Journal of Machine Learning Research 11 (2010) 1323-1351 Submitted 6/09; Revised 11/09; Published 4/10

Unsupervised Supervised Learning I: Estimating Classification and Regression Errors without Labels

Pinar Donmez PINARD@CS.CMU.EDU
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213, USA

Guy Lebanon LEBANON@CC.GATECH.EDU
Krishnakumar Balasubramanian KRISHNAKUMAR3@GATECH.EDU
College of Computing
Georgia Institute of Technology
Atlanta, GA 30332, USA

Editor: Nicolò Cesa-Bianchi

Abstract

Estimating the error rates of classifiers or regression models is a fundamental task in machine learning which has thus far been studied exclusively using supervised learning techniques. We propose a novel unsupervised framework for estimating these error rates using only unlabeled data and mild assumptions. We prove consistency results for the framework and demonstrate its practical applicability on both synthetic and real world data.

Keywords: classification and regression, maximum likelihood, latent variable models

1. Introduction

A common task in machine learning is predicting a response variable $y \in \mathcal{Y}$ based on an explanatory variable $x \in \mathcal{X}$. Assuming a joint distribution $p(x,y)$ and a loss function $L(y,\hat{y})$, a predictor $f : \mathcal{X} \to \mathcal{Y}$ is characterized by an expected loss or risk function

$$R(f) = \mathsf{E}_{p(x,y)}\{L(y, f(x))\}.$$

For example, in classification we may have $\mathcal{X} = \mathbb{R}^d$, $\mathcal{Y} = \{1,\ldots,l\}$, and $L(y,\hat{y}) = I(y \neq \hat{y})$ where $I(A) = 1$ if A is true and 0 otherwise. The resulting risk is known as the 0-1 risk or simply the classification error rate

$$R(f) = P(f \text{ predicts the wrong class}).$$

In regression we may have $\mathcal{X} = \mathcal{Y} = \mathbb{R}$, and $L(y,\hat{y}) = (y-\hat{y})^2$. The resulting risk is the mean squared error

$$R(f) = \mathsf{E}_{p(x,y)}(y - f(x))^2.$$

We consider the case where we are provided with k predictors $f_i : \mathcal{X} \to \mathcal{Y}$, $i = 1,\ldots,k$ ($k \geq 1$) whose risks are unknown. The main task we are faced with is estimating the risks $R(f_1),\ldots,R(f_k)$ without using any labeled data whatsoever. The estimation of $R(f_i)$ is rather based on an estimator $\hat{R}(f_i)$ that uses unlabeled data $x^{(1)},\ldots,x^{(n)} \overset{\text{iid}}{\sim} p(x)$.

A secondary task that we consider is obtaining effective schemes for combining k predictors $f_1,\ldots,f_k$ in a completely unsupervised manner. We refer to these two tasks of risk estimation and predictor combination as unsupervised-supervised learning since they refer to unsupervised analysis of supervised prediction models.

It may seem surprising that unsupervised risk estimation is possible at all. After all in the absence of labels there is no ground truth that guides us in estimating the risks. However, as we show in this paper, if the marginal $p(y)$ is known it is possible in some cases to obtain a consistent estimator for the risks using only unlabeled data, that is,

$$\lim_{n\to\infty} \hat{R}(f_i;x^{(1)},\ldots,x^{(n)}) = R(f_i) \quad \text{with probability } 1, \quad i=1,\ldots,k.$$

In addition to demonstrating consistency, we explore the asymptotic variance of the risk estimators and how it is impacted by changes in n (amount of unlabeled data), k (number of predictors), and $R(f_1),\ldots,R(f_k)$ (risks). We also demonstrate that the proposed estimation technique works well in practice on both synthetic and real world data.

The assumption that $p(y)$ is known seems restrictive, but there are plenty of cases where it holds. Examples include medical diagnosis ($p(y)$ is the well known marginal disease frequency), handwriting recognition/OCR ($p(y)$ is the easily computable marginal frequencies of different English letters), regression model for life expectancy ($p(y)$ is the well known marginal life expectancy tables). In these and other examples $p(y)$ is obtained from extremely accurate histograms.

There are several reasons that motivate our approach of using exclusively unlabeled data to estimate the risks. Labeled data may be unavailable due to privacy considerations where the predictors are constructed by organizations using training sets with private labels. For example, in medical diagnosis prediction, the predictors $f_1,\ldots,f_k$ may be obtained by k different hospitals, each using a private internal labeled set. Following the training stage, each hospital releases its predictor to the public who then proceed to estimate $R(f_1),\ldots,R(f_k)$ using a separate unlabeled data set.

Another motivation for using unlabeled data is domain adaptation where predictors that are trained on one domain, are used to predict data from a new domain from which we have only unlabeled data. For example, predictors are often trained on labeled examples drawn from the past but are used at test time to predict data drawn from a new distribution associated with the present. Here the labeled data used to train the predictors will not provide an accurate estimate due to differences in the test and train distributions.

Another motivation is companies releasing predictors to clients as black boxes (without their training data) in order to protect their intellectual property. This is the situation in business analytics and consulting. In any case, it is remarkable that without labels we can still accurately estimate supervised risks.

The collaborative nature of this diagnosis is especially useful for multiple predictors as the predictor ensemble $\{f_1,\ldots,f_k\}$ diagnoses itself. However, our framework is not restricted to a large k and works even for a single predictor with $k=1$. It may further be extended to the case of active learning where classifiers are queried for specific data and the case of semi-supervised learning where a small amount of labeled data is augmented by massive unlabeled data.

We proceed in the next section to describe the general framework and some important special cases. In Section 3 we discuss extensions to the general framework and in Section 4-5 we discuss the theory underlying our estimation process. In Section 6 we discuss practical optimization algorithms. Section 7 contains an experimental study. We conclude with a discussion in Section 8.

2. Unsupervised Risk Estimation Framework

We adopt the framework presented in Section 1 with the added requirement that the predictors $f_1,\ldots,f_k$ are stochastic, that is, their prediction $\hat{y} = f_i(x)$ (conditioned on x) is a random variable. Such stochasticity occurs if the predictors are conditional models predicting values according to their estimated probability, that is, f_i models a conditional distribution q_i and predicts y' with probability $q_i(y'|x)$.

As mentioned previously our goal is to estimate the risk associated with classification or regression models $f_1,\ldots,f_k$ based on unlabeled data $x^{(1)},\ldots,x^{(n)} \overset{\text{iid}}{\sim} p(x)$. The testing marginal and conditional distributions $p(x), p(y|x)$ may differ from the distributions used at training time for the different predictors. In fact, each predictor may have been trained on a completely different training distribution, or may have been designed by hand with no training data whatsoever. We consider the predictors as black boxes and do not assume any knowledge of their modeling assumptions or training processes.

At the center of our framework is the idea to define a parameter vector $\theta \in \Theta$ which characterizes the risks $R(f_1),\ldots,R(f_k)$, that is, $R(f_j) = g_j(\theta)$ for some function $g_j : \Theta \to \mathbb{R}$, $j = 1,\ldots,k$. The parameter vector θ is estimated from data by connecting it to the probabilities

$$p_j(y'|y) \overset{\text{def}}{=} p(f_j \text{ predicts } y' |\text{ true label is } y).$$

More specifically, we use a plug-in estimate $\hat{R}(f_j) = g_j(\hat{\theta})$ where $\hat{\theta}$ maximizes the likelihood of the predictor outputs $\hat{y}_j^{(i)} = f_j(x^{(i)})$ with respect to the model $p_\theta(\hat{y}) = \int p_\theta(\hat{y}|y)p(y)\,dy$. The precise equations are:

$$\boxed{\begin{aligned}
\hat{R}(f_j;\hat{y}^{(1)},\ldots,\hat{y}^{(n)}) &= g_j(\hat{\theta}^{\text{mle}}(\hat{y}^{(1)},\ldots,\hat{y}^{(n)})) \quad \text{where} && (1)\\
\hat{y}^{(i)} &\overset{\text{def}}{=} (\hat{y}_1^{(i)},\ldots,\hat{y}_k^{(i)})\\
\hat{y}_j^{(i)} &\overset{\text{def}}{=} f_j(x^{(i)}),\\
\hat{\theta}^{\text{mle}}(\hat{y}^{(1)},\ldots,\hat{y}^{(n)}) &= \arg\max \ell(\theta;\hat{y}^{(1)},\ldots,\hat{y}^{(n)}), && (2)\\
\ell(\theta;\hat{y}^{(1)},\ldots,\hat{y}^{(n)}) &= \sum_{i=1}^{n} \log p_\theta(\hat{y}_1^{(i)},\ldots,\hat{y}_k^{(i)}) && (3)\\
&= \sum_{i=1}^{n} \log \int_{\mathcal{Y}} p_\theta(\hat{y}_1^{(i)},\ldots,\hat{y}_k^{(i)}|y^{(i)})p(y^{(i)})\,d\mu(y^{(i)}).
\end{aligned}}$$

The integral in (3) is over the unobserved label $y^{(i)}$ associated with $x^{(i)}$. It should be a continuous integral $\int_{y^{(i)}=-\infty}^{\infty}$ for regression and a finite summation $\sum_{y^{(i)}=1}^{l}$ for classification. For notational simplicity we maintain the integral sign for both cases with the understanding that it is over a continuous or discrete measure μ, depending on the topology of $\mathcal{Y}$. Note that (3) and its maximizer are computable without any labeled data. All that is required are the classifiers (as black boxes), unlabeled data $x^{(1)},\ldots,x^{(n)}$, and the marginal label distribution $p(y)$.

Besides being a diagnostic tool for the predictor accuracy, $\hat{\theta}^{\text{mle}}$ can be used to effectively aggregate $f_1,\dots,f_j$ to predict the label of a new example x^{new}

$$\begin{aligned}\hat{y}^{\text{new}} &= \arg\max_{y\in\mathcal{Y}} p_{\hat{\theta}^{\text{mle}}}(y \mid f_1(x^{\text{new}}),\dots,f_k(x^{\text{new}})) \\ &= \arg\max_{y\in\mathcal{Y}} p(y)\prod_{j=1}^{k} p_{\hat{\theta}_j^{\text{mle}}}(f_j(x^{\text{new}}) \mid y). \end{aligned} \tag{4}$$

As a result, our framework may be used to combine existing classifiers or regression models in a completely unsupervised manner.

There are three important research questions concerning the above framework. First, what are the statistical properties of $\hat{\theta}^{\text{mle}}$ and $\hat{R}$ (consistency, asymptotic variance). Second, how can we efficiently solve the maximization problem (2). And third, how does the framework work in practice. We address these three questions in Sections 4-5, 6, 7 respectively, We devote the rest of the current section to examine some important special cases of (2)-(3) and consider some generalizations in the next section.

2.1 Non-Collaborative Estimation of the Risks

In the non-collaborative case we estimate the risk of each one of the predictors $f_1,\dots,f_k$ separately. This reduces the problem to that of estimating the risk of a single predictor, which is repeated k times for each one of the predictors. We thus assume in this subsection the framework (1)-(3) with $k=1$ with no loss of generality. For simplicity we denote the single predictor by f rather than f_1 and denote $g=g_1$ and $\hat{y}^{(i)}=\hat{y}_1^{(i)}$. The corresponding simplified expressions are

$$\begin{aligned}\hat{R}(f;\hat{y}^{(1)},\dots,\hat{y}^{(n)}) &= g(\hat{\theta}^{\text{mle}}(\hat{y}^{(1)},\dots,\hat{y}^{(n)})), \\ \hat{\theta}^{\text{mle}}(\hat{y}^{(1)},\dots,\hat{y}^{(n)}) &= \arg\max_{\theta}\sum_{i=1}^{n}\log\int_{\mathcal{Y}} p_\theta(\hat{y}^{(i)}|y^{(i)})p(y^{(i)})\,d\mu(y^{(i)}) \end{aligned} \tag{5}$$

where $\hat{y}^{(i)}=f(x^{(i)})$.

We consider below several important special cases.

2.1.1 CLASSIFICATION

Assuming l labels $\mathcal{Y}=\{1,\dots,l\}$, the classifier f defines a multivariate Bernoulli distribution $p_\theta(\hat{y}|y)$ mapping the true label y to $\hat{y}$

$$p_\theta(\hat{y}|y)=\theta_{\hat{y},y}. \tag{6}$$

where θ is the stochastic confusion matrix or noise model corresponding to the classifier f. In this case, the relationship between the risk $R(f)$ and the parameter θ is

$$R(f)=1-\sum_{y\in\mathcal{Y}}\theta_{yy}\,p(y). \tag{7}$$

Equations (6)-(7) may be simplified by assuming a symmetric error distribution (Cover and Thomas, 2005)

$$\begin{aligned} p_\theta(\hat{y}|y) &= \theta^{I(\hat{y}=y)} \left(\frac{1-\theta}{l-1}\right)^{I(\hat{y}\neq y)}, \\ R(f) &= 1-\theta \end{aligned} \tag{8}$$

where I is the indicator function and $\theta \in [0,1]$ is a scalar corresponding to the classifier accuracy. Estimating θ by maximizing (5), with (6) or (8) substituting p_θ completes the risk estimation task.

In the simple binary case $l = 2, \mathcal{Y} = \{1,2\}$ with the symmetric noise model (8) the loglikelihood

$$\ell(\theta) = \sum_{i=1}^{n} \log \sum_{y^{(i)}=1}^{2} \theta^{I(\hat{y}^{(i)}=y^{(i)})}(1-\theta)^{I(\hat{y}^{(i)}\neq y^{(i)})} p(y^{(i)})$$

may be shown to have the following closed form maximizer

$$\hat{\theta}^{\text{mle}} = \frac{p(y=1) - m/n}{2p(y=1)-1}. \tag{9}$$

where $m \stackrel{\text{def}}{=} |\{i \in \{1,\ldots,n\} : \hat{y}^{(i)} = 2\}|$. The estimator (9) works well in practice and is shown to be a consistent estimator in the next section (i.e., it converges to the true parameter value). In cases where the symmetric noise model (8) does not hold, using (9) to estimate the classification risk may be misleading. For example, in some cases (9) may be negative. In these cases, using the more general model (6) instead of (8) should provide more accurate results. We discuss this further from theoretical and experimental perspectives in Sections 4-5, and 7 respectively.

2.1.2 REGRESSION

Assuming a regression equation

$$y = ax + \varepsilon, \qquad \varepsilon \sim N(0, \tau^2)$$

and an estimated regression model or predictor $\hat{y} = a'x$ we have

$$\hat{y} = a'x = a'a^{-1}(y-\varepsilon) = \theta y - \theta\varepsilon$$

where $\theta = a'a^{-1}$. Thus, in the regression case the distribution $p_\theta(\hat{y}|y)$ and the relationship between the risk and the parameter $R(f) = g(\theta)$ are

$$\begin{aligned} p_\theta(\hat{y}|y) &= (2\pi\theta^2\tau^2)^{-1/2} \exp\left(-\frac{(\hat{y}-\theta y)^2}{2\theta^2\tau^2}\right), \\ R(f|y) &= \mathsf{bias}^2(f) + \mathsf{Var}(f) = (1-\theta)^2 y^2 + \theta^2\tau^2, \\ R(f) &= \theta^2\tau^2 + (1-\theta)^2 \mathsf{E}_{p(y)}(y^2). \end{aligned} \tag{10}$$

Note that we consider regression as a stochastic estimator in that it predicts $y = a'x + \varepsilon$ or $y|x \sim N(a'x, \tau^2)$.

Assuming $p(y) = N(\mu_y, \sigma_y^2)$ (as is often done in regression analysis) we have

$$\begin{aligned} p_\theta(\hat{y}^{(i)}) &= \int_{\mathbb{R}} p_\theta(\hat{y}^{(i)}|y)p(y)dy = (2\pi\theta^2\tau^2 2\pi\sigma_y^2)^{-1/2}\int_{\mathbb{R}} \exp\left(-\frac{(\hat{y}-\theta y)^2}{2\theta^2\tau^2} - \frac{(y-\mu_y)^2}{2\sigma_y^2}\right)dy \qquad (11)\\ &= \frac{1}{\theta\sqrt{2\pi(\tau^2+\sigma_y^2)}} \exp\left(\frac{(\hat{y}^{(i)})^2}{2\theta^2\tau^2}\left(\frac{\sigma_y^2}{\sigma_y^2+\tau^2}-1\right)+\frac{\mu_y^2}{2\sigma_y^2}\left(\frac{\tau^2}{\sigma_y^2+\tau^2}-1\right)+\frac{\hat{y}^{(i)}\mu_y}{\theta\left(\tau^2+\sigma_y^2\right)}\right) \end{aligned}$$

where we used the following lemma in the last equation.

Lemma 1 (e.g., Papoulis, 1984)

$$\int_{-\infty}^{\infty} Ae^{-Bx^2+Cx+D}dx = A\sqrt{\frac{\pi}{B}}\exp\left(C^2/4B+D\right)$$

where A, B, C, D are constants that do not depend on x.

In this case the loglikelihood simplifies to

$$\ell(\theta) = -n\log\left(\theta\sqrt{2\pi(\tau^2+\sigma_y^2)}\right) - \left(\frac{\sum_{i=1}^n(\hat{y}^{(i)})^2}{2(\tau^2+\sigma_y^2)}\right)\frac{1}{\theta^2} + \left(\frac{\mu_y\sum_{i=1}^n\hat{y}^{(i)}}{\tau^2+\sigma_y^2}\right)\frac{1}{\theta} - n\frac{\mu_y^2}{2(\sigma_y^2+\tau^2)}$$

which can be shown to have the following closed form maximizer

$$\hat{\theta}^{\text{mle}} = -\frac{\mu_y\sum_{i=1}^n\hat{y}^{(i)}}{2n(\tau^2+\sigma_y^2)} \pm \sqrt{\frac{\left(\mu_y\sum_{i=1}^n\hat{y}^{(i)}\right)^2}{4n^2(\tau^2+\sigma_y^2)^2} + \frac{\sum_{i=1}^n(\hat{y}^{(i)})^2}{n(\tau^2+\sigma_y^2)}}$$

where the two roots correspond to the two cases where $\theta = a'/a > 0$ and $\theta = a'/a < 0$.

The univariate regression case described above may be extended to multiple explanatory variables, that is, $y = Ax + \varepsilon$ where y, x, ε are vectors and A is a matrix. This is an interesting extension which falls beyond the scope of the current paper.

2.1.3 NOISY GAUSSIAN CHANNEL

In this case our predictor f corresponds to a noisy channel mapping a real valued signal y to its noisy version $\hat{y}$. The aim is to estimate the mean squared error or noise level $R(f) = \mathsf{E}\,\|y-\hat{y}\|^2$. In this case the distribution $p_\theta(\hat{y}|y)$ and the relationship between the risk and the parameter $R(f) = g(\theta)$ are

$$\begin{aligned} p_\theta(\hat{y}|y) &= (2\pi\theta^2)^{-1/2}\exp\left(-\frac{(\hat{y}-y)^2}{2\theta^2}\right),\\ R(f|y) &= \theta^2,\\ R(f) &= \theta^2\mathsf{E}_{p(y)}(y). \end{aligned}$$

The loglikelihood and other details in this case are straightforward variations on the linear regression case described above. We therefore concentrate in this paper on the classification and linear regression cases.

As mentioned above, in both classification and regression, estimating the risks for $k \geq 2$ predictors rather than a single one may proceed by repeating the optimization process described above for each predictor separately. That is $\hat{R}(f_j) = g_j(\hat{\theta}_j^{\text{mle}})$ where $\hat{\theta}_1^{\text{mle}}, \ldots, \hat{\theta}_k^{\text{mle}}$ are estimated by maximizing k different loglikelihood functions. In some cases the convergence rate to the true risks can be accelerated by jointly estimating the risks $R(f_1), \ldots, R(f_k)$ in a collaborative fashion. Such collaborative estimation is possible under some assumptions on the statistical dependency between the noise processes defining the k predictors. We describe below such an assumption followed by a description of more general cases.

2.2 Collaborative Estimation of the Risks: Conditionally Independent Predictors

We have previously seen how to estimate the risks of k predictors by separately applying (1) to each predictor. If the predictors are known to be conditionally independent given the true label, that is, $p_\theta(\hat{y}_1, \ldots, \hat{y}_k | y) = \prod_j p_{\theta_j}(\hat{y}_j | y)$ the loglikelihood (3) simplifies to

$$\ell(\theta) = \sum_{i=1}^{n} \log \int_{\mathcal{Y}} \prod_{j=1}^{k} p_{\theta_j}(\hat{y}_j^{(i)} | y^{(i)}) p(y^{(i)}) \, d\mu(y^{(i)}), \quad \text{where} \quad \hat{y}_j^{(i)} = f_j(x^{(i)}) \qquad (12)$$

and p_{θ_j} above is (6) or (8) for classification and (10) for regression. Maximizing the loglikelihood (12) jointly over $\theta_1, \ldots, \theta_k$ results in estimators $\hat{R}(f_1), \ldots, \hat{R}(f_k)$ that converge to the true value faster than the non-collaborative MLE (5) (more on this in Section 7). Equation (12) does not have a closed form maximizer requiring the use of iterative computational techniques.

The conditional independence of the predictors is a much weaker condition than the independence of the predictors which is very unlikely to hold. In our case, each predictor f_j has its own stochastic noise operator $T_j(r,s) = p(\hat{y} = r | y = s)$ (regression) or matrix $[T_j]_{rs} = p_j(\hat{y} = r | y = s)$ (classification) where $T_1, \ldots, T_k$ may be arbitrarily specified. In particular, some predictors may be similar, for example, $T_i \approx T_j$, and some may be different, for example, $T_i \not\approx T_j$. The conditional independence assumption that we make in this subsection is that conditioned on the latent label y the predictions of the predictors proceed stochastically according to $T_1, \ldots, T_k$ in an independent manner.

Figure 1 displays the loglikelihood functions $\ell(\theta)$ for three different data set sizes $n = 100, 250, 500$. As the size n of the unlabeled data grows the curves become steeper and $\hat{\theta}_n^{\text{mle}}$ approach θ^{true}. Figure 2 displays a similar figure for $k = 1$ in the case of regression.

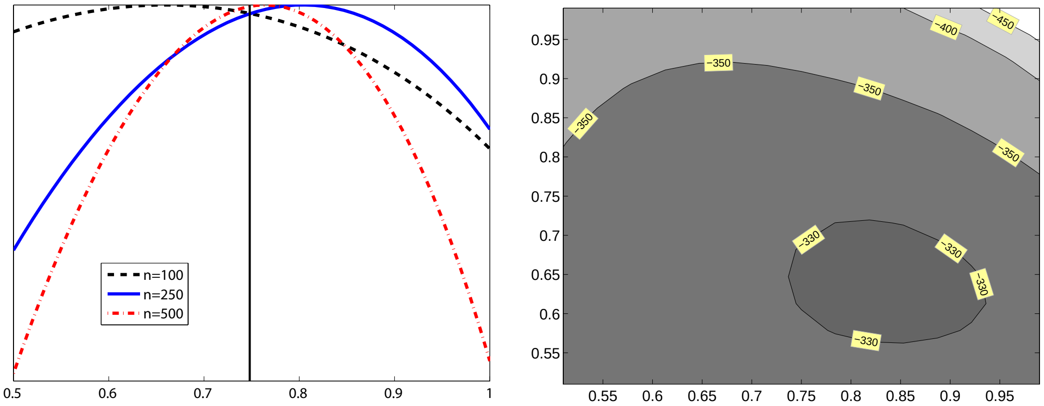


Figure 1: A plot of the loglikelihood functions $\ell(\theta)$ in the case of classification for $k=1$ (left, $\theta^{\text{true}}=0.75$) and $k=2$ (right, $\theta^{\text{true}}=(0.8,0.6)^\top$). The loglikelihood was constructed based on random samples of unlabeled data with sizes $n=100,250,500$ (left) and $n=250$ (right) and $p(y=1)=0.75$. In the left panel the y values of the curves were scaled so their maxima would be aligned. For $k=1$ the estimators $\hat{\theta}^{\text{mle}}$ (and their errors $|\hat{\theta}^{\text{mle}}-0.75|$) for $n=100,250,500$ are 0.6633 (0.0867), 0.8061 (0.0561), 0.765 (0.0153). As additional unlabeled examples are added the loglikelihood curves become steeper and their maximizers become more accurate and closer to θ^{true}.

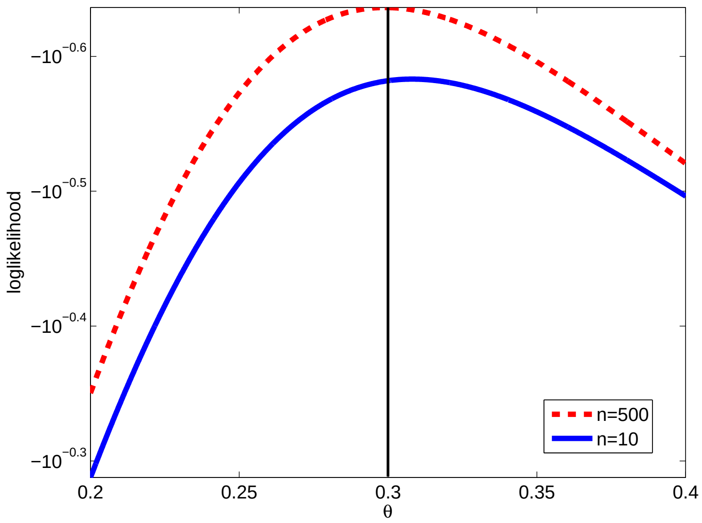


Figure 2: A plot of the loglikelihood function $\ell(\theta)$ in the case of regression for $k=1$ with $\theta^{\text{true}}=0.3$, $\tau=1$, $\mu_y=0$ and $\sigma_y=0.2$. As additional unlabeled examples are added the loglikelihood curve become steeper and their maximizers get closer to the true parameter θ^{true} resulting in a more accurate risk estimate.

In the case of regression (12) involves an integral over a product of $k+1$ Gaussians, assuming that $y \sim N(\mu_y, \sigma_y^2)$. In this case the integral in (12) simplifies to

$$\begin{aligned}
p_\theta(\hat{y}_1^{(i)}, \dots, \hat{y}_k^{(i)}) &= \int_{-\infty}^{\infty} \left(\prod_{j=1}^{k} \frac{1}{\theta_j \tau \sqrt{2\pi}} e^{-\left(\hat{y}_j^{(i)} - \theta_j y^{(i)}\right)^2 \big/ 2\theta_j^2 \tau^2} \right) \frac{1}{\sigma_y \sqrt{2\pi}} e^{-\left(y^{(i)} - \mu_y\right)^2 \big/ 2\sigma_y^2} \, dy^{(i)} \\
&= \frac{1}{\tau^k (\sqrt{2\pi})^{k+1} \sigma_y \prod_{j=1}^k \theta_j} \int_{-\infty}^{\infty} \exp\left[-\frac{1}{2} \left(\left(\frac{y^{(i)} - \mu_y}{\sigma_y} \right)^2 + \sum_{j=1}^{k} \left(\frac{y^{(i)}}{\tau} - \frac{\hat{y}_j^{(i)}}{\tau\theta_j} \right)^2 \right) \right] dy^{(i)} \\
&= \frac{\int_{-\infty}^{\infty} \exp\left(-\frac{1}{2}\left(\frac{1}{\sigma_y^2} + \frac{k}{\tau^2} \right) (y^{(i)})^2 + \left(\frac{\mu_y}{\sigma_y^2} + \sum_{j=1}^k \frac{\hat{y}_j^{(i)}}{\tau^2 \theta_j} \right) y^{(i)} - \frac{1}{2} \left(\frac{\mu_y^2}{\sigma_y^2} + \sum_{j=1}^k \frac{(\hat{y}_j^{(i)})^2}{\tau^2 \theta_j^2} \right) \right)}{\tau^k (\sqrt{2\pi})^{k+1} \sigma_y \prod_{j=1}^k \theta_j} \\
&= \frac{\sqrt{\pi} \left[\frac{1}{2} \left(\frac{1}{\sigma_y^2} + \frac{k}{\tau^2} \right) \right]^{-1/2}}{\tau^k (\sqrt{2\pi})^{k+1} \sigma_y \prod_{j=1}^k \theta_j} \exp\left(\frac{\left(\frac{\mu_y}{\sigma_y^2} + \sum_{j=1}^k \frac{\hat{y}_j^{(i)}}{\tau^2 \theta_j} \right)^2}{2\left(\frac{1}{\sigma_y^2} + \frac{k}{\tau^2} \right)} - \sum_{j=1}^{k} \frac{(\hat{y}_j^{(i)})^2}{2\tau^2 \theta_j^2} - \frac{\mu_y^2}{2\sigma_y^2} \right) \qquad (13)
\end{aligned}$$

where the last equation was obtained using Lemma 1 concerning Gaussian integrals. Note that this equation does not have a closed form maximizer requiring the use of iterative computational techniques.

2.3 Collaborative Estimation of the Risks: Conditionally Correlated Predictors

In some cases the conditional independence assumption made in the previous subsection does not hold and the factorization (12) is violated. In this section, we discuss how to relax this assumption in the classification case. A similar approach may also be used for regression. We omit the details here due to notational clarity.

There are several ways to relax the conditional independence assumption. Most popular, perhaps, is the mechanism of hierarchical loglinear models for categorical data (Bishop et al., 1975). For example, generalizing our conditional independence assumption to second-order interaction log-linear models we have

$$\log p(\hat{y}_1, \dots, \hat{y}_k | y) = \alpha_y + \sum_{i=1}^{l} \beta_{i,\hat{y}_i,y} + \sum_{i<j} \gamma_{i,j,\hat{y}_i,\hat{y}_j,y} \qquad (14)$$

where the following ANOVA-type parameter constraints are needed (Bishop et al., 1975)

$$\begin{aligned}
0 &= \sum_{\hat{y}_i} \beta_{i,\hat{y}_i,y} \quad \forall i, y, \qquad (15) \\
0 &= \sum_{\hat{y}_i} \gamma_{i,j,\hat{y}_i,\hat{y}_j,y} = \sum_{\hat{y}_j} \gamma_{i,j,\hat{y}_i,\hat{y}_j,y} \quad \forall i, j, y.
\end{aligned}$$

The β parameters in (14) correspond to the order-1 interaction between the variables $\hat{y}_1, \dots, \hat{y}_k$, conditioned on y. They correspond to the θ_i in the independent formulation (6)-(8). The γ parameters capture two-way interactions which do not appear in the conditionally independent case. Indeed, setting $\gamma_{i,j,\hat{y}_i,\hat{y}_j,y} = 0$ retrieves the independent models (6)-(8).

In the case of classification, the number of degrees of freedom or free unconstrained parameters in (14) depends on whether the number of classes is 2 or more and what additional assumptions exist on β and γ. For example, assuming that the probability of f_i, f_j making an error depends on the true class y but not on the predicted classes $\hat{y}_i, \hat{y}_j$ results in a $k+k^2$ parameters. Relaxing that assumption but assuming binary classification results in $2k+4k^2$ parameters. The estimation and aggregation techniques described in Section 2.1 work as before with a slight modification of replacing (6)-(8) with variations based on (14) and enforcing the constraints (15).

Equation (14) captures two-way interactions but cannot model higher order interactions. However, three-way and higher order interaction models are straightforward generalizations of (14) culminating in the full loglinear model which does not make any assumption on the statistical dependency of the noise operators $T_1, \dots, T_k$. However, as we weaken the assumptions underlying the loglinear models and add higher order interactions the number of parameters increases adding to the difficulty in estimating the risks $R(f_1), \dots, R(f_k)$.

In our experiments on real world data (see Section 7), it is often the case that maximizing the loglikelihood under the conditionally independent assumption (12) provides adequate accuracy and there is no need for the more general (14)-(15). Nevertheless, we include here the case of loglinear models as it may be necessary in some situations.

3. Extensions: Missing Values, Active Learning, and Semi-Supervised Learning

In this section, we discuss extensions to the current framework. Specifically, we consider extending the framework to the cases of missing values, active and semi-supervised learning.

Occasionally, some predictors are unable to provide their output over specific data points. That is assuming a data set $x^{(1)}, \dots, x^{(n)}$ each predictor may provide output on an arbitrary subset of the data points $\{f_j(x^{(i)}) : i \in S_j\}$, where $S_j \subset \{1, \dots, n\}$, $j = 1, \dots, k$.

Commonly referred to as a missing value situation, this scenario may apply in cases where different parts of the unlabeled data are available to the different predictors at test time due to privacy, computational complexity, or communication cost. Another example where this scenario applies is active learning where operating f_j involves a certain cost $c_j \geq 0$ and it is not advantageous to operate all predictors with the same frequency for the purpose of estimating the risks $R(f_1), \dots, R(f_k)$. Such is the case when f_j corresponds to judgments obtained from human experts or expensive machinery that is busy serving multiple clients. Active learning fits into this situation with S_j denoting the set of selected data points for each predictor.

We proceed in this case by defining indicators β_{ji} denoting whether predictor j is available to emit $f_j(x^{(i)})$. The risk estimation proceeds as before with the observed likelihood modified to account for the missing values.

In the case of collaborative estimation with conditional independence, the estimator and loglikelihood become

$$\begin{aligned}
\hat{\theta}_n^{\text{mle}} &= \arg\max_{\theta} \ell(\theta), \\
\ell(\theta) &= \sum_{i=1}^{n} \log \sum_{r:\beta_{ri}=0} \int_{\mathcal{Y}} p_\theta(\hat{y}_1^{(i)}, \dots, \hat{y}_k^{(i)})\, d\mu(\hat{y}_r^{(i)}) \qquad (16) \\
&= \sum_{i=1}^{n} \log \sum_{r:\beta_{ri}=0} \iint_{\mathcal{Y}^2} p_\theta(\hat{y}_1^{(i)}, \dots, \hat{y}_k^{(i)} | y^{(i)}) p(y^{(i)})\, d\mu(\hat{y}_r^{(i)}) d\mu(y^{(i)})
\end{aligned}$$

where p_θ may be further simplified using the non-collaborative approach, or using the collaborative approach with conditional independence or loglinear model assumptions.

In the case of semi-supervised learning a small set of labeled data is augmented by a large set of unlabeled data. In this case our framework remains as before with the likelihood summing over the observed labeled and unlabeled data. For example, in the case of collaborative estimation with conditional independence we have

$$\ell(\theta) = \sum_{i=1}^{n} \log \int_{\mathcal{Y}} \prod_{j=1}^{k} p_{\theta_j}(\hat{y}_j^{(i)}|y^{(i)}) p(y^{(i)})\, d\mu(y^{(i)}) + \sum_{i=n+1}^{m} \log \prod_{j=1}^{k} p_{\theta_j}(\hat{y}_j^{(i)}|y^{(i)}) p(y^{(i)}).$$

The different variations concerning missing values, active learning, semi-supervised learning, and non-collaborative or collaborative estimation with conditionally independent or correlated noise processes can all be combined in different ways to provide the appropriate likelihood function. This provides substantial modeling flexibility.

4. Consistency of $\hat{\theta}_n^{\mathbf{mle}}$ and $\hat{R}(f_j)$

In this and the next section we consider the statistical behavior of the estimator $\hat{\theta}_n^{\text{mle}}$ defined in (2) and the risk estimator $\hat{R}(f_j) = g_j(\hat{\theta}^{\text{mle}})$ defined in (1). The analysis is conducted under the assumption that the vectors of observed predictors outputs $\hat{y}^{(i)} = (\hat{y}_1^{(i)}, \dots, \hat{y}_k^{(i)})$ are iid samples from the distribution

$$p_\theta(\hat{y}) = p_\theta(\hat{y}_1, \dots, \hat{y}_k) = \int_{\mathcal{Y}} p_\theta(\hat{y}_1, \dots, \hat{y}_k|y) p(y)\, d\mu(y).$$

We start by investigating whether estimator $\hat{\theta}^{\text{mle}}$ in (2) converges to the true parameter value. More formally, strong consistency of the estimator $\hat{\theta}_n^{\text{mle}} = \hat{\theta}(\hat{y}^{(1)}, \dots, \hat{y}^{(n)})$, $\hat{y}^{(1)}, \dots, \hat{y}^{(n)} \overset{\text{iid}}{\sim} p_{\theta_0}$ is defined as strong convergence of the estimator to θ_0 as $n \to \infty$ (Ferguson, 1996)

$$\lim_{n\to\infty} \hat{\theta}_n^{\text{mle}}(\hat{y}^{(1)}, \dots, \hat{y}^{(n)}) = \theta_0 \text{ with probability 1.}$$

In other words as the number of samples n grows, the estimator will surely converge to the true parameter θ_0 governing the data generation process.

Assuming that the risks $R(f_j) = g_j(\theta)$ are defined using continuous functions g_j, strong consistency of $\hat{\theta}^{\text{mle}}$ implies strong convergence of $\hat{R}(f_j)$ to $R(f_j)$. This is due to the fact that continuity preserves limits. Indeed, as the g_j functions are continuous in both the classification and regression cases, strong consistency of the risk estimators $\hat{R}(f_j)$ reduces to strong consistency of the estimators $\hat{\theta}^{\text{mle}}$.

It is well known that the maximum likelihood estimator is often strongly consistent. Consider, for example, the following theorem.

Proposition 2 (e.g., Ferguson, 1996) *Let $\hat{y}^{(1)}, \dots, \hat{y}^{(n)} \overset{\text{iid}}{\sim} p_{\theta_0}$, $\theta_0 \in \Theta$. If the following conditions hold*

1. *Θ is compact* *(compactness)*
2. *$p_\theta(\hat{y})$ is upper semi-continuous in θ for all $\hat{y}$* *(continuity)*
3. *There exists a function $K(\hat{y})$ such that $E_{p_{\theta_0}}|K(\hat{y})| < \infty$ and $\log p_\theta(\hat{y}) - \log p_{\theta_0}(\hat{y}) \leq K(\hat{y}) \quad \forall \hat{y} \quad \forall \theta$* *(boundedness)*
4. *For all θ and sufficiently small $\rho > 0$, $\sup_{|\theta'-\theta|<\rho} p_{\theta'}(\hat{y})$ is measurable in $\hat{y}$* *(measurability)*
5. *$p_\theta \equiv p_{\theta_0} \Rightarrow \theta = \theta_0$* *(identifiability)*

then the maximum likelihood estimator is strongly consistent, that is, $\hat{\theta}^{mle} \to \theta_0$ as $n \to \infty$ with probability 1.

Note that $p_\theta(\hat{y})$ in the proposition above corresponds to $\int_{\mathcal{Y}} p_\theta(\hat{y}|y)p(y)\,d\mu(y)$ in our framework. That is the MLE operates on the observed data or predictor output $\hat{y}^{(1)},\dots,\hat{y}^{(n)}$ that is sampled iid from the distribution $p_{\theta_0}(\hat{y}) = \int_{\mathcal{Y}} p_{\theta_0}(\hat{y}|y)p(y)\,d\mu(y)$.

Of the five conditions above, the last condition of identifiability is the only one that is truly problematic. The first condition of compactness is trivially satisfied in the case of classification. In the case of regression it is satisfied assuming that the regression parameter and model parameter are finite and $a \neq 0$ as the estimator $\hat{\theta}^{\text{mle}}$ will eventually lie in a compact set. The second condition of continuity is trivially satisfied in both classification and regression as the function $\int_{\mathcal{Y}} p_\theta(\hat{y}|y)p(y)\,d\mu(y)$ is continuous in θ once $\hat{y}$ is fixed. The third condition is trivially satisfied for classification (finite valued y). In the case of regression due to conditions 1,2 (compactness and semi-continuity) we can replace the quantifier $\forall\theta$ with a particular value $\theta' \in \Theta$ representing worst case situation in the bound of the logarithm difference. Then, the bound K may be realized by the difference of log terms (with respect to that worst case θ') whose expectation converges to the KL divergence which in turn is never ∞ for Gaussian distributions or its derivatives. The fourth condition of measurability follows as p_θ is specified in terms of compositions, summations, multiplications, and point-wise limits of well-known measurable functions.

The fifth condition of identifiability states that if $p_\theta(\hat{y})$ and $p_{\theta_0}(\hat{y})$ are identical as functions, that is, they are identical for every value of $\hat{y}$, then necessarily $\theta = \theta_0$. This condition does not hold in general and needs to be verified in each one of the special cases.

We start with establishing consistency in the case of classification where we rely on a symmetric noise model (8). The non-symmetric case (6) is more complicated and is treated afterwards. We conclude the consistency discussion with an examination of the regression case.

4.1 Consistency of Classification Risk Estimation

Proposition 3 *Let $f_1,\dots,f_k$ be classifiers $f_i : \mathcal{X} \to \mathcal{Y}$, $|\mathcal{Y}| = l$, with conditionally independent noise processes described by* (8). *If the classifiers are weak learners, that is, $1/l < 1 - err(f_i) < 1$ and $p(y)$ is not uniform the unsupervised collaborative diagnosis model is identifiable.*

Corollary 4 *Let $f_1,\dots,f_k$ be classifiers $f_i : \mathcal{X} \to \mathcal{Y}$ with $|\mathcal{Y}| = l$ and noise processes described by* (8). *If the classifiers are weak learners, that is, $1/l < 1 - err(f_i) < 1$, and $p(y)$ is not uniform the unsupervised non-collaborative diagnosis model is identifiable.*

Proof Proving identifiability in the non-collaborative case proceeds by invoking Proposition 3 (whose proof is given below) with $k = 1$ separately for each classifier. The conditional independence assumption in Proposition 3 becomes redundant in this case of a single classifier, resulting in identifiability of $p_{\theta_j}(\hat{y}_j)$ for each $j = 1,\dots,k$ ■

Corollary 5 *Under the assumptions of Proposition 3 or Corollary 4 the unsupervised maximum likelihood estimator is consistent, that is,*

$$P\left(\lim_{n\to\infty} \hat{\theta}_n^{mle}(\hat{y}^{(1)},\dots,y^{(n)}) = (\theta_1^{true},\dots,\theta_k^{true})\right) = 1.$$

Consequentially, assuming that $R(f_j) = g_j(\theta), j = 1, \dots, k$ *with continuous* g_j *we also have*

$$P\left(\lim_{n\to\infty} \hat{R}(f_j; y^{(1)}, \dots, y^{(n)}) = R(f_j), \quad \forall j = 1, \dots, k\right) = 1.$$

Proof Proposition 3 or Corollary 4 establishes identifiability, which in conjunction with Proposition 2 proves the corollary. ■

Proof (**for Proposition 3**) We prove identifiability by induction on k. In the base case of $k = 1$, we have a set of l equations, corresponding to $i = 1, 2 \dots l$,

$$\begin{aligned} p_\theta(\hat{y}_1 = i) &= p(y = i)\theta_1 + \left(\sum_{j\neq i} p(y = j)\right)\frac{(1-\theta_1)}{(l-1)} \\ &= p(y = i)\theta_1 + (1 - p(y = i))\frac{(1-\theta_1)}{(l-1)} \\ &= \frac{\theta_1(lp(y = i) - 1) + 1 - p(y = i)}{(l-1)} \end{aligned}$$

from which we can see that if $\eta \neq \theta$ and $p(y = i) \neq 1/l$ then $p_\theta(\hat{y}_1) \neq p_\eta(\hat{y}_1)$. This proves identifiability for the base case of $k = 1$.

Next, we assume identifiability holds for k and prove that it holds for $k+1$. We do so by deriving a contradiction from the assumption that identifiability holds for k but not for $k+1$. We denote the parameters corresponding to the k labelers by the vectors $\theta, \eta \in [0,1]^k$ and the parameters corresponding the additional $k+1$ labeler by θ_{k+1}, η_{k+1}.

In the case of k classifiers we have

$$p_\theta(\hat{y}_1, \dots, \hat{y}_k) = \sum_{i=1}^{l} p_\theta(\hat{y}_1, \dots, \hat{y}_k | y = i) p(y = i) = \sum_{i=1}^{l} G(\mathcal{A}_i, \theta)$$

where

$$G(\mathcal{A}_i, \theta) \overset{\text{def}}{=} p(y = i) \prod_{j\in\mathcal{A}_i} \theta_j \cdot \prod_{j\notin\mathcal{A}_i} \frac{(1-\theta_j)}{(l-1)},$$
$$\mathcal{A}_i \overset{\text{def}}{=} \{j \in \{1, 2..., k\} : \hat{y}_j = i\}.$$

Note that the $\mathcal{A}_1, \dots, \mathcal{A}_l$ form a partition of $\{1, \dots, k\}$, that is, they are disjoint and their union is $\{1, \dots, k\}$.

In order to have unidentifiability for the $k+1$ classifiers we need $(\theta, \theta_{k+1}) \neq (\eta, \eta_{k+1})$ and the following l equations (corresponding to $\hat{y}_{k+1} = 1, 2, \dots, l$) to hold for any $\hat{y}_1, \dots, \hat{y}_k$ which corresponds to any partition $\mathcal{A}_1,\ldots,\mathcal{A}_l$

$$\begin{aligned}
\theta_{k+1}G(\mathcal{A}_1,\theta)+\frac{(1-\theta_{k+1})}{(l-1)}\sum_{i\neq 1}G(\mathcal{A}_i,\theta) &= \eta_{k+1}G(\mathcal{A}_1,\eta)+\frac{(1-\eta_{k+1})}{(l-1)}\sum_{i\neq 1}G(\mathcal{A}_i,\eta),\\
\theta_{k+1}G(\mathcal{A}_2,\theta)+\frac{(1-\theta_{k+1})}{(l-1)}\sum_{i\neq 2}G(\mathcal{A}_i,\theta) &= \eta_{k+1}G(\mathcal{A}_2,\eta)+\frac{(1-\eta_{k+1})}{(l-1)}\sum_{i\neq 2}G(\mathcal{A}_i,\eta),\\
&\vdots\\
\theta_{k+1}G(\mathcal{A}_l,\theta)+\frac{(1-\theta_{k+1})}{(l-1)}\sum_{i\neq l}G(\mathcal{A}_i,\theta) &= \eta_{k+1}G(\mathcal{A}_l,\eta)+\frac{(1-\eta_{k+1})}{(l-1)}\sum_{i\neq l}G(\mathcal{A}_i,\eta).
\end{aligned}$$

We consider two cases in which $(\theta,\theta_{k+1})\neq(\eta,\eta_{k+1})$: (a) $\theta\neq\eta$, and (b) $\theta=\eta,\theta_{k+1}\neq\eta_{k+1}$. In the case of (a) we add the l equations above which marginalizes $\hat{y}_{k+1}$ out of $p_\theta(\hat{y}_1,\ldots,\hat{y}_k,\hat{y}_{k+1})$ and $p_\eta(\hat{y}_1,\ldots,\hat{y}_k,\hat{y}_{k+1})$ to provide

$$\sum_{i=1}^{l}G(\mathcal{A}_i,\theta)=\sum_{i=1}^{l}G(\mathcal{A}_i,\eta)$$

which together with $\theta\neq\eta$ contradicts the identifiability for the case of k classifiers.

In case (b) we have from the l equations above

$$\begin{aligned}
\theta_{k+1}G(\mathcal{A}_t,\theta)+\frac{1-\theta_{k+1}}{l-1}\left(\sum_{i=1}^{l}G(\mathcal{A}_i,\theta)-G(\mathcal{A}_t,\theta)\right)&\\
=\eta_{k+1}G(\mathcal{A}_t,\eta)+\frac{1-\eta_{k+1}}{l-1}&\left(\sum_{i=1}^{l}G(\mathcal{A}_i,\eta)-G(\mathcal{A}_t,\eta)\right)
\end{aligned}$$

for any $t\in\{1,\ldots,l\}$ which simplifies to

$$0=(\theta_{k+1}-\eta_{k+1})\left(lG(\mathcal{A}_t,\theta)-\sum_{i=1}^{l}G(\mathcal{A}_i,\theta)\right)\quad t=1,\ldots,k.$$

As we assume at this point that $\theta_{k+1}\neq\eta_{k+1}$ the above equality entails

$$lG(\mathcal{A}_t,\theta)=\sum_{i=1}^{l}G(\mathcal{A}_i,\theta). \tag{17}$$

We show that (17) cannot hold by examining separately the cases $p(y=t)>1/l$ and $p(y=t)<1/l$. Recall that there exists a t for which $p(y=t)\neq 1/l$ since the proposition requires that $p(y)$ is not uniform.

If $p(y=t)>1/l$ we choose $\mathcal{A}_t=\{1,\ldots,k\}$ and obtain

$$\begin{aligned}
lp(y=t)\prod_{j=1}^{k}\theta_j&=\sum_{i\neq t}p(y=i)\prod_{j=1}^{k}\frac{1-\theta_j}{l-1}+p(y=t)\prod_{j=1}^{k}\theta_j\\
(l-1)p(y=t)\prod_{j=1}^{k}\theta_j&=(1-p(y=t))\prod_{j=1}^{k}\frac{1-\theta_j}{l-1}\\
p(y=t)\prod_{j=1}^{k}\theta_j&=\frac{(1-p(y=t))}{(l-1)}\prod_{j=1}^{k}\frac{1-\theta_j}{l-1}
\end{aligned}$$

which cannot hold as the term on the left hand side is necessarily larger than the term on the right hand side (if $p(y=t) > 1/l$ and $\theta_j > 1/l$). In the case $p(y=t) < 1/l$ we choose $\mathcal{A}_s = \{1,\dots,k\}$, $s \neq t$ to obtain

$$l p(y=t)\prod_{j=1}^{k}\frac{1-\theta_j}{l-1} = \sum_{i\neq s} p(y=i)\prod_{j=1}^{k}\frac{1-\theta_j}{l-1} + p(y=s)\prod_{j=1}^{k}\theta_j$$

$$(l p(y=t) - p(y\neq s))\prod_{j=1}^{k}\frac{1-\theta_j}{l-1} = p(y=s)\prod_{j=1}^{k}\theta_j$$

which cannot hold as the term on the left hand side is necessarily smaller than the term on the right hand side (if $p(y=t) < 1/l$ and $\theta_j > 1/l$).

Since we derived a contradiction to the fact that we have k-identifiability but not $k+1$ identifiability, the induction step is proven which establishes identifiability for any $k \geq 1$. ■

The conditions asserted above that $p(y) \neq 1/l$ and $1/l < 1 - \mathrm{err}(f_i) < 1$ are intuitive. If they are violated a certain symmetry may emerge which renders the model non-identifiable and the MLE estimator not consistent.

In the case of the non-collaborative estimation for binary classification with the non-symmetric noise model, the matrix θ in (6) is a 2×2 matrix with two degrees of freedom as each row sums to one. In particular we have $\theta_{11} = p_\theta(\hat{y}=1|y=1)$, $\theta_{12} = p_\theta(\hat{y}=1|y=2)$, $\theta_{21} = p_\theta(\hat{y}=2|y=1)$, $\theta_{22} = p_\theta(\hat{y}=2|y=2)$ with the overall risk $R(f) = 1-\theta_{11}p(y=1)-\theta_{22}p(y=2)$. Unfortunately, the matrix θ is not identifiable in this case and neither is the scalar parameter $\theta_{11}p(y=1)+\theta_{22}p(y=2)$ that can be used to characterize the risk.

We can, however, obtain a consistent estimator for θ (and therefore for $R(f)$) by first showing that the parameter $\theta_{11}p(y=1)-\theta_{22}p(y=2)$ is identifiable and then taking the intersection of two such estimators.

Lemma 6 *In the case of the non-collaborative estimation for binary classification with the non-symmetric noise model and $p(y)\neq 0$, the parameter $\theta_{11}p(y=1)-\theta_{22}p(y=2)$ is identifiable.*

Proof For two different parameterizations θ, η we have

$$p_\theta(\hat{y}=1) = p(y=1)\theta_{11} + (1-p(y=1))(1-\theta_{22}), \tag{18}$$

$$p_\theta(\hat{y}=2) = p(y=1)(1-\theta_{11}) + (1-p(y=1))\theta_{22} \tag{19}$$

and

$$p_\eta(\hat{y}=1) = p(y=1)\eta_{11} + (1-p(y=1))(1-\eta_{22}), \tag{20}$$

$$p_\eta(\hat{y}=2) = p(y=1)(1-\eta_{11}) + (1-p(y=1))\eta_{22}. \tag{21}$$

Equating the two Equations (18) and (20) we have

$$p(y=1)(\theta_{11}+\theta_{22}) + 1 - p(y=1) - \theta_{22} = p(y=1)(\eta_{11}+\eta_{22}) + 1 - p(y=1) - \eta_{22}$$

$$p(y=1)\theta_{11} - (1-p(y=1))\theta_{22} = p(y=1)\eta_{11} - (1-p(y=1))\eta_{22}$$

$$p(y=1)\theta_{11} - p(y=2)\theta_{22} = p(y=1)\eta_{11} - p(y=2)\eta_{22}$$

Similarly, equating Equation (19) and Equation (21) also results in $p(y=1)\theta_{11} - p(y=2)\theta_{22} = p(y=1)\eta_{11} - p(y=2)\eta_{22}$. As a result, we have

$$p_\theta \equiv p_\eta \quad \Rightarrow \quad p(y=1)\theta_{11} - p(y=2)\theta_{22} = p(y=1)\eta_{11} - p(y=2)\eta_{22}.$$

■

The above lemma indicates that we can use the maximum likelihood method to obtain a consistent estimator for the parameter $\theta_{11}p(y=1) - \theta_{22}p(y=2)$. Unfortunately the parameter $\theta_{11}p(y=1) - \theta_{22}p(y=2)$ does not have a clear probabilistic interpretation and does not directly characterize the risk. As the following proposition shows we can obtain a consistent estimator for the risk $R(f)$ if we have two populations of unlabeled data drawn from distributions with two distinct marginals $p_1(y)$ and $p_2(y)$.

Proposition 7 *Consider the case of the non-collaborative estimation of binary classification risk with the non-symmetric noise model. If we have access to two unlabeled data sets drawn independently from two distributions with different marginals, that is,*

$$x^{(1)}, \dots, x^{(n)} \overset{\text{iid}}{\sim} p_1(x) = \sum_y p(x|y)p_1(y),$$
$$x'^{(1)}, \dots, x'^{(m)} \overset{\text{iid}}{\sim} p_2(x) = \sum_y p(x|y)p_2(y)$$

we can obtain a consistent estimator for the classification risk $R(f)$.

Proof Operating the classifier f on both sets of unlabeled data we get two sets of observed classifier outputs $\hat{y}^{(1)}, \dots, \hat{y}^{(n)}$, $\hat{y}'^{(1)}, \dots, \hat{y}'^{(m)}$ where $\hat{y}^{(i)} \overset{\text{iid}}{\sim} \sum_y p_\theta(\hat{y}|y)p_1(y)$ and $\hat{y}'^{(i)} \overset{\text{iid}}{\sim} \sum_y p_\theta(\hat{y}|y)p_2(y)$. In particular, note that the marginal distributions $p_1(y)$ and $p_2(y)$ are different but the parameter matrix θ is the same in both cases as we operate the same classifier on samples from the same class conditional distribution $p(x|y)$.

Based on Lemma 6 we construct a consistent estimator for $p_1(y=1)\theta_{11} - p_1(y=2)\theta_{22}$ by maximizing the likelihood of $\hat{y}^{(1)}, \dots, \hat{y}^{(n)}$. Similarly, we construct a consistent estimator for $p_2(y=1)\theta_{11} - p_2(y=2)\theta_{22}$ by maximizing the likelihood of $\hat{y}'^{(1)}, \dots, \hat{y}'^{(m)}$. Note that $p_1(y=1)\theta_{11} - p_1(y=2)\theta_{22}$ and $p_2(y=1)\theta_{11} - p_2(y=2)\theta_{22}$ describe two lines in the 2-D space $(\theta_{11}, \theta_{22})$. Since the true value of θ_{11}, θ_{22} represent a point in that 2-D space belonging to both lines, it is necessarily the intersection of both lines (the lines cannot be parallel since their linear coefficients are distributions which are assumed to be different).

As n and m increase to infinity, the two estimators converge to the true parameter values. As a result, the intersection of the two lines described by the two estimators converges to the true values of $(\theta_{11}, \theta_{22})$ thus allowing reconstruction of the matrix θ and the risk $R(f)$. ■

Clearly, the conditions for consistency in the asymmetric case are more restricted than in the symmetric case. However, situations such as in Proposition 7 are not necessarily unrealistic. In many cases it is possible to identify two unlabeled sets with different distributions. For example, if y denotes a medical condition, it may be possible to obtain two unlabeled sets from two different hospitals or two different regions with different marginal distribution corresponding to the frequency of the medical condition.

As indicated in the previous section, the risk estimation framework may be extended beyond non-collaborative estimation and collaborative conditionally independent estimation. In these extensions, the conditions for identifiability need to be determined separately, in a similar way to Corollary 4. A systematic way to do so may be obtained by noting that the identifiability equations

$$0 = p_\theta(\hat{y}_1,\dots,\hat{y}_k) - p_\eta(\hat{y}_1,\dots,\hat{y}_k) \quad \forall \hat{y}_1,\dots,\hat{y}_k$$

is a system of polynomial equations in (θ,η). As a result, demonstrating lack of identifiability becomes equivalent to obtaining a solution to a system of polynomial equations. Using Hilbert's Nullstellensatz theorem we have that a solution to a polynomial system exists if the polynomial system defines a proper ideal of the ring of polynomials (Cox et al., 2006). As k increases the chance of identifiability failing decays dramatically as we have a system of l^k polynomials with $2k$ variables. Such an over-determined system with substantially more equations than variables is very unlikely to have a solution.

These observations serve as both an interesting theoretical connection to algebraic geometry as well as a practical tool due to the substantial research in computational algebraic geometry. See Sturmfels (2002) for a survey of computational algorithms and software associated with systems of polynomial equations.

4.2 Consistency of Regression Risk Estimation

In this section, we prove the consistency of the maximum likelihood estimator $\hat{\theta}^{\text{mle}}$ in the regression case. As in the classification case our proof centers on establishing identifiability.

Proposition 8 *Let $f_1,\dots,f_k$ be regression models $f_i(x) = a_i'x$ with $y \sim N(\mu_y,\sigma_y^2)$, $y = ax+\epsilon$. Assuming that $a \neq 0$ the unsupervised collaborative estimation model assuming conditionally independent noise processes* (12) *is identifiable.*

Corollary 9 *Let $f_1,\dots,f_k$ be regression models $f_i(x) = a_i'x$ with $y \sim N(\mu_y,\sigma_y^2)$, $y = ax+\epsilon$. Assuming that $a \neq 0$ the unsupervised non-collaborative estimation model* (12) *is identifiable.*

Proof Proving identifiability in the non-collaborative case proceeds by invoking Proposition 8 (whose proof is given below) with $k=1$ separately for each regression model. The conditional independence assumption in Proposition 8 becomes redundant in this case of a single predictor, resulting in identifiability of $p_{\theta_j}(\hat{y}_j)$ for each $j = 1,\dots,k$. ■

Corollary 10 *Under the assumptions of Proposition 8 or Corollary 9 the unsupervised maximum likelihood estimator is consistent, that is,*

$$P\left(\lim_{n\to\infty} \hat{\theta}_n^{mle}(\hat{y}^{(1)},\dots,y^{(n)}) = (\theta_1^{true},\dots,\theta_k^{true})\right) = 1.$$

Consequentially, assuming that $R(f_j) = g_j(\theta)$, $j = 1,\dots,k$ with continuous g_j we also have

$$P\left(\lim_{n\to\infty} \hat{R}(f_j;y^{(1)},\dots,y^{(n)}) = R(f_j), \quad \forall j = 1,\dots,k\right) = 1.$$

Proof Proposition 8 or Corollary 9 establish identifiability, which in conjunction with Proposition 2 completes the proof. ■

Proof (of Proposition 8).

We will proceed, as in the case of classification, with induction on the number of predictors k. In the base case of $k = 1$ we have derived $p_{\theta_1}(\hat{y}_1)$ in Equation (11). Substituting in it $\hat{y}_1 = 0$ we get

$$P_{\theta_1}(\hat{y}_1 = 0) = \frac{1}{\theta_1\sqrt{2\pi(\tau^2+\sigma_y^2)}} \exp\left(\frac{\mu_y^2}{2\sigma_y^2}\left(\frac{\tau^2}{\sigma_y^2+\tau^2}-1\right)\right),$$

$$P_{\eta_1}(\hat{y}_1 = 0) = \frac{1}{\eta_1\sqrt{2\pi(\tau^2+\sigma_y^2)}} \exp\left(\frac{\mu_y^2}{2\sigma_y^2}\left(\frac{\tau^2}{\sigma_y^2+\tau^2}-1\right)\right).$$

The above expression leads to $\theta_1 \neq \eta_1 \Rightarrow p_{\theta_1}(\hat{y}_1 = 0) \neq p_{\eta_1}(\hat{y}_1 = 0)$ which implies identifiability.

In the induction step we assume identifiability holds for k and we prove that it holds also for $k+1$ by deriving a contradiction to the assumption that it does not hold. We assume that identifiability fails in the case of $k+1$ due to differing parameter values, that is,

$$p_{(\theta,\theta_{k+1})}(\hat{y}_1,\dots,\hat{y}_k,\hat{y}_{k+1}) = p_{(\eta,\eta_{k+1})}(\hat{y}_1,\dots,\hat{y}_k,\hat{y}_{k+1}) \ \forall \hat{y}_j \in \mathbb{R} \ j = 1,\dots,k+1 \qquad (22)$$

with $(\theta,\theta_{k+1}) \neq (\eta,\eta_{k+1})$ where $\theta,\eta \in \mathbb{R}^k$. There are two cases which we consider separately: (a) $\theta \neq \eta$ and (b) $\theta = \eta$.

In case (a) we marginalize both sides of (22) with respect to $\hat{y}_{k+1}$ which leads to a contradiction to our assumption that identifiability holds for k

$$\int_{-\infty}^{\infty} p_{(\theta,\theta_{k+1})}(\hat{y}_1,\dots,\hat{y}_k,\hat{y}_{k+1})d\hat{y}_{k+1} = \int_{-\infty}^{\infty} p_{(\eta,\eta_{k+1})}(\hat{y}_1,\dots,\hat{y}_k,\hat{y}_{k+1})d\hat{y}_{k+1}$$
$$p_\theta(\hat{y}_1,\dots,\hat{y}_k) = p_\eta(\hat{y}_1,\dots,\hat{y}_k).$$

In case (b) $\theta = \eta$ and $\theta_{k+1} \neq \eta_{k+1}$. Substituting $\hat{y}_1 = \cdots = \hat{y}_{k+1} = 0$ in (22) (see (13) for a derivation) we have

$$P_{(\theta,\theta_{k+1})}(\hat{y}_1 = 0,\dots,\hat{y}_{k+1} = 0) = P_{(\eta,\eta_{k+1})}(\hat{y}_1 = 0,\dots,\hat{y}_{k+1} = 0)$$

or

$$\frac{\sqrt{\pi}\left[\frac{1}{2}\left(\frac{1}{\sigma_y^2}+\frac{k+1}{\tau^2}\right)\right]^{-1/2}}{\tau^{k+1}(\sqrt{2\pi})^{k+2}\sigma_y\theta_{k+1}\prod_{j=1}^k\theta_j} \exp\left(\frac{\left(\frac{\mu_y}{\sigma_y^2}\right)^2}{2\left(\frac{1}{\sigma_y^2}+\frac{k+1}{\tau^2}\right)} - \frac{\mu_y^2}{2\sigma_y^2}\right)$$
$$= \frac{\sqrt{\pi}\left[\frac{1}{2}\left(\frac{1}{\sigma_y^2}+\frac{k+1}{\tau^2}\right)\right]^{-1/2}}{\tau^{k+1}(\sqrt{2\pi})^{k+2}\sigma_y\eta_{k+1}\prod_{j=1}^k\eta_j} \exp\left(\frac{\left(\frac{\mu_y}{\sigma_y^2}\right)^2}{2\left(\frac{1}{\sigma_y^2}+\frac{k+1}{\tau^2}\right)} - \frac{\mu_y^2}{2\sigma_y^2}\right)$$

which cannot hold if $\theta = \eta$ but $\theta_{k+1} \neq \eta_{k+1}$. ■

5. Asymptotic Variance of $\hat{\theta}_n^{\mathbf{mle}}$ and $\hat{R}$

A standard result from statistics is that the MLE has an asymptotically normal distribution with mean vector θ^{true} and variance matrix $(nJ(\theta^{\text{true}}))^{-1}$, where $J(\theta)$ is the $r \times r$ Fisher information matrix

$$J(\theta) = \mathsf{E}_{p_\theta}\{\nabla \log p_\theta(\hat{y})(\nabla \log p_\theta(\hat{y}))^\top\}$$

with $\nabla \log p_\theta(\hat{y})$ represents the $r \times 1$ gradient vector of $\log p_\theta(\hat{y})$ with respect to θ. Stated more formally, we have the following convergence in distribution as $n \to \infty$ (Ferguson, 1996)

$$\sqrt{n}\,(\hat{\theta}_n^{\text{mle}} - \theta_0) \rightsquigarrow N(0, J^{-1}(\theta^{\text{true}})). \tag{23}$$

It is instructive to consider the dependency of the Fisher information matrix, which corresponds to the asymptotic estimation accuracy, on $n, k, p(y), \theta^{\text{true}}$.

In the case of classification considering (8) with $k = 1$ and $\mathcal{Y} = \{1, 2\}$ it can be shown that

$$J(\theta) = \frac{\alpha(2\alpha-1)^2}{(\theta(2\alpha-1)-\alpha+1)^2} - \frac{(2\alpha-1)^2(\alpha-1)}{(\alpha-\theta(2\alpha-1))^2} \tag{24}$$

where $\alpha = P(y = 1)$. As Figure 3 (right) demonstrates, the asymptotic accuracy of the MLE (as indicated by J) tends to increase with the degree of non-uniformity of $p(y)$. Recall that since identifiability fails for a uniform $p(y)$ the risk estimate under a uniform $p(y)$ is not consistent. The above derivation (24) is a quantification of that fact reflecting the added difficulty in estimating the risk as we move closer to a uniform label distribution $\alpha \to 1/2$. The dependency of the asymptotic accuracy on θ^{true} is more complex, tending to favor θ^{true} values close to 1 or 0.5. Figure 3 (left) displays the empirical accuracy of the estimator as a function of $p(y)$ and θ^{true} and shows remarkable similarity to the contours of the Fisher information (see Section 7 for more details on the experiments). In particular, whenever the estimation error is high the asymptotic variance of the estimator is high (or equivalently, the Fisher information is low). For instance, the top contours in the left panel have smaller estimation error on the top right than in the top left. Similarly, the top contours in the right panel have smaller asymptotic variance on the top right than on the top left. We thus conclude that the Fisher information provides practical, as well as theoretical insight into the estimation accuracy.

Similar calculations of $J(\theta^{\text{true}})$ for collaborative classification case or for the regression case result in more complicated but straightforward derivations. It is important to realize that consistency is ensured for any identifiable $\theta^{\text{true}}, p(y)$. The value $(J(\theta^{\text{true}}))^{-1}$ is the constant dominating that consistency convergence.

A similar distributional analysis can be derived for the risk estimator. Applying Cramer's theorem (Ferguson, 1996) to $\hat{R}(f_j) = g_j(\hat{\theta}^{\text{mle}})$, $j = 1, \ldots, k$ and (23) we have

$$\sqrt{n}(\hat{R}(f) - R(f)) \rightsquigarrow N\left(0, \nabla g(\theta^{\text{true}}) J(\theta^{\text{true}}) \nabla g(\theta^{\text{true}})^\top\right)$$

where $R(f), \hat{R}(f)$ are the vectors of true risk and risk estimates for the different predictors $f_1, \ldots, f_k$ and $\nabla g(\theta^{\text{true}})$ is the Jacobian matrix of the mapping $g = (g_1, \ldots, g_k)$ evaluated at θ^{true}.

For example, in the case of classification with $k = 1$ we have $R(f_j) = 1 - \theta_j$ and the Jacobian matrix is -1, leading to an identical asymptotic distribution to that of the MLE (23)-(24)

$$\sqrt{n}(\hat{R}(f) - R(f)) \rightsquigarrow N\left(0, \left(\frac{\alpha(2\alpha-1)^2}{(\theta(2\alpha-1)-\alpha+1)^2} - \frac{(2\alpha-1)^2(\alpha-1)}{(\alpha-\theta(2\alpha-1))^2}\right)^{-1}\right).$$

6. Optimization Algorithms

Recall that we obtained closed forms for the likelihood maximizers in the cases of non-collaborative estimation for binary classifiers and non-collaborative estimation for one dimensional regression models. The lack of closed form maximizers in the other cases necessitates iterative optimization techniques.

One class of technique for optimizing nonlinear loglikelihoods is the class of gradient based methods such as gradient descent, conjugate gradients, and quasi Newton methods. These techniques proceed iteratively following a search direction; they often have good performance and are easy to derive. The main difficulty with their implementation is the derivation of the loglikelihood and its derivatives. For example, in the case of collaborative estimation of classification ($l \geq 2$) with symmetric noise model and missing values the loglikelihood gradient is

$$\frac{\partial \ell}{\partial \theta_j} = \sum_{i=1}^{n} \frac{\sum_{y^{(i)}} p(y^{(i)}) \sum_{r:\beta_{ri}=0} \sum_{\hat{y}_r^{(i)}} \prod_{p \neq j} h_{pi} (I(\hat{y}_j^{(i)} = y^{(i)}) - \theta_j)((l-1)\theta_j)^{I(\hat{y}_j^{(i)}=y^{(i)})-1} (1-\theta_j)^{-I(\hat{y}_j^{(i)}=y^{(i)})}}{\sum_{y^{(i)}} p(y^{(i)}) \sum_{r:\beta_{ri}=0} \sum_{\hat{y}_r^{(i)}} \prod_{p=1}^{k} h_{pi}},$$

$$h_{pi} = \theta_p^{I(\hat{y}_p^{(i)}=y^{(i)})} \left(\frac{1-\theta_p}{l-1}\right)^{I(\hat{y}_p^{(i)} \neq y^{(i)})}$$

Similar derivations may be obtained in the other cases in a straightforward manner.

An alternative iterative optimization technique for finding the MLE is expectation maximization (EM). The derivation of the EM update equations is again relatively straightforward. For example in the above case of collaborative estimation of classification ($l \geq 2$) with symmetric noise model and missing values the EM update equations are

$$\begin{aligned}
\theta^{(t+1)} &= \arg\max_{\theta} \sum_{i=1}^{n} \sum_{y^{(i)}} \sum_{r:\beta_{ri}=0} \sum_{\hat{y}_r^{(i)}} q^{(t)}(\hat{y}_r^{(i)}, y^{(i)}) \sum_{j=1}^{k} \log p_j(\hat{y}_j^{(i)} | y^{(i)}) \\
&= \frac{1}{n} \sum_{i=1}^{n} \sum_{y^{(i)}} \sum_{r:\beta_{ri}=0} \sum_{\hat{y}_r^{(i)}} q^{(t)}(\hat{y}_r^{(i)}, y^{(i)}) I(\hat{y}_j^{(i)} = y^{(i)}), \\
q^{(t)}(\hat{y}_r^{(i)}, y^{(i)}) &= \frac{p(y^{(i)}) \prod_{j=1}^{k} p_j(\hat{y}_j^{(i)} | y^{(i)}, \theta^{(t)})}{\sum_{y^{(i)}} \sum_{r:\beta_{ri}=0} \sum_{\hat{y}_r^{(i)}} p(y^{(i)}) \prod_{j=1}^{k} p_j(\hat{y}_j^{(i)} | y^{(i)}, \theta^{(t)})}.
\end{aligned}$$

where $q^{(t)}$ is the conditional distribution defining the EM bound over the loglikelihood function.

If all the classifiers are always observed, that is, $\beta_{ri} = 1 \ \forall r, i$ Equation (16) reverts to (12), and the loglikelihood and its gradient may be efficiently computed in $O(nlk^2)$. In the case of missing classifier outputs a naive computation of the gradient or EM step is exponential in the number of missing values $R = \max_i \sum_r \beta_{ri}$. This, however, can be improved by careful dynamic programming. For example, the nested summations over the unobserved values in the gradient may be computed using a variation of the elimination algorithm in $O(nlk^2R)$ time.

7. Empirical Evaluation

We start with some experiments demonstrating our framework using synthetic data. These experiments are meant to examine the behavior of the estimators in a controlled setting. We then describe

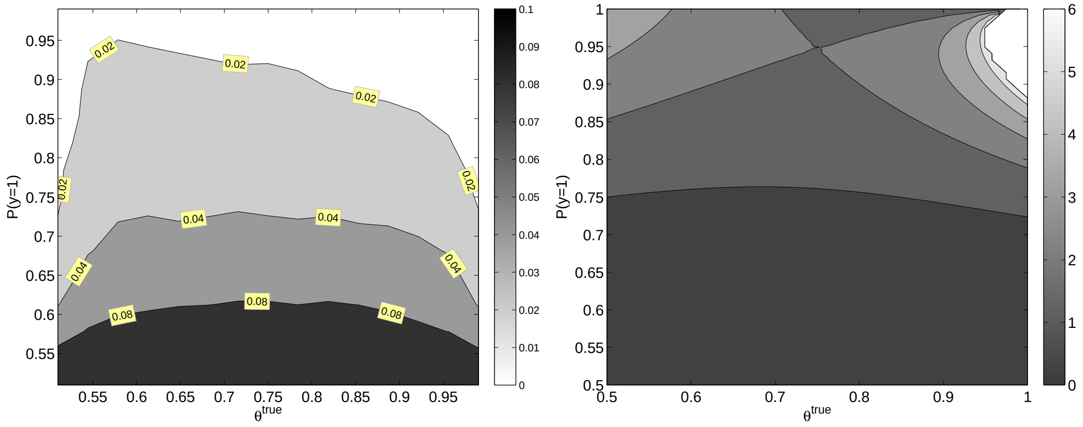


Figure 3: Left: Average value of $|\hat{\theta}_n^{\text{mle}} - \theta^{\text{true}}|$ as a function of θ^{true} and $p(y=1)$ for $k=1$ classifier and $n=500$ (computed over a uniform spaced grid of 15×15 points). The plot illustrates the increased accuracy obtained by a less uniform $P(y)$. Right: Fisher information $J(\theta)$ for $k=1$ as a function of θ^{true} and $P(y)$. The asymptotic variance of the estimator is $J^{-1}(\theta)$ which closely matches the experimental result in the left panel.

some experiments using several real world data sets. In these experiments we examine the behavior of the estimators in an uncontrolled setting where some of the underlying assumptions may be violated. In most of the experiments we consider the mean absolute error (mae) or the ℓ_1 error as a metric that measures the estimation quality

$$\text{mae}(\hat{\theta}^{\text{mle}}, \theta^{\text{true}}) = \frac{1}{k}\sum_{i=1}^{k} \left|\theta_i^{\text{true}} - \hat{\theta}_i^{\text{mle}}\right|.$$

In the non-collaborative case (which is equivalent to the collaborative case with $k=1$) this translates into the absolute deviation of the estimated parameter from the true parameter.

In Figure 3 (left) we display $\text{mae}(\hat{\theta}^{\text{mle}}, \theta^{\text{true}})$ for classification with $k=1$ as a function of θ^{true} and $p(y)$ for $n=500$ simulated data points. The estimation error, while overall relatively small, decays as $p(y)$ diverges from the uniform distribution. The dependency on θ^{true} indicates that the error is worst for θ^{true} around 0.75 and it decays as $|\theta^{\text{true}} - 0.75|$ increases with a larger decay attributed to higher θ^{true}. These observations are remarkably consistent with the developed theory as Figure 3 (right) shows by demonstrating the value of the inverse asymptotic variance $J(\theta)$ which agrees nicely with the empirical measurement in the left panel.

Figure 4 (left) contains a scatter plot contrasting values of θ^{true} and $\hat{\theta}^{\text{mle}}$ for $k=1$ classifier and $p(y=1)=0.8$. The estimator was constructed based on 500 simulated data points. We observe a symmetric Gaussian-like distribution of estimated values $\hat{\theta}^{\text{mle}}$, conditioned on specific values of θ^{true}. This is in agreement with the theory predicting an asymptotic Gaussian distribution for the mle, centered around the true value θ^{true}. A similar observation is made in Figure 5 (left) which contains a similar scatter plot in the regression case ($k=1$, $\sigma_y=1$, $n=1000$). In both figures, the striped effect is due to selection of θ^{true} over a discrete grid with a small perturbation for increased visibility. Similar plots of larger and smaller n values (not shown) verify that the variation of $\hat{\theta}^{\text{mle}}$

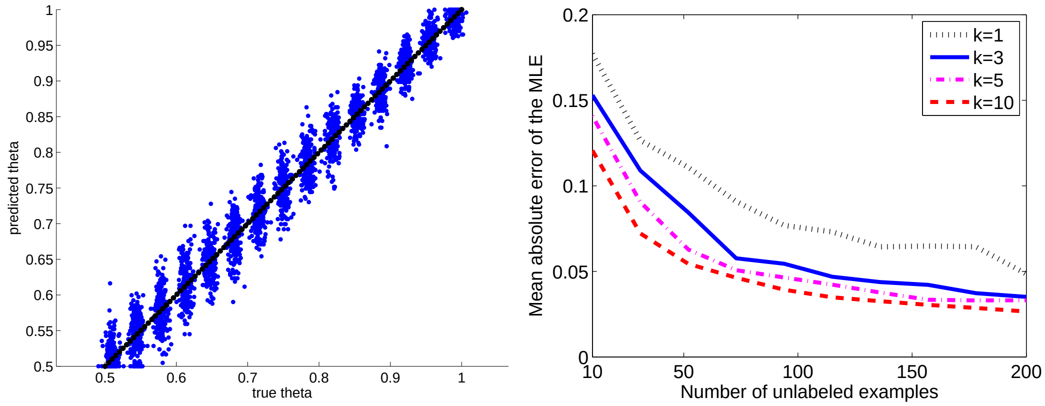


Figure 4: Left: Scatter plot contrasting the true and predicted values of θ in the case of a single classifier $k = 1$, $p(y = 1) = 0.8$, and $n = 500$ unlabeled examples. The displayed points were perturbed for improved visualization and the striped effect is due to empirical evaluation over a discrete grid of θ^{true} values. Right: $\text{mae}(\hat{\theta}^{\text{mle}}, \theta^{\text{true}})$ as a function of the number of unlabeled examples for different number of classifiers ($\theta_i^{\text{true}} = p(y = 1) = 0.75$) in the collaborative case. The estimation error decreases as more classifiers are used due to the collaborative nature of the estimation process.

around θ^{true} decreases as n increases. This agrees with the theory that indicates a $O(n^{-1})$ rate of decay for the variance of the asymptotic distribution.

Figures 4 and 5 (right) show the $\text{mae}(\hat{\theta}^{\text{mle}}, \theta^{\text{true}})$ for various k values in classification and regression, respectively. In classification, $\hat{\theta}^{\text{mle}}$ was obtained by sampling data from $p(y = 1) = 0.75 = \theta_i^{\text{true}}, \forall i$. In regression, the data was sampled from the regression equation with $\theta_i^{\text{true}} = 1$ and $p(y) = N(0, 1)$. In both cases, the mae error decays with n as expected from the consistency proof and with k as a result of the collaborative estimation effect.

To further illustrate the effect of the collaboration on the estimation accuracy, we estimated the error rates individually (non-collaboratively) for 10 predictors and compared their mae to that of the collaborative estimation case in Figure 6. This shows that each of the classifiers have a similar mae curve when non-collaborative estimation is used. However, all of these curves are higher than the collaborative mae curve (solid black line in Figure 6) demonstrating the improvement of the collaborative process.

We compare in Figure 7 the proposed unsupervised estimation framework with supervised estimation that takes advantage of labeled information to determine the classifier accuracy. We conducted this study using equal number of examples for both supervised and unsupervised cases. Clearly, this is an unfair comparison if we assume that labeled data is unavailable or is difficult to obtain. The unsupervised estimation does not perform as well as the supervised version especially in general. Nevertheless, the unsupervised estimation accuracy improves significantly with increasing number of classifiers and finally reaches the performance level of the supervised case due to collaborative estimation.

In Figure 8 we report the effect of misspecification of the marginal $p(y)$ on the estimation accuracy. More specifically, we generated synthetic data using a true marginal distribution but

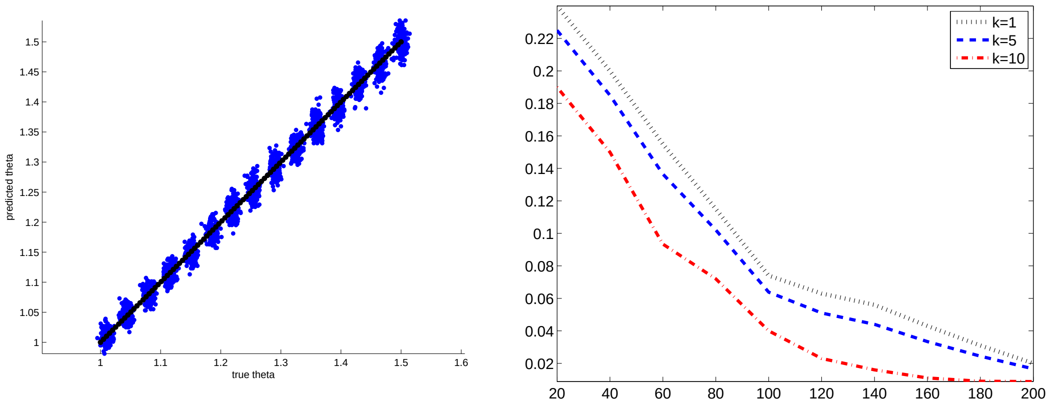


Figure 5: Left: Scatter plot contrasting the true and predicted values of θ in the case of a single regression model $k = 1$, $\sigma_y = 1$, and $n = 1000$ unlabeled examples. The displayed points were perturbed for improved visualization and the striped effect is due to empirical evaluation over a discrete grid of θ^{true} values. Right: $\text{mae}(\hat{\theta}^{\text{mle}}, \theta^{\text{true}})$ as a function of the number of unlabeled examples for different number of regression models ($\theta_i^{\text{true}} = \sigma_y = 1$) in the collaborative case. The estimation error decreases as more regression models are used due to the collaborative nature of the estimation process.

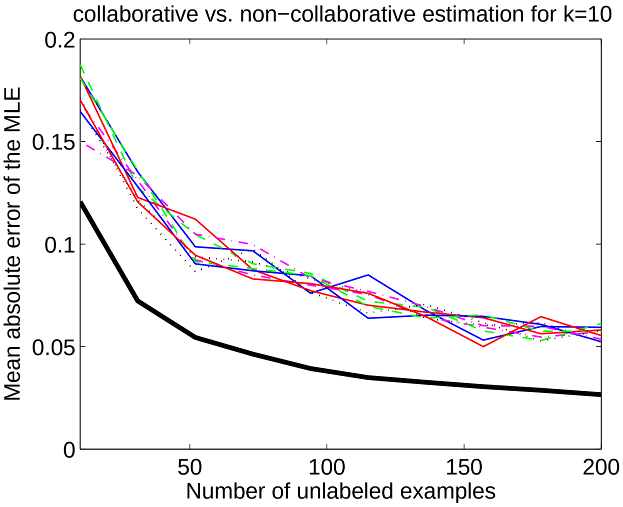


Figure 6: Comparison of collaborative and non-collaborative estimation for $k = 10$ classifiers. $\text{mae}(\hat{\theta}^{\text{mle}}, \theta^{\text{true}})$ as a function of n is reported for $\theta_i^{\text{true}} = 0.75\ \forall k_i$ and $P(y = 1) = 0.75$. The colored lines represent the estimation error for each individual classifier and the solid black line represents the collaborative estimation for all classifiers. The estimation converges to the truth faster in the collaborative case than in the non-collaborative case.

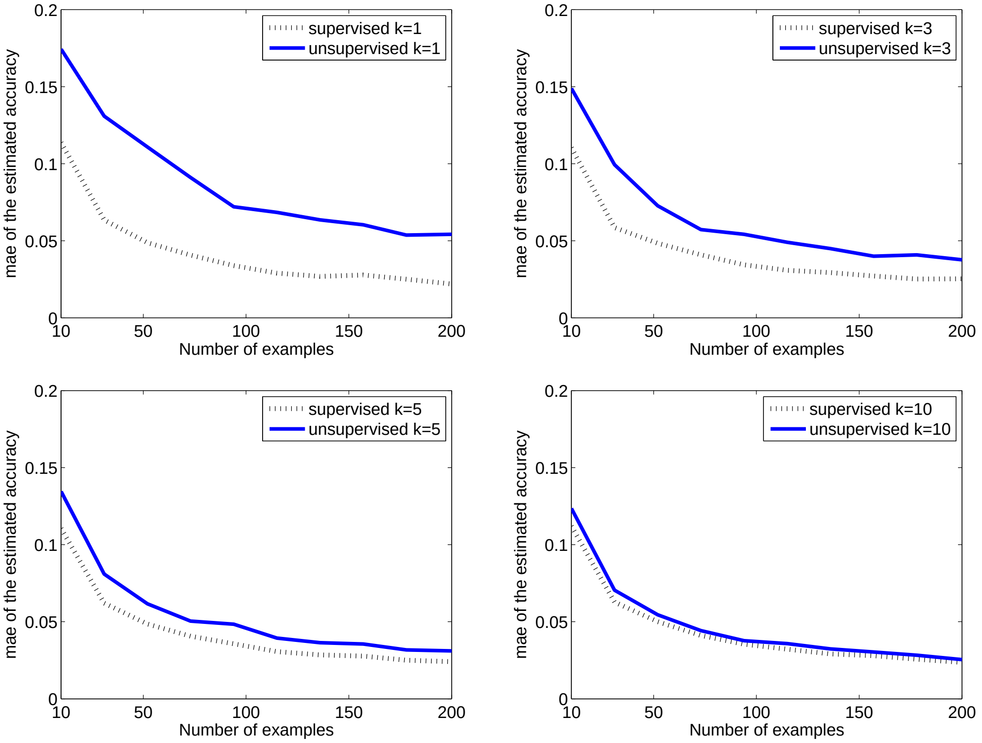


Figure 7: Comparison of supervised and unsupervised estimation for different values of classifiers with $k = 1,3,5,10$. Supervised estimation uses the true labels to determine the accuracy of the classifiers whereas in the unsupervised case the estimation proceeds according to the collaborative estimation framework. Despite the fact that the supervised case uses labels the unsupervised framework reaches similar levels by increasing the number of classifiers.

estimated the classifier accuracy on this data assuming a misspecified marginal. Generally, the estimation framework is robust to small perturbations while over-specifying tends to hurt less than under-specifying (misspecification closer to uniform distribution).

Figure 9 shows the mean prediction accuracy for the unsupervised predictor combination scheme in (4) for synthetic data. The left panel displays classification accuracy and the right panel displays the regression accuracy as measured by $1 - \frac{1}{m}\sum_{i=1}^{m}(\hat{y}_i^{\text{new}} - y_i^{\text{new}})^2$. The graphs show that in both cases the accuracy increases with k and n in accordance with the theory and the risk estimation experiments. The parameter θ_i^{true} was chosen uniformly in the range $(0.5,1)$, and $P(y = 1) = 0.75$ for classification and $\theta_i^{\text{true}} = 0.3$, $p(y) = N(0,1)$ in the case of regression.

We also experimented with the natural language understanding data set introduced in Snow et al. (2008). This data was created using the Amazon Mechanical Turk (AMT) for data annotation. AMT is an online tool that uses paid employees to complete small labeling and annotation tasks. We selected two binary tasks from this data: the textual entailment recognition (RTE) and temporal event recognition (TEMP) tasks. In the former task, the annotator is presented with two sentences for each question. He needs to decide whether the second sentence can be inferred from the first.

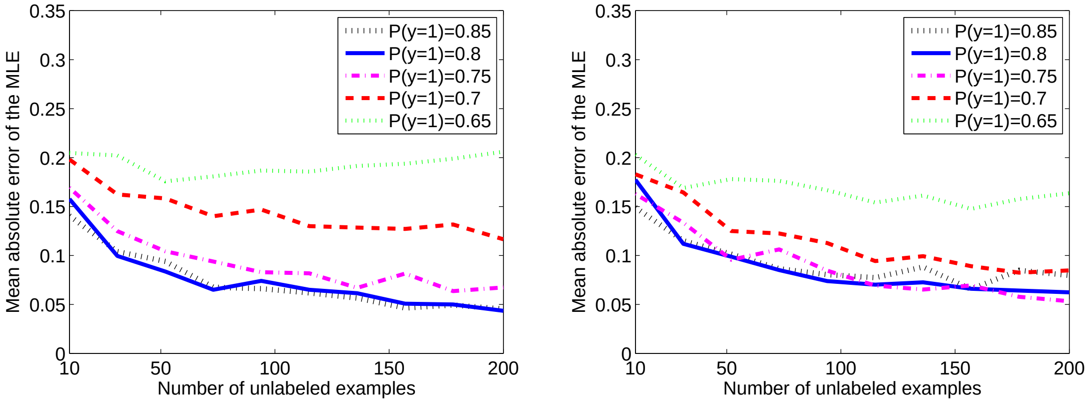


Figure 8: The figure compares the estimator accuracy assuming that the marginal $p(y)$ is misspecified. The plots draw mae$(\hat{\theta}^{\text{mle}}, \theta^{\text{true}})$ as a function of n for $k=1$ and $\theta^{\text{true}} = 0.75$ when $P^{\text{true}}(y=1) = 0.8$ (left) and $P^{\text{true}}(y=1) = 0.75$ (right). Small perturbations in $P^{\text{true}}(y)$ do not affect the results significantly; interestingly over-specifying $P^{\text{true}}(y=1)$ leads to more accurate estimates than under-specifying (misspecification closer to uniform distribution)

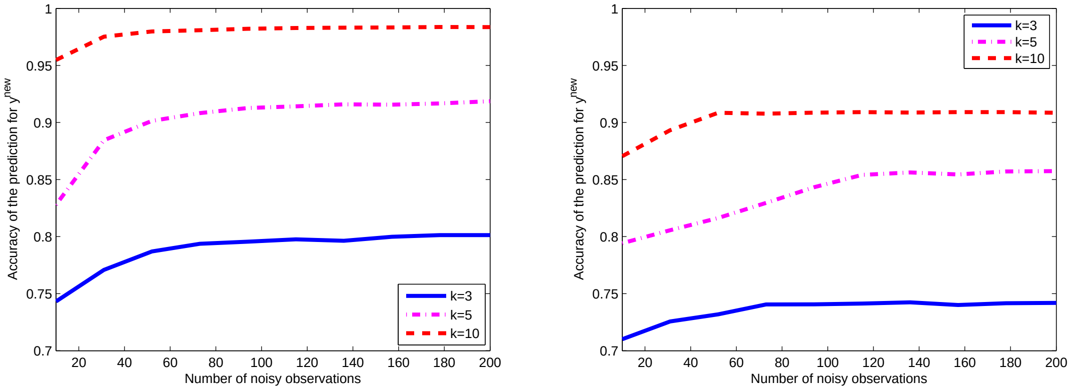


Figure 9: Mean prediction accuracy for the unsupervised predictor combination scheme in (4) for synthetic data. The left panel displays classification accuracy and the right panel displays the regression accuracy as measured by $1 - \frac{1}{m}\sum_{i=1}^{m}(\hat{y}_i^{\text{new}} - y_i^{\text{new}})^2$. The graphs show that in both cases the accuracy increases with k and n in accordance with the theory and the risk estimation experiments.

The original data set contains 800 sentence pairs with a total of 165 annotators. The latter task involves recognizing the temporal relation in verb-event pairs. The annotator is forced to decide whether the event described by the first verb occurs before or after the second. The original data set contains 462 pairs and 76 annotators. In both data sets, most of the annotators have completed only a handful of tasks. Therefore, we selected a subset of these annotators for each task such that each annotator has completed at least 100 problems and has differing accuracies. The data sets contain ground truth labels which are used solely to calculate the annotator accuracy and not used

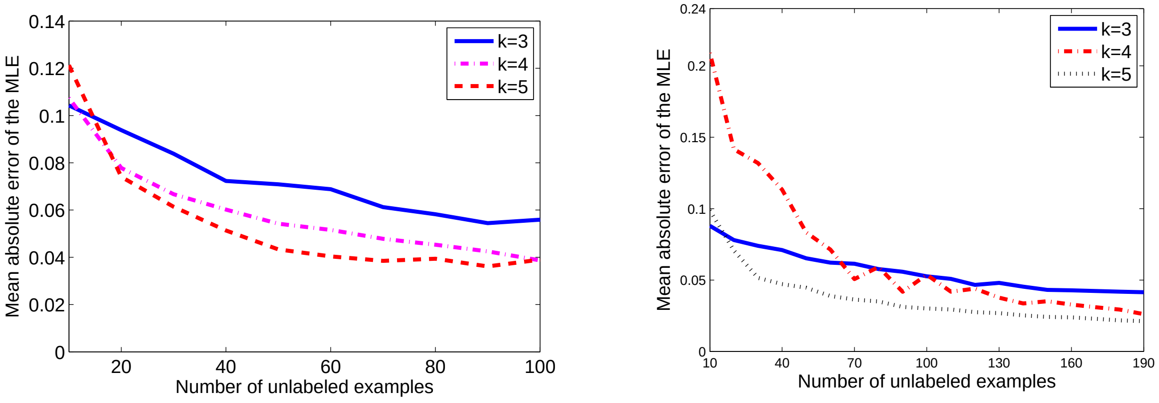


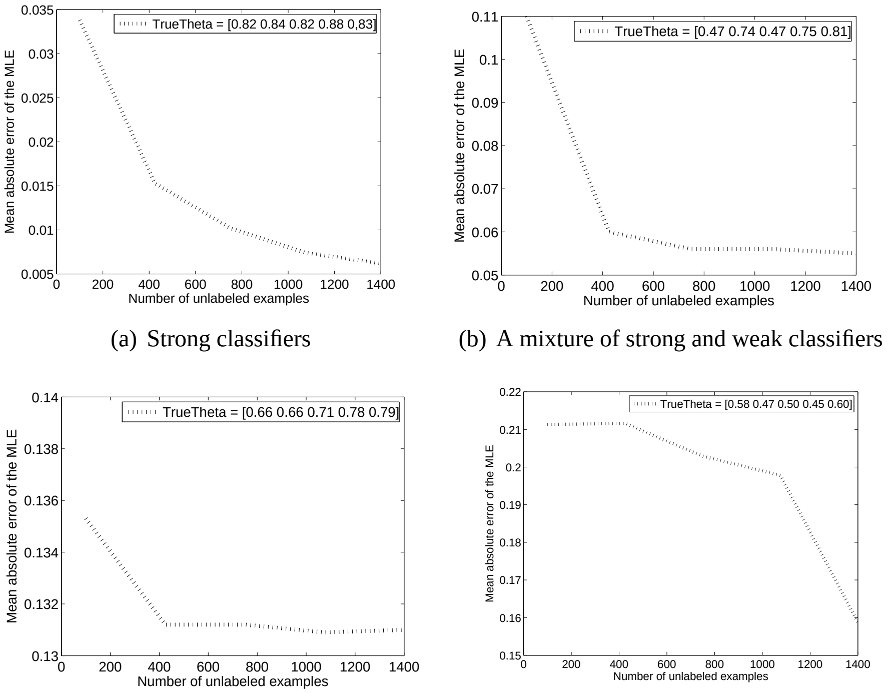
Figure 10: $\mathrm{mae}(\hat{\theta}^{\mathrm{mle}}, \theta^{\mathrm{true}})$ as a function of n for different number of annotators k on RTE (left) and TEMP (right) data sets. Left: $n = 100$, $P(y = 1) = 0.5$ and $\theta^{\mathrm{true}} = \{0.85, 0.92, 0.58, 0.5, 0.51\}$. Right: $n = 190$, $P(y = 1) = 0.56$ and $\theta^{\mathrm{true}} = \{0.93, 0.92, 0.54, 0.44, 0.92\}$. The classifiers were added in the order specified.

at all during the estimation process. For efficiency, we selected only the instances for which all annotators provide an answer. This resulted in $n = 100, 190$ for RTE and TEMP, respectively.

In Figure 10 we display $\mathrm{mae}(\theta^{\mathrm{true}}, \hat{\theta}^{\mathrm{mle}})$ for these data sets as function of n for different values of k. These plots generated from real-world data show similar trend to the synthetic experiments. The estimation errors decay to 0 as n increases and generally tend to decrease as k increases. This correspondence is remarkable since two of the labelers have worse than random accuracy and since it is not clear whether the conditional independence assumption actually holds in reality for these data sets. Nevertheless, the collaborative estimation error behaves in accordance with the synthetic data experiments and the theory. This shows that the estimation framework is robust to the breakdown of the assumption that the classifier accuracy must be higher than random choice. Also, whether the conditional independence assumption holds or not is not crucial in this case.

We further experimented with classifiers trained on different representations of the same data set and estimated their error rates. We adopted the Ringnorm data set generated by Breiman (1996). Ringnorm is a 2-class artificial data set with 20 dimensions where each class is drawn from a multivariate normal distribution. One class has zero mean and a covariance $\Sigma = 4I$ where I is the identity matrix. The other class has unit covariance and a mean $\mu = (\frac{2}{\sqrt{20}}, \frac{2}{\sqrt{20}}, \ldots, \frac{2}{\sqrt{20}})$. The total size is 7400. We created 5 different representations of the data by projecting it onto mutually exclusive sets of principal components obtained by Principal Component Analysis (PCA). We trained an SVM classifier (with 2-degree polynomial kernel) (Vapnik, 2000; Joachims, 1999) on samples from each representation while holding out 1400 examples as the test set resulting in a total of 5 classifiers. We tested each of the 5 classifiers on the test set and used their outputs to estimate the corresponding parameters. The true labels of the test set examples were used as ground truth to calculate the mae of the mle estimators.

The mae curves for this data set appear in Figure 11 as a function of the number n of unlabeled examples. When all classifiers are highly accurate (upper left panel), the collaborative unsupervised estimator is reliable, see Figure 11(a). With a mixture of weak and strong classifiers (upper right panel), the collaborative unsupervised estimator is also reliable. This is despite the fact that some of



(a) Strong classifiers

(b) A mixture of strong and weak classifiers

(c) Mostly weak classifiers

(d) Very weak classifiers

Figure 11: $\text{mae}(\theta^{\text{true}}, \hat{\theta}^{\text{mle}})$ as a function of the test set size on the Ringnorm data set. $p(y = 1) = 0.47$, and θ^{true} is indicated in the legend in each plot. The four panels represent mostly strong classifiers (upper left), a mixture of strong and weak classifiers (upper right), mostly weak classifiers (bottom left), and mostly very weak classifiers (bottom right). The figure shows that the framework is robust to occasional deviations from the assumption regarding better than random guess classification accuracy (upper right panel). However, as most of the classifiers become weak or very weak, the collaborative unsupervised estimation framework results in worse estimation error.

the weak classifiers in Figure 11(b) have worse than random accuracy which violates the assumptions in the consistency proposition. This shows again that the estimation framework is robust to occasional deviations from the requirement concerning better than random classification accuracies. On the other hand, as most of the classifiers become worse (bottom row), the accuracy of the unsupervised estimator decreases, in accordance with the theory developed in Sections 5 (recall the Fisher information contour plot).

Our experiments thus far assumed the symmetric noise model (8). Despite it not being always applicable for real world data and classifiers, it did result in good estimation accuracy in some of the cases described thus far. However, in some cases this assumption is grossly violated and the more general noise model is needed (6). For this reason, we conducted two experiments using real world data assuming the more general (6).

The first experiment concerned domain adaptation (Blitzer et al., 2007) for Amazon's product reviews in four different product domains: books, DVDs, electronics and kitchen appliances. Each

	book	dvd	kitchen	electronics	20newsgroup
training error	0.22	0.23	0.26	0.30	0.028
non-collaborative	**0.04**	**0.04**	**0.08**	**0.06**	**0.006**
collaborative	0.10	0.10	0.09	0.08	n/a

Figure 12: mae$(\hat{\theta}^{\text{mle}}, \theta^{\text{true}})$ for the domain adaptation ($n = 1000$, $p(y = 1) = 0.75$) and 20 newsgroup ($n = 15,000$, $p(y = 1) = 0.05$ for each one-vs-all data). The unsupervised non-collaborative estimator outperforms the collaborative estimator due to violation of the conditional independence assumption. Both unsupervised estimators perform substantially better than the baseline training error rate estimator. In both cases the results were averaged over 50 random train test splits.

domain consists of positive ($y = 1$) and negative ($y = 2$) reviews with $p(y = 1) = 0.75$. The task was to estimate the error rates of classifiers (linear SVM, Vapnik, 2000; Joachims, 1999) that are trained on 300 examples from one domain but tested on other domains. The mae values for the classification risks are displayed in Figure 12 with the columns indicating the test domain. In this case, the unsupervised non-collaborative estimator outperforms the collaborative estimator due to violation of the conditional independence assumption. Both unsupervised estimators perform substantially better than the baseline estimator that uses the training error on one domain to predict testing error on another domain.

In the second experiment using (6) we estimated the risk (non-collaboratively) of 20 one vs. all classifiers (trained to predict one class) on the 20 newsgroup data (Lang, 1995). The train set size was 1000 and the unlabeled data size was 15000. In this case the unsupervised non-collaborative estimator returned extremely accurate risk estimators. As a comparison, the risk estimates obtained from the training error are four times larger than the unsupervised MLE estimator (See Figure 12).

8. Discussion

We have demonstrated a collaborative framework for the estimation of classification and regression error rates for $k \geq 1$ predictors. In contrast to previous supervised risk estimation methods such as cross validation (Duda et al., 2001), bootstrap (Efron and Tibshirani, 1997), and others (Hand, 1986), our approach is fully unsupervised and thus able to use vast collections of unlabeled data. Other related work includes Smyth et al. (1995) and Sheng et al. (2008) which consider repeated labeling where each instance is labeled by multiple experts and the final label is decided based on a majority voting scheme. However, Smyth et al. and Sheng et al. fail to address estimating the risks of the predictors which is the main focus of our work.

We prove statistical consistency in the unsupervised case and derive the asymptotic variance. Our experiments on synthetic data demonstrate the effectiveness of the framework and verify the theoretical results. Experiments on real world data show robustness to underlying assumptions. The framework may be applied to estimate additional quantities in an unsupervised manner, including noise level in noisy communication channels (Cover and Thomas, 2005) and error rates in structured prediction problems.

Acknowledgments

The authors thank Jian Zhang and the reviewers for helpful discussions and suggestions. K. Balasubramanian and G. Lebanon were partially supported by NSF grant IIS-0906643.

References

Y. Bishop, S. Fienberg, and P. Holland. *Discrete Multivariate Analysis: Theory and Practice*. MIT press, 1975.

J. Blitzer, M. Dredze, and F. Pereira. Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In *Proc. of ACL '07*, 2007.

L. Breiman. Bias, variance, and arcing classifiers. Technical Report 460, Statistics department, University of California, 1996.

T. M. Cover and J. A. Thomas. *Elements of Information Theory*. John Wiley & Sons, second edition, 2005.

D. Cox, J. Little, and D. O'Shea. *Ideals, Varieties, and Algorithms: An Introduction to Computational Algebraic Geometry and Commutative Algebra*. Springer, 2006.

R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification*. Wiley New York, 2001.

B. Efron and R. J. Tibshirani. *An Introduction to the Bootstrap*. Chapman & Hall, 1997.

T. S. Ferguson. *A Course in Large Sample Theory*. Chapman & Hall, 1996.

D. J. Hand. Recent advances in error rate estimation. *Pattern Recognition Letters*, 4(5):335–346, 1986.

T. Joachims. Making large-scale svm learning practical. In B. Schölkopf, C. Burges, and A. Smola, editors, *Advances in Kernel Methods - Support Vector Learning*. MIT Press, 1999.

K. Lang. Newsweeder: Learning to filter netnews. In *International Conference on Machine Learning*, 1995.

A. Papoulis. *Probability, Random Variables, and Stochastic Processes*. McGraw-Hill, 1984.

V. S. Sheng, F. Provost, and P. G. Ipeirotis. Get another label? improving data quality and data mining using multiple, noisy labelers. In *Proc. of the 14th ACM SIGKDD Internation Conference on Knowledge Discovery and Data Mining*, pages 614–622, 2008.

P. Smyth, U. Fayyad, M. Burl, P. Perona, and P. Baldi. Inferring ground truth from subjective labelling of venus images. In *Advances in Neural Information Processing Systems 7*, 1995.

R. Snow, B. O'Connor, D. Jurafsky, and A. Y. Ng. Cheap and fast-but is it good? evaluating non-expert annotations for natural language tasks. In *Proc. of EMNLP*, 2008.

B. Sturmfels. *Solving Systems of Polynomial Equations*. American Mathematical Society, 2002.

V. N. Vapnik. *The Nature of Statistical Learning Theory*. Springer, second edition, 2000.