	0.62 ± 0.03	0.33 ± 0.03	0.30 ± 0.03
SPECTF	0.32 ± 0.03	0.24 ± 0.01	$\mathbf{0.28 \pm 0.03}$	$\mathbf{0.28 \pm 0.05}$	0.34 ± 0.03	0.33 ± 0.03
Transfusion	0.34 ± 0.03	0.23 ± 0.00	$\mathbf{0.32 \pm 0.03}$	0.52 ± 0.09	0.37 ± 0.05	0.33 ± 0.03
WDBC	0.11 ± 0.01	0.04 ± 0.00	$\mathbf{0.09 \pm 0.01}$	0.38 ± 0.05	$\mathbf{0.09 \pm 0.01}$	$\mathbf{0.08 \pm 0.01}$
BCI	0.21 ± 0.01	0.16 ± 0.01	$\mathbf{0.20 \pm 0.01}$	0.21 ± 0.02	0.21 ± 0.02	$\mathbf{0.20 \pm 0.01}$

TABLE III. Average 10-fold cross-validation negative log-likelihood (loss) and its standard deviation over 10 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 three semi-supervised classifiers considered.

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

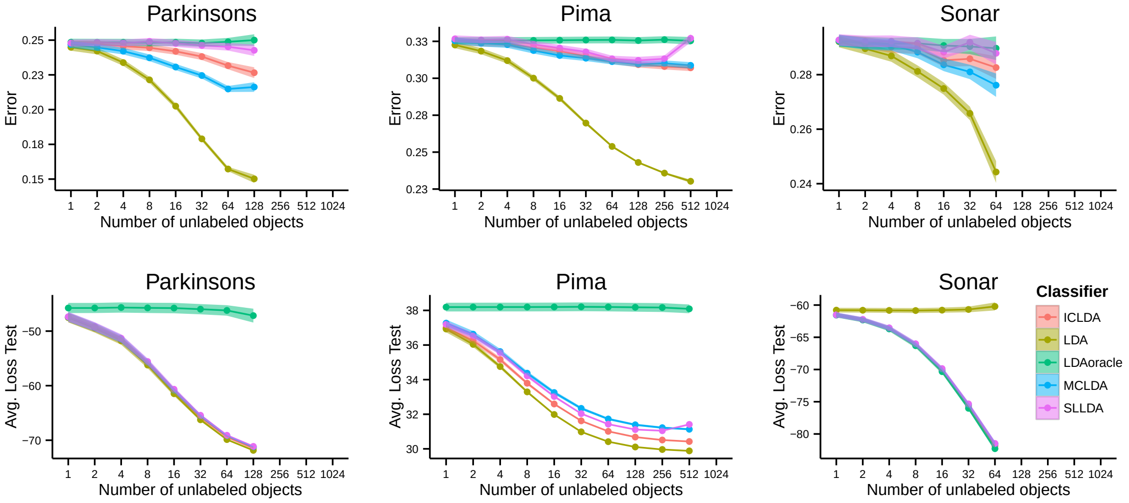

Fig. 3. Learning curves for increasing amounts of unlabeled data for the error rate as well as the loss (negative log likelihood) for three datasets using 500 repeats. The shaded regions indicate one standard error around the mean.

and has the best performance on 2 of the datasets, it is conservative in that it does not show improvements in terms of the classification error for many datasets where the other classifiers do offer improvements.

The picture is markedly different when we consider the log likelihood criterion that supervised LDA optimizes, when evaluated on the test set (Table III). Here ICLDA outperforms all other methods on the majority of the datasets.

V. Discussion

The results indicate that ICLDA forms the safest option that we considered for a semi-supervised version of LDA. An interesting observation is that its constraints are apparently fairly different from those of MCLDA. This is interesting because both approaches attempt to optimize the supervised loss function, under constraints. Furthermore, the constraints of ICLDA always imply that the MCLDA constraints also hold. ICLDA constraint may therefore be stricter then MCLDA, while they were derived in a more principled way. It is important to note that all of the methods considered here adhere to the constraints in Equation (5) and (6). However, the objective functions that are minimized are different.

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 were MCLDA or the EM approaches do offer improvements. In terms of the loss on the test set, however, ICLDA is the clear 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. We found similar empirical and theoretical performance results in terms of improvements in the loss on the test set when applying the implicitly constrained framework to the least squares classifier [7]. How then, this improvement in "surrogate" loss relates to the eventual goal of classification error, is unclear, especially for a non-margin based loss function such as the log likelihood [21]. However, since ICLDA does offer the best behaviour of supervised LDA's loss on the test set, ICLDA could be considered a step towards a principled semi-supervised version of LDA.

An open question is how the optimization of ICLDA behaves. The solution in terms of the responsibility vector $\mathbf{y}_u$ is non-unique: different labelings of the points can lead to the same parameters. In terms of the parameters however, we the optimization seems to converge to a unique global optimum. While we do not have a formal proof of this, such as in the case of implicitly constrained least squares classification, it is interesting that implicit constraints also lead to a convex problem for LDA.

We find that the behaviour of EMLDA is more erratic than that of SLLDA. The hard label assignments could have a regularizing effect on the semi-supervised solutions, making self-learning a good and fast alternative to self-learning. Note that safer versions of SLLDA and EMLDA could be obtained by introducing a weight parameter to control the influence of the unlabeled data [8]. In the limited labeled data setting, it is hard to correctly set this parameter. While this may help when dealing with larger sample sizes, the constraint approaches bring us closer to methods that always perform at least as well as their supervised counterpart.

VI. Conclusion

ICLDA is a principled and robust adaptation of LDA to the semi-supervised setting. In terms of error rates, it may be overly conservative. When measured in terms of the loss on the training set, however, it outperforms other semi-supervised methods. It therefore seems that there are opportunities for robust semi-supervised learners, although the performance criterion that we should consider may not be the error rate, but rather the loss that the supervised learner minimizes.

Acknowledgment

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

References

[1] K. Nigam, A. K. McCallum, S. Thrun, and T. Mitchell, "Text classification from labeled and unlabeled documents using EM," *Machine learning*, vol. 34, pp. 1–34, 2000. [Online]. Available: http://link.springer.com/article/10.1023/A:1007692713085

[2] D. Elworthy, "Does Baum-Welch re-estimation help taggers?" in *Proceedings of the fourth conference on Applied natural language processing*, 1994, pp. 53–58. [Online]. Available: http://dl.acm.org/citation.cfm?id=974371

[3] J. Weston, C. Leslie, E. Ie, D. Zhou, A. Elisseeff, and W. S. Noble, "Semi-supervised protein classification using cluster kernels." *Bioinformatics*, vol. 21, no. 15, pp. 3241–7, Aug. 2005. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/15905279

[4] O. Chapelle, B. Schölkopf, and A. Zien, *Semi-supervised learning*. MIT press, 2006.

[5] F. Cozman and I. Cohen, "Risks of Semi-Supervised Learning," in *Semi-Supervised Learning*, O. Chapelle, B. Schölkopf, and A. Zien, Eds. MIT press, 2006, ch. 4, pp. 56–72.

[6] M. Loog, "Semi-supervised linear discriminant analysis through moment-constraint parameter estimation," *Pattern Recognition Letters*, vol. In press, Mar. 2013. [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/S0167865513000913

[7] J. H. Krijthe and M. Loog, "Implicitly Constrained Semi-Supervised Least Squares Classification," Tech. Rep., 2013. [Online]. Available: www.jessekrijthe.com/papers/krijthe2013.pdf

[8] G. J. McLachlan, "Iterative Reclassification Procedure for Constructing an Asymptotically Optimal Rule of Allocation in Discriminant Analysis," vol. 70, no. 350, pp. 365–369, 1975.

[9] P. Taylor and G. J. Mclachlan, "Estimating the Linear Discriminant Function from Initial Samples Containing a Small Number of Unclassified Observations Estimating the Linear Discriminant Function from Initial Samples Containing a Small Number of Unclassified Observations," *Journal of the American Statistical Association*, vol. 72, no. 358, pp. 403–406, 1977.

[10] D. Yarowsky, "Unsupervised word sense disambiguation rivaling supervised methods," *Proceedings of the 33rd annual meeting on Association for Computational Linguistics -*, pp. 189–196, 1995. [Online]. Available: http://portal.acm.org/citation.cfm?doid=981658.981684

[11] S. Abney, "Understanding the yarowsky algorithm," *Computational Linguistics*, vol. 30, no. 3, pp. 365–395, 2004. [Online]. Available: http://www.mitpressjournals.org/doi/pdf/10.1162/0891201041850876

[12] A. Dempster, N. Laird, and D. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society. Series B*, vol. 39, no. 1, pp. 1–38, 1977. [Online]. Available: http://www.jstor.org/stable/10.2307/2984875

[13] X. Zhu, Z. Ghahramani, and J. Lafferty, "Semi-supervised learning using gaussian fields and harmonic functions," in *Proceedings of the 20th International Conference on Machine Learning*, 2003, pp. 912–919. [Online]. Available: http://www.aaai.org/Papers/ICML/2003/ICML03-118.pdf

[14] K. P. Bennett and A. Demiriz, "Semi-supervised support vector machines," in *Advances in Neural Information Processing Systems 11*, 1998.

[15] T. Joachims, "Transductive inference for text classification using support vector machines," in *Proceedings of the 16th International Conference on Machine Learning*. Morgan Kaufmann Publishers, 1999, pp. 200–209.

[16] Y. Grandvalet and Y. Bengio, "Semi-supervised learning by entropy minimization," in *Advances in Neural Information Processing Systems 17*, L. K. Saul, Y. Weiss, and L. Bottou, Eds. Cambridge, MA: MIT Press, 2005. [Online]. Available: http://eprints.pascal-network.org/archive/00001978/

[17] B. Fan, Z. Lei, and S. Z. Li, "Normalized LDA for semi-supervised learning," in *8th IEEE International Conference on Automatic Face & Gesture Recognition*. Ieee, Sep. 2008, pp. 1–6. [Online]. Available: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=4813329

[18] D. Cai, X. He, and J. Han, "Semi-supervised Discriminant Analysis," *IEEE 11th International Conference on Computer Vision*, pp. 1–7, 2007. [Online]. Available: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=4408856

[19] M. Loog and A. C. Jensen, "Constrained log-likelihood-based semi-supervised linear discriminant analysis," in *Structural, Syntactic, and Statistical Pattern Recognition*, 2012. [Online]. Available: http://www.springerlink.com/index/U16X1L3015777162.pdf

[20] K. Bache and M. Lichman, "{UCI} Machine Learning Repository," 2013. [Online]. Available: http://archive.ics.uci.edu/ml

[21] P. L. Bartlett, M. I. Jordan, and J. D. McAuliffe, "Convexity, Classification, and Risk Bounds," *Journal of the American Statistical Association*, vol. 101, no. 473, pp. 138–156, Mar. 2006. [Online]. Available: http://www.tandfonline.com/doi/abs/10.1198/016214505000000907