The Asymptotics of Semi-Supervised Learning in Discriminative Probabilistic Models

Nataliya Sokolovska SOKOLOVSKA@TELECOM-PARISTECH.FR
Olivier Cappé CAPPE@TELECOM-PARISTECH.FR
LTCI, TELECOM ParisTech and CNRS, 46 rue Barrault, 75013 Paris, France

François Yvon YVON@LIMSI.FR
Université Paris-Sud 11 and LIMSI-CNRS, 91403 Orsay, France

Abstract

Semi-supervised learning aims at taking advantage of unlabeled data to improve the efficiency of supervised learning procedures. For discriminative models however, this is a challenging task. In this contribution, we introduce an original methodology for using unlabeled data through the design of a simple semi-supervised objective function. We prove that the corresponding semi-supervised estimator is asymptotically optimal. The practical consequences of this result are discussed for the case of the logistic regression model.

1. Introduction

In most real-world pattern classification problems (e.g., for text, image or audio data), unannotated data is plentiful and can be collected at almost no cost, whereas labeled data are comparatively rarer, and more costly to gather. A sensible question is thus to find ways to exploit the unlabeled data in order to improve the performance of supervised training procedures. Many proposals have been made in the recent years to devise effective semi-supervised training schemes (see (Chapelle et al., 2006) for an up-to-date panorama). In this contribution, we focus on methods applicable to probabilistic classifiers, that is, classifiers designed to provide a probabilistic confidence measure associated with each decision. These classifiers do not necessarily perform better than other alternatives – particularly since probabilistic classification and minimum error classification are related, but different, tasks – but are important in some applications, for instance when it comes to predicting the generalization error, dealing with uneven error costs, ranking, combining decisions from multiple sources, etc.

Probabilistic generative models fare easily with the use of unlabeled data, usually through Expectation-Maximization (see, e.g., (Nigam et al., 2000; Klein & Manning, 2004) for successful implementations of this idea). It is however an extensively documented fact that discriminative models perform better than Generative models for classification tasks (Ng & Jordan, 2002). Integrating unlabeled data into discriminative models is however a much more challenging issue. Put in probabilistic terms, when learning to predict an output y from an observation x, a discriminative model attempts to fit $P(y|x;\theta)$, where θ denotes the parameter. The role to be played by any available prior knowledge about the marginal probability $P(x)$ in this context is not obvious. Several authors indeed claim that knowledge of $P(x)$ is basically useless (Seeger, 2002; Lasserre et al., 2006), although one of the contribution of this paper will be to show that this intuition relies on the implicit assumption that the model is *well-specified*, in the sense of allowing a perfect fit of the conditional probability.

The most common approach is to make the unknown parameter vector θ depend on the unlabeled data, either directly or indirectly. One way to achieve this goal is to use the unlabeled data to enforce constraints on the shape of $P(y|x)$: the *cluster assumption*, for instance, stipulates that the decision boundary should be located in low density regions (Seeger, 2002; Chapelle & Zien, 2005). (Grandvalet & Bengio, 2004) use this intuition to devise a semi-supervised training method (termed *entropy regularization*), which combines the usual log-likelihood term with an entropy-based penalty; see also (Jiao et al., 2006), who extend this methodology to Conditional Random Fields, (Laf-

Appearing in *Proceedings of the 25th International Conference on Machine Learning*, Helsinki, Finland, 2008. Copyright 2008 by the author(s)/owner(s).

ferty et al., 2001), or (Corduneanu & Jaakkola, 2003) for related ideas. This approach, as any attempt to distort the supervised training criterion with supplementary terms faces two risks: (i) to turn a well-behaved convex optimization problem into a non-convex one, fraught with local optima, thus making the results highly dependent of a proper initialization; (ii) to loose the asymptotic consistency property of the usual (conditional maximum likelihood) estimator. As a result, these methods are not guaranteed to improve over a trivial baseline which would only use the available annotated data. They furthermore require a fine tuning of the various optimization parameters (Mann & McCallum, 2007). The cluster assumption is also used in graph-based methods, which exploit the intuition that unlabeled data points should receive the same label as their labeled neighbors: in (Zhu & Ghahramani, 2002), a neighborhood graph is used to iteratively propagate labels from labeled to unlabeled data points until convergence.

(Lasserre et al., 2006) explores yet another avenue, introducing two sets of parameters: one for the conditional $P(y|x;\theta)$, and one for the marginal $P(x;\nu)$: the case where θ and ν are unrelated corresponds to the purely discriminative model, where unlabeled data are of no help; taking $\theta = \nu$ recovers the traditional generative model; introducing (via their Bayesian prior distribution) dependencies between (θ, ν) allows to build a full range of hybrid models. Finally, we also mention (Mann & McCallum, 2007) who try to also exploit prior knowledge on the distribution of the labels Y, which may be available in some specific applications.

In this paper, we try to challenge the view that unlabeled data cannot help purely discriminative models by exhibiting a semi-supervised estimator of the parameter θ which is asymptotically optimal and, in some situations, preferable to the usual maximum (conditional) likelihood estimator. To this aim, we make the simplifying assumption that the marginal $P(x)$ is fully known, which is true in the limit of infinitely many unlabeled data. An interesting observation about the proposed method is that it is most efficient when the Bayes error is very small which correlates well with the intuition underlying most semi-supervised approaches that unlabeled data is most useful if one can assume that the classes are "well-separated". In addition to the asymptotic results, we also discuss a number of empirical findings pertaining to logistic regression.

This paper is organized as follows: in Section 2, we introduce our formal framework and formulate the main result of the paper (Theorem 1), which is first exposed in its full generality, then particularized to the case of the logistic regression. Experiments with the logistic regression model are discussed in Section 3. Concluding remarks and perspectives close the paper.

2. Semi-Supervised Estimator

Let $g(y|x;\theta)$ denote the conditional probability density function (pdf) corresponding to a discriminative probabilistic model parameterized by $\theta \in \Theta$. In the following, we will always assume that the class variable Y takes its values in a finite set, $\mathcal{Y}$, with a special interest for the binary case where $\mathcal{Y} = \{0, 1\}$. We will further assume that the input (or explanatory) variable X also takes its values in a finite set $\mathcal{X}$, which may be arbitrary large.

The training procedure has access to a set of n i.i.d. *labeled* observations, $(X_i, Y_i)_{1 \le i \le n}$, as well as to a potentially unlimited number of unlabeled observations, where the quantity of unlabeled data is so large that we can consider that the marginal probability of X is fully known.

Finally, for a function $f : \mathbb{R}^p \mapsto \mathbb{R}$, we denote by $\nabla_z f(z_\star)$ the $p \times 1$ gradient vector and by $\nabla_{z^\mathrm{T}} \nabla_z f(z_\star)$ the $p \times p$ Hessian matrix in $z_\star$. When $f : \mathbb{R}^p \mapsto \mathbb{R}^r$, the notation $\nabla_{z^\mathrm{T}} f(z_\star)$ will be used to denote the $r \times p$ Jacobian matrix in $z_\star$.

2.1. Preliminary: A Simple Case

We first consider the case where the "model" of interest is very basic and simply consists in estimating the complete joint probability of X and Y, which is denoted by $\pi(x, y)$. We will also denote by $\eta(y|x)$ and $q(x)$, respectively, the conditional and the marginal probabilities associated with π. Although this case is not directly of interest for real-life statistical learning tasks, it highlights the role played by the knowledge of the marginal q in semi-supervised learning.

It is well known that the maximum-likelihood estimator of $\pi(x, y)$ defined by

$$\hat{\pi}_n(x, y) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}\{X_i = x, Y_i = y\} \tag{1}$$

is asymptotically efficient with asymptotic variance $v(x, y) = \pi(x, y)(1 - \pi(x, y))$ (assuming that $0 < \pi(x, y) < 1$).

Assume now that we are given $q(x)$, the marginal distribution of X, and that $0 < q(x) < 1$. It is easily checked that the maximum-likelihood estimator of $\pi(x, y)$ subject to the marginal constraint that

$\sum_{y\in\mathcal{Y}} \pi(x,y) = q(x)$ is given by

$$\hat{\pi}^s_n(x,y) = \frac{\sum_{i=1}^n \mathbf{1}\{X_i = x, Y_i = y\}}{\sum_{i=1}^n \mathbf{1}\{X_i = x\}} q(x) \qquad (2)$$

where the superscript s stands for "semi-supervised" and the ratio is recognized as the maximum-likelihood estimate of the *conditional probability* $\eta(y|x)$. As $\hat{\pi}^s_n(x,y)$ is a ratio of two simple estimators, its asymptotic variance can be computed using the δ-method, yielding

$$v^s(x,y) = \pi(x,y)(1 - \pi(x,y)/q(x))$$

As $0 < \pi(x,y) \leq q(x) < 1$, $v^s(x,y)$ is less than $v(x,y)$. Hence, in general the semi-supervised estimator $\hat{\pi}^s_n(x,y)$ and $\hat{\pi}_n(x,y)$ are not asymptotically equivalent, and $\hat{\pi}^s_n(x,y)$ is preferable. More precisely, $v^s(x,y)/v(x,y) = (1 - \pi(x,y)/q(x))/(1 - \pi(x,y))$ which tends to zero as $\pi(x,y)$ gets closer to $q(x)$. In other words, the performance of $\hat{\pi}^s_n(x,y)$ is all the more appreciable, compared to that of $\hat{\pi}_n(x,y)$, that y is a frequent label for x. In this case, knowledge of the marginal $q(x)$ makes it possible to obtain a precise estimate of $\hat{\pi}^s_n(x,y) \approx q(x)$ even with a very limited number of observations of x.

2.2. General Discriminative Model

We now consider the extension of the previous simple observation to the case of a general discriminative probabilistic model; the main difference being the fact that a given parametric model $\{g(y|x;\theta)\}_{\theta\in\Theta}$ will generally not be able to fit exactly the actual conditional distribution $\eta(y|x)$ of the data. As in the fully-specified case above, it is nonetheless possible to exhibit a semi-supervised estimator which is asymptotically optimal and preferable to the usual conditional maximum likelihood estimator defined by

$$\hat{\theta}_n = \arg\min_{\theta\in\Theta} \frac{1}{n}\sum_{i=1}^n \ell(Y_i|X_i;\theta) \qquad (3)$$

where $\ell(y|x;\theta) = -\log g(y|x;\theta)$ denotes the inverse of the conditional log-likelihood function.

Under the (classical) assumptions of Theorem 1 below, $\frac{1}{n}\sum_{i=1}^n \ell(Y_i|X_i;\theta)$ tends, uniformly in θ, to $\mathrm{E}_\pi[\ell(Y|X;\theta)]$ and thus the limiting value of $\hat{\theta}_n$ is given by

$$\theta_\star = \arg\min_{\theta\in\Theta} \mathrm{E}_\pi[\ell(Y|X;\theta)] \qquad (4)$$

The maximum likelihood estimator in (3) may also be interpreted as $\hat{\theta}_n = \arg\min_{\theta\in\Theta} \mathrm{E}_{\hat{\pi}_n}[\ell(Y|X;\theta)]$ where

$$\hat{\pi}_n(x,y) = \frac{1}{n}\sum_{i=1}^n \mathbf{1}\{X_i = x, Y_i = y\}$$

denotes the empirical measure associated with the sample $(X_i, Y_i)_{1\leq i\leq n}$, which also coincides with the maximum likelihood estimate of $\pi(x,y)$ defined in (1).

If we now assume that the marginal $q(x)$ is available, we know that $\hat{\pi}_n(x,y)$ is dominated (asymptotically) by the estimator $\hat{\pi}^s_n(x,y)$ defined in (2), which we here particularize to

$$\hat{\pi}^s_n(x,y) = \begin{cases} \frac{\sum_{i=1}^n \mathbf{1}\{X_i=x,Y_i=y\}}{\sum_{i=1}^n \mathbf{1}\{X_i=x\}} q(x) & \text{if } \sum_{i=1}^n \mathbf{1}\{X_i = x\} > 0 \\ 0 & \text{otherwise} \end{cases} \qquad (5)$$

By analogy with the construction used in the absence of information on q, we now define the corresponding semi-supervised estimator as $\hat{\theta}^s_n = \arg\min_{\theta\in\Theta} \mathrm{E}_{\hat{\pi}^s_n}[\ell(Y|X;\theta)]$, where the notation $\mathrm{E}_{\hat{\pi}^s_n}[f(Y,x)] = \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} \hat{\pi}^s_n(x,y) f(x,y)$ is used somewhat loosely here as it may happen that, for finite n, $\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} \hat{\pi}_n(x,y) < 1$, although $\hat{\pi}_n(x,y)$ sums to one with probability one, for sufficiently large n. It is easily checked that $\hat{\theta}^s_n$ may also be rewritten as

$$\hat{\theta}^s_n = \arg\min_{\theta\in\Theta} \sum_{i=1}^n \frac{q(X_i)}{\sum_{j=1}^n \mathbf{1}\{X_j = X_i\}} \ell(Y_i|X_i;\theta) \qquad (6)$$

Eq. (6) is a weighted version of (3) where the weight given to observations that share the same input x is common and reflects our prior knowledge on the marginal $q(x)$.

Theorem 1 *Let the joint probability of X and Y factorize as $\pi(x,y) = \eta(y|x)q(x)$, where q is known, and define the following matrices*

$$H(\theta_\star) = \mathrm{E}_q\left(\mathrm{V}_\eta\left[\nabla_\theta \ell(Y|X;\theta_\star)|X\right]\right) \qquad (7)$$

$$I(\theta_\star) = \mathrm{E}_q\left[\nabla_\theta \ell(Y|X;\theta_\star)\left\{\nabla_\theta \ell(Y|X;\theta_\star)\right\}^{\mathrm{T}}\right] \qquad (8)$$

$$J(\theta_\star) = \mathrm{E}_q\left[\nabla_{\theta^{\mathrm{T}}}\nabla_\theta \ell(Y|X;\theta_\star)\right] \qquad (9)$$

Assume that (1) $\mathcal{X}$ and $\mathcal{Y}$ are finite sets; (2) $\pi(x,y) > 0$ for all $(x,y) \in \mathcal{X}\times\mathcal{Y}$; (3) for all $(x,y) \in \mathcal{X}\times\mathcal{Y}$, $\ell(y|x;\theta)$ is bounded on Θ; (4) $\theta_\star$ is the unique minimizer of $\mathrm{E}_\pi[\ell(Y|X;\theta)]$ on Θ; (5) for all $(x,y) \in \mathcal{X}\times\mathcal{Y}$, $\ell(y|x;\theta)$ is twice continuously differentiable on Θ; (6) the matrices $H(\theta_\star)$ and $J(\theta_\star)$ are non singular.

Then, $\hat{\theta}_n$ and $\hat{\theta}^s_n$ are consistent and asymptotically normal estimators of $\theta_\star$, which satisfy

$$\sqrt{n}\left(\hat{\theta}_n - \theta_\star\right) \xrightarrow{L} \mathcal{N}\left(0, J^{-1}(\theta_\star)I(\theta_\star)J^{-1}(\theta_\star)\right) \quad (10)$$

$$\sqrt{n}\left(\hat{\theta}^s_n - \theta_\star\right) \xrightarrow{L} \mathcal{N}\left(0, J^{-1}(\theta_\star)H(\theta_\star)J^{-1}(\theta_\star)\right) \quad (11)$$

Furthermore, $\hat{\theta}^s_n$ is asymptotically efficient.

Theorem 1 asserts that the asymptotic covariance matrix associated with $\hat{\theta}_n^s$ is optimal. Understanding the relations between $H(\theta_\star)$ and $I(\theta_\star)$ is thus important to assess the asymptotic performance achievable by *any* semi-supervised training method which assumes prior knowledge of $q(x)$. Indeed, the well-known Rao-Blackwell variance decomposition shows that

$$I(\theta_\star) - H(\theta_\star) = \mathrm{V}_q\left(\mathrm{E}_\eta\left[\nabla_\theta \ell(Y|X;\theta_\star)|X\right]\right)$$

As a result, the difference between both estimators will mostly depend on whether $\mathrm{E}_\eta\left[\nabla_\theta \ell(Y|X;\theta_\star)|X=x\right]$ varies significantly or not around 0 as a function of x, given that, by definition, $\theta_\star$ is such that $\mathrm{E}_q\left(\mathrm{E}_\eta\left[\nabla_\theta \ell(Y|X;\theta_\star)|X\right]\right) = 0$.

Note that in the particular case where the model is *well-specified*, in the sense that $\theta_\star$ is such that $g(y|x;\theta_\star) = \eta(y|x)$ for all $(x,y) \in \mathcal{X} \times \mathcal{Y}$, not only is $\mathrm{E}_q\left(\mathrm{E}_\eta\left[\nabla_\theta \ell(Y|X;\theta_\star)|X\right]\right)$ null but one indeed has the stronger result that *for all* $x \in \mathcal{X}$, $\mathrm{E}_\eta\left[\nabla_\theta \ell(Y|X;\theta_\star)|X=x\right] = 0$. This is the only case for which $H(\theta_\star) = I(\theta_\star)$, and hence, where both estimators are asymptotically equivalent; it is also well known that in this case $J(\theta_\star) = I(\theta_\star)$ so that all asymptotic covariance matrices coincide with the usual expression of the inverse of the Fisher information matrix for θ. Theorem 1 gives formal support to the intuition that it is impossible to improve over the classic maximum likelihood estimator for large n's when the model is well-specified, even when the marginal q is known.

The results of Theorem 1 are stated in terms of parameter estimation which is usually not the primary interest for statistical learning tasks. Due to the non-differentiability of the 0–1 loss, it is not directly possible to derive results pertaining to the error probability from Theorem 1. One may however state the following result in terms of the *logarithmic risk*, in which the negated log-likelihood $\ell(y|x;\theta)$ is interpreted as a loss function.

Corollary 2 *In addition to the assumptions of Theorem 1, assume that $\ell(y|x;\theta)$ has bounded second derivative on Θ. Then, the logarithmic risk admits the following asymptotic equivalent:* $\mathrm{E}_{\pi^{\otimes n}}\{\mathrm{E}_\pi[\ell(Y|X;\hat{\theta}_n)]\} = \mathrm{E}_\pi[\ell(Y|X;\theta_\star)] + \frac{1}{2n}\operatorname{trace}\left\{I(\theta_\star)J^{-1}(\theta_\star)\right\} + o\left(\frac{1}{n}\right)$, *where $\mathrm{E}_{\pi^{\otimes n}}$ denotes the expectation with respect to the training data $(X_i,Y_i)_{1\le i\le n}$; for the semi-supervised estimator $\hat{\theta}_n^s$, the first order term is given by* $\frac{1}{2n}\operatorname{trace}\left\{H(\theta_\star)J^{-1}(\theta_\star)\right\}$.

As a final comment on Theorem 1, note that the form of the semi-supervised estimator in (6) shows that $\hat{\theta}_n^s$ will be consistent also in the presence of covariate shift (i.e., when the marginal distribution of the training sample differs from q), whereas the logistic regression estimates can only be consistent in this case if we assume that the model is well-specified (Shimodaira, 2000). In the presence of covariate shift however, the expressions of the asymptotic covariance matrices will be different.

2.3. Application to Logistic Regression

To gain further insight into the results summarized in Theorem 1, we consider the example of the logistic regression model with binary labels Y and input variables X in $\mathbb{R}^p$; the parameter θ is thus p-dimensional. In this model, the negative log-likelihood function is given by $\ell(y|x;\theta) = -y\theta^{\mathrm{T}}x + \log(1+\mathrm{e}^{\theta^{\mathrm{T}}x})$[1]. Thus, the estimation equation which implicitly defines the value of the optimal fit $\theta_\star$ as the value for which $\mathrm{E}_\pi\left[\nabla_\theta \ell(Y|X;\theta_\star)\right] = 0$ may be rewritten as

$$\mathrm{E}_q\left[X\left(g(1|X;\theta_\star) - \eta(1|X)\right)\right] = 0 \tag{12}$$

Similar direct calculations yield

$$H(\theta_\star) = \mathrm{E}_q\left[\eta(1|X)(1-\eta(1|X))XX^{\mathrm{T}}\right] \tag{13}$$

$$I(\theta_\star) = \mathrm{E}_q\big[\{\eta(1|X)(1-\eta(1|X)) + (\eta(1|X) - g(1|X;\theta_\star))^2\}XX^{\mathrm{T}}\big] \tag{14}$$

$$J(\theta_\star) = \mathrm{E}_q\left[g(1|X;\theta_\star)\{1-g(1|X;\theta_\star)\}XX^{\mathrm{T}}\right] \tag{15}$$

$J(\theta_\star)$ is the Fisher information matrix traditionally found in logistic regression. Interestingly, $H(\theta_\star)$ is recognized as the Fisher information matrix for $\theta_\star$ corresponding to the fully supervised logistic regression model in the well-specified case (i.e. assuming that $g(y|x;\theta_\star) = \eta(y|x)$), although we made no such assumption here.

For logistic regression, the difference

$$I(\theta_\star) - H(\theta_\star) = \mathrm{E}_q\left[\{\eta(1|X) - g(1|X;\theta_\star)\}^2 XX^{\mathrm{T}}\right]$$

is clearly a term that is all the more significant that the fit achievable by the model is poor. The second important factor that can lead to substantial differences between the asymptotic performances of $\hat{\theta}_n$ and $\hat{\theta}_n^s$ is revealed by the following observation: for a given distribution π, the largest (in a matrix sense) achievable value for $I(\theta_\star)$ is given by

$$I(\theta_\star) = \mathrm{E}_q\left[\max\{\eta(1|X), 1-\eta(1|X)\}XX^{\mathrm{T}}\right]$$

whereas $H(\theta_\star)$ in (13) may be rewritten as

$$H(\theta_\star) = \mathrm{E}_q\big[\max\{\eta(1|X), 1-\eta(1|X)\} \min\{\eta(1|X), 1-\eta(1|X)\}XX^{\mathrm{T}}\big]$$

[1] Or $\log(1+\mathrm{e}^{-\theta^{\mathrm{T}}yx})$ when the labels are coded as $\{-1,1\}$ rather than $\{0,1\}$.

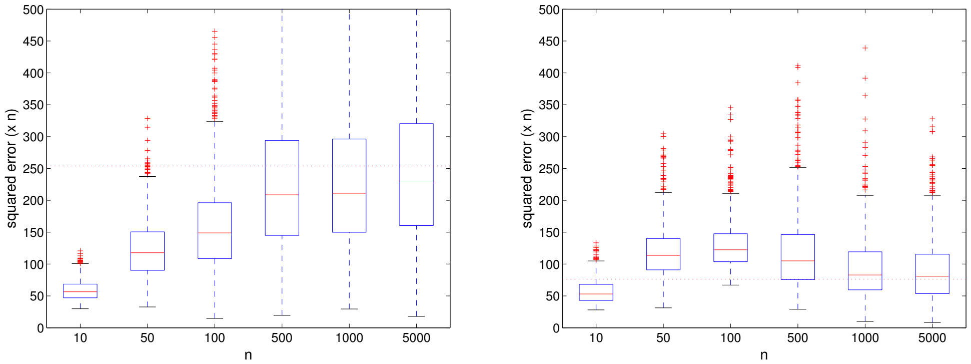

Figure 1. Boxplots of the scaled squared parameter estimation error as a function of the number of observations. Left: for logistic regression, $n\|\hat{\theta}_n - \theta_\star\|^2$; right: for the semi-supervised estimator, $n\|\hat{\theta}^s_n - \theta_\star\|^2$.

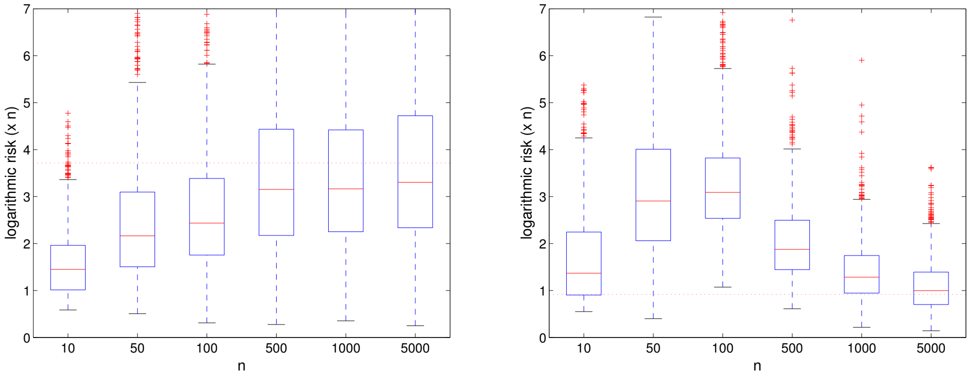

Figure 2. Boxplots of the scaled excess logarithmic risk as a function of the number of observations. Left: for logistic regression, $n(\mathrm{E}_{\pi\otimes n}\{\mathrm{E}_\pi[\ell(Y|X;\hat{\theta}_n)]\} - \mathrm{E}_\pi[\ell(Y|X;\theta_\star)])$; right: for the semi-supervised estimator, $n(\mathrm{E}_{\pi\otimes n}\{\mathrm{E}_\pi[\ell(Y|X;\hat{\theta}^s_n)]\} - \mathrm{E}_\pi[\ell(Y|X;\theta_\star)])$.

Hence the difference between $I(\theta_\star)$ and $H(\theta_\star)$ can only become very significant in cases where $\min\{\eta(1|X = x), 1 - \eta(1|X = x)\}$ is small, that is, when the probability of incorrect decision is small, for some values of x. The overall effect will be all the more significant that this situation happens for many values of x, or, in other words, that the Bayes error associated with π is small.

3. Experiments

3.1. A Small Scale Experiment

We consider here experiments on artificial data which correspond to the case of binary logistic regression discussed in Section 2.3. We focus on a small-scale problem where it is possible to exactly compute error probabilities and risks so as to completely bypass the empirical evaluation of trained classifiers. This setting makes it possible to obtain an accurate assessment of the performance as the only source of Monte Carlo error lies in the choice of the training corpus. More precisely, we consider the case where each observation consists of a vector of $p = 10$ positive counts which sums to $k = 3$. Hence the logistic regression parameter θ is ten-dimensional and the set $\mathcal{X}$ of possible count vectors contains exactly $\frac{(p+k-1)!}{(p-1)!k!} = 220$ different vectors.

In this case, it is well-known that one can simulate data from well-specified logistic models by resorting to mixture of multinomial distributions. Denote by α_1 the prior probability of class 1, and by β_0 and β_1 the vectors of multinomial parameters. Count vectors X generated from the mixture of multinomial have marginal probabilities $q(x) = \alpha_1 \operatorname{mult}(x; \beta_1) + (1 - \alpha_1) \operatorname{mult}(x; \beta_0)$ and conditional probabilities $\mathrm{P}(Y = 1|X = x) = \{1 + \exp -[(\log \beta_1 - \log \beta_0)^{\mathrm{T}} x + \log \frac{\alpha_1}{1-\alpha_1}]\}^{-1}$, where the log is to be understood componentwise. In the following, we take $\alpha_1 = 0.5$, i.e., balanced classes, so as to avoid the bias term.

In order to generate misspecified scenarios, we simply

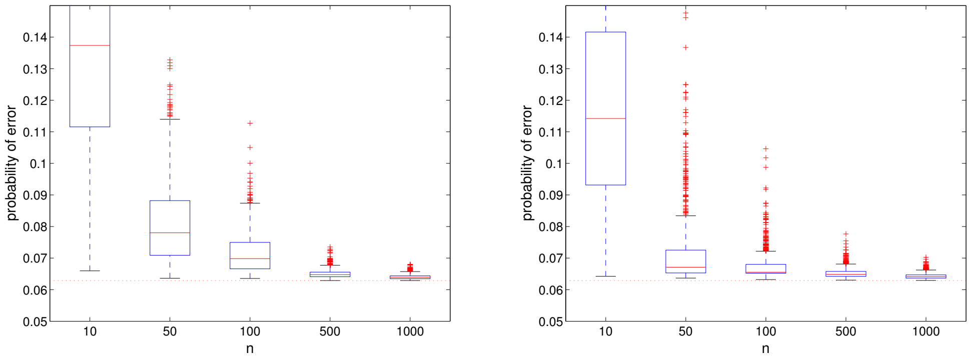

Figure 3. Boxplots of the probability of error as a function of the number of observations for a well-specified model. Left: for the logistic regression; right: for the semi-supervised estimator.

flipped the labels of a few (three in the case shown on figures. 1–2) x's taken among the most likely ones. This label flipping transformation leaves the Bayes error unchanged to that of the underlying unperturbed logistic model but the performance achievable by logistic regression is of course reduced. Figures 1 and 2 correspond to a case where the underlying unperturbed logistic model has a Bayes error of 1.7% and the probability of error associated with the best fitting logistic model is of 9.4%. Remember that in these figures, the only source of randomness is due to the choice of the training sample, which is repeated 1000 times independently for each size of the training sample, from $n = 10$ to $n = 5000$ observations.

As logistic regression is very sensitive to the use of regularization for small sample sizes (here, when n is less than one thousand), both (3) and (6) were regularized by adding a L^2 penalty term of the form $\rho_n\|\theta\|^2$, where ρ_n has been calibrated independently for each value of n. This being said, the optimal regularization parameter was always found to be within a factor 2 of $\rho_n = 1/n$ for (3) and $\rho_n = \frac{1}{n}\sum_{\{x:\sum_{i=1}^{1}\mathbf{1}\{X_i=x\}>0\}} q(x)$ for (6). The effect of regularization is also negligible for the two rightmost boxplots in each graph (i.e., when n is greater than 1000). On figures 1 and 2, the superimposed horizontal dashed lines correspond to the theoretical averages computed from Theorem 1 and Corollary 2, respectively.

When n is larger than one thousand, figures 1 and 2 perfectly correlate with the theory which predicts some advantage for the semi-supervised estimator as we are considering a case where the Bayes error is small and the model misspecification is significant. For large values of n, the semi-supervised estimator not only achieves better average performance but also does so more constantly, with a reduced variability. For smaller values of n, the picture is more contrasted, particularly when n ranges from 50 to 100 where the semi-supervised estimator may perform comparatively worse than the logistic regression. In this example, in terms of the probability of error, the semi-supervised estimator performs marginally better than logistic regression when $n = 10$ and $n = 5000$ (although the difference is bound to be very small in the latter case) and somewhat worse in between.

As expected, the difference between both approaches for large values of n decreases for scenarios with larger error probabilities. In those scenarios, the semi-supervised estimator performs worse than logistic regression for smaller values of n and equivalently for large values of n. A finding of interest is the fact that for well-specified models (i.e., with data generated from a multinomial mixture model) with low Bayes error, the semi-supervised approach does perform better than logistic regression, *for small values of* n. This effect can be significant even when considering the probability of error of the trained classifiers, as exemplified on Figure 3 in a case where the Bayes error is 6.3%. This observation is promising and deserves further investigation as the analysis of Section 2 only explains the behavior observed for large values of n, which in the case of well-specified models results in the two approaches being equivalent.

3.2. Text Classification Experiment

To evaluate our methodology on a more realistic test bed, we have used a simple binary classification task, consisting in classifying mails as spam or ham based on their textual content. The corpus used is the SpamAssassin corpus (Mason, 2002), which contains approximately 6 000 documents. Adapting our technique to real-world data requires to provide an estimate for the marginal $q(x)$. This was carried out by performing a discrete quantification of the data vectors as follows. We first use unsupervised clustering techniques to partition the available unlabeled collection of documents in k clusters. More specifically, we used a mixture of multinomial model as in (Nigam et al., 2000) with $k = 10$ components. We then simply adapt (6) by replacing $q(X_i)$ by the empirical frequency of the cluster to which X_i belongs, likewise the denominator $\sum_{j=1}^{n} \mathbf{1}\{X_j = X_i\}$ is replaced by the number *of training documents* belonging to the same cluster as X_i. We believe that this methodology is very general and makes the proposed approach applicable to a large variety of data. In effect, observations belonging to clusters which are underrepresented in the training corpus have higher relative weights, while the converse if true for observations belonging to overrepresented clusters. Note that, at this stage, no attempts have been made at tuning the number k of clusters, although intuition suggests that it would probably be reasonable to increase k (slowly) with n.

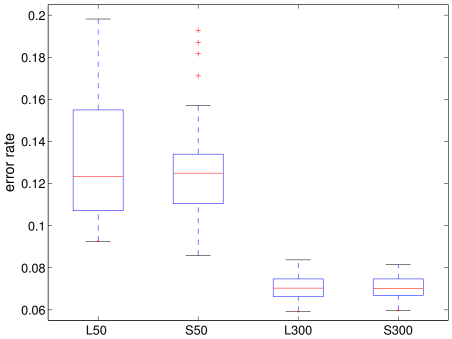

Figure 4. Boxplots of the error rates for, L50: logistic regression with $n = 50$; S50: semi-supervised estimator with $n = 50$; L300 and S300, idem with $n = 300$.

We tested the method with $n = 50$ and $n = 300$ randomly chosen training documents, the remaining mails serving as the test set; each trial gave rise to 50 Monte Carlo replications. For each value of n, the best regularization parameter was determined experimentally both for the usual logistic regression and the semi-supervised estimator. Each document is here represented as a count vector of dimension 1500. The resulting error rates are plotted as boxplots on Figure 4. Although the difference between both methods is certainly not very significant in this preliminary experiment, we note that, as in the simple case of Section 3.1, the semi-supervised estimator provides a more less variable performance when n is small.

4. Conclusion

In this contribution, we have tried to address the problem of semi-supervised learning without using any prior idea on what type of information is to be provided by the unlabeled data. The result of Theorem 1 provides both proper theoretical support for the claim that the unlabeled data does not matter asymptotically *when the model is well-specified* and a better understanding of the cases where the unlabeled data does matter. In particular, it confirms the intuition that unlabeled data is most useful when the Bayes error is small. One advantage of the proposed method is that it does not compromise the simplicity of the maximum likelihood approach because the weighted semi-supervised criterion stays convex. In addition, one could easily incorporate prior knowledge as used in other semi-supervised approaches: for instance the "cluster assumption" can be implemented by modifying (5) so as to incorporate a Bayesian prior that connects conditional probabilities for neighboring values of the input vector. In Section 3.2, we suggested a means by which the method can be extended to larger scales problem, including applications in which the feature vector is either continuous or has a more complex structure. We are in particular currently investigating the extension of the proposed approach to the case of sequence labelling with conditional random fields. Another open issue is the theoretical analysis of the behavior of the proposed criterion when n is small, which cannot be deduced from the asymptotic analysis presented here.

Appendix: Sketch of Proofs

First note that (10) is the well-known result that pertains to the behavior of the maximum likelihood estimator in misspecified models – see, for instance, (White, 1982) or Lemma 1 of (Shimodaira, 2000).

Now, the fact that $\hat{\theta}_n^s = \arg\min_{\theta\in\Theta} \mathrm{E}_{\hat{\pi}_n^s}[\ell(Y|X;\theta)]$ implicitly defines the semi-supervised estimator $\hat{\theta}_n^s$ as a function of the maximum-likelihood estimator of the conditional probabilities

$$\hat{\eta}_n(y|x) = \frac{\sum_{i=1}^{n} \mathbf{1}\{X_i = x, Y_i = y\}}{\sum_{i=1}^{n} \mathbf{1}\{X_i = x\}}$$

In our setting, the conditional probability η may be represented by a finite dimensional vector block defined by $\boldsymbol{\eta} = (\boldsymbol{\eta}(x_1), \ldots, \boldsymbol{\eta}(x_d))^{\mathrm{T}}$, where $\boldsymbol{\eta}(x_i) = (\eta(y_1|x_i), \ldots, \eta(y_k|x_i))^{\mathrm{T}}$, $\{x_1, \ldots, x_d\}$ denote the elements of $\mathcal{X}$, and, $\{y_0, \ldots, y_k\}$ denote the elements of $\mathcal{Y}$. As usual in polytomous regression models, we omit one of the possible values of Y (by convention, y_0) due to the constraint that $\sum_{y\in\mathcal{Y}} \eta(y|x) = 1$, for all $x \in \mathcal{X}$. The estimator $\hat{\boldsymbol{\eta}}_n$ is defined similarly with $\hat{\eta}_n(y|x)$ substituted for $\eta_n(y|x)$. $\hat{\boldsymbol{\eta}}_n$ is the maximum likelihood estimator of $\boldsymbol{\eta}$ and it is asymptotically efficient with asymptotic covariance matrix given by $K^{-1}(\boldsymbol{\eta})$, the inverse of the Fisher information matrix for $\boldsymbol{\eta}$, block-defined by

$$K^{-1}(\boldsymbol{\eta}) = \operatorname{diag}\left(K^{-1}(x_1;\boldsymbol{\eta}), \dots, K^{-1}(x_d;\boldsymbol{\eta})\right)$$

where

$$K^{-1}(x_i;\boldsymbol{\eta}) = q(x_i)^{-1}\left\{\operatorname{diag}\left(\boldsymbol{\eta}(x_i)\right) - \boldsymbol{\eta}(x_i)\boldsymbol{\eta}^{\mathrm{T}}(x_i)\right\}$$

To obtain the asymptotic behavior of the semi-supervised estimator $\hat{\theta}_n^s$, remark that $\hat{\theta}_n^s$ is obtained as a function ψ of $\hat{\boldsymbol{\eta}}_n$, where ψ is implicitly defined by the optimality equation $s(\boldsymbol{\eta}, \psi(\boldsymbol{\eta})) = 0$ where s is the (negative of the) score function defined by

$$s(\boldsymbol{\eta},\theta) = \nabla_\theta \mathrm{E}_\pi\left[\nabla_\theta \ell(Y|X;\theta)\right] = \sum_{x\in\mathcal{X}} q(x) \sum_{y\in\mathcal{Y}} \eta(y|x) \nabla_\theta \ell(y|x;\theta) \quad (16)$$

Because $\theta_\star = \psi(\boldsymbol{\eta})$ and $\hat{\theta}_n^s = \psi(\hat{\boldsymbol{\eta}}_n)$, $\hat{\theta}_n^s$ is an asymptotically efficient estimator of $\theta_\star$ with asymptotic covariance matrix given by $\nabla_{\boldsymbol{\eta}^{\mathrm{T}}}\psi(\boldsymbol{\eta}) K^{-1}(\boldsymbol{\eta}) \left\{\nabla_{\boldsymbol{\eta}^{\mathrm{T}}}\psi(\boldsymbol{\eta})\right\}^{\mathrm{T}}$. The Jacobian matrix $\nabla_{\boldsymbol{\eta}^{\mathrm{T}}}\psi(\boldsymbol{\eta})$ may be evaluated thanks to the implicit function theorem as

$$\nabla_{\boldsymbol{\eta}^{\mathrm{T}}}\psi(\boldsymbol{\eta}) = \left\{\nabla_{\theta^{\mathrm{T}}} s(\boldsymbol{\eta},\theta_\star)\right\}^{-1} \nabla_{\boldsymbol{\eta}^{\mathrm{T}}} s(\boldsymbol{\eta},\theta_\star)$$

From the definition of the score function in (16), it is obvious that $\nabla_{\theta^{\mathrm{T}}} s(\boldsymbol{\eta},\theta_\star) = J(\theta_\star)$. In order to calculate $\nabla_{\boldsymbol{\eta}^{\mathrm{T}}} s(\boldsymbol{\eta},\theta_\star)$, we differentiate the rightmost expression in (16) using the fact that $\eta(y_0|x) = 1 - \sum_{y\neq y_0}\eta(y|x)$ to obtain

$$\frac{\partial s(\boldsymbol{\eta},\theta)}{\partial \eta(x|y)} = q(x)\left[\nabla_\theta \ell(y|x;\theta) - \nabla_\theta \ell(y_0|x;\theta)\right]$$

The expression given in Theorem 1 or the asymptotic variance of $\hat{\theta}_n^s$ follows by computing the product $\nabla_{\boldsymbol{\eta}^{\mathrm{T}}} s(\boldsymbol{\eta},\theta_\star) K^{-1}(x_i;\boldsymbol{\eta}) \left\{\nabla_{\boldsymbol{\eta}^{\mathrm{T}}} s(\boldsymbol{\eta},\theta_\star)\right\}^{\mathrm{T}}$ – which factories into blocks of size k – and using the fact that $\eta(y_0|x) = 1 - \sum_{y\neq y_0}\eta(y|x)$.

Corollary 2 is based on the classical asymptotic expansion of $\mathrm{E}_\pi[\ell(Y|X;\hat{\theta}_n)] - \mathrm{E}_\pi[\ell(Y|X;\theta_\star)]$ as $\frac{1}{2}(\hat{\theta}_n - \theta_\star)^T J(\theta_\star)(\hat{\theta}_n - \theta_\star) + o_p(\frac{1}{n})$, see, for instance, (Bach, 2006).

References

Bach, F. (2006). Active learning for misspecified generalized linear models. *NIPS.*

Chapelle, O., Schölkopf, B., & Zien, A. (2006). *Semi-supervised learning.* MIT Press.

Chapelle, O., & Zien, A. (2005). Semi-supervised classification by low density separation. *In Proc. of the Tenth International Workshop on Artificial Intelligence and Statistics.*

Corduneanu, A., & Jaakkola, T. (2003). On information regularization. *In Proc. of the 19th conference on Uncertainty in Artificial Intelligence (UAI).*

Grandvalet, Y., & Bengio, Y. (2004). Semi-supervised learning by entropy minimization. *NIPS.*

Jiao, F., Wang, S., Lee, C. H., Greiner, R., & Schuurmans, D. (2006). Semi-supervised conditional random fields for improved sequence segmentation and labeling. *ACL/COLING.*

Klein, D., & Manning, C. D. (2004). Corpus-based induction of syntactic structure: models of dependency and constituency. *ACL.*

Lafferty, J., McCallum, A., & Pereira, F. (2001). Conditional random fields: probabilistic models for segmenting and labeling sequence data. *ICML.*

Lasserre, J. A., Bishop, C. M., & Minka, T. P. (2006). Principled hybrids of generative and discriminative models. *IEEE CVPR.*

Mann, G., & McCallum, A. (2007). Simple, robust, scalable semi-supervised learning via expectation regularization. *ICML.*

Mason, J. (2002). SpamAssassin corpus.

Ng, A., & Jordan, M. (2002). On discriminative vs. generative classifiers: A comparison of logistic regression and naive Bayes. *NIPS.*

Nigam, K., McCallum, A. K., Thrun, S., & Mitchell, T. (2000). Text classification from labeled and unlabeled documents using EM. *Machine Learning*, *39*, 103–134.

Seeger, M. (2002). *Learning with labeled and unlabeled data* (Technical Report). University of Edinburgh, Institute for Adaptive and Neural Computation.

Shimodaira, H. (2000). Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of Statistical Planning and Inference*, *90*, 227–244.

White, H. (1982). Maximum likelihood estimation in misspecified models. *Econometrica*, *50*, 1–25.

Zhu, X., & Ghahramani, Z. (2002). *Learning from labeled and unlabeled data with label propagation* (Technical Report). Carnegie Mellon University.