Journal of Machine Learning Research 12 (2011) 3119-3145 Submitted 7/10; Revised 04/11; Published

Unsupervised Supervised Learning II: Margin-Based Classification Without Labels

Krishnakumar Balasubramanian KRISHNAKUMAR3@GATECH.EDU
College of Computing
Georgia Institute of Technology
266 Ferst Dr.
Atlanta, GA 30332, USA

Pinar Donmez PINARD@YAHOO-INC.COM
Yahoo! Labs
701 First Ave.
Sunnyale CA 94089, USA

Guy Lebanon LEBANON@CC.GATECH.EDU
College of Computing
Georgia Institute of Technology
266 Ferst Dr.
Atlanta, GA 30332, USA

Editor: Ingo Steinwart

Abstract

Many popular linear classifiers, such as logistic regression, boosting, or SVM, are trained by optimizing a margin-based risk function. Traditionally, these risk functions are computed based on a labeled data set. We develop a novel technique for estimating such risks using only unlabeled data and the marginal label distribution. We prove that the proposed risk estimator is consistent on high-dimensional data sets and demonstrate it on synthetic and real-world data. In particular, we show how the estimate is used for evaluating classifiers in transfer learning, and for training classifiers with no labeled data whatsoever.

Keywords: classification, large margin, maximum likelihood

1. Introduction

Many popular linear classifiers, such as logistic regression, boosting, or SVM, are trained by optimizing a margin-based risk function. For standard linear classifiers $\hat{Y} = \operatorname{sign} \sum \theta_j X_j$ with $Y \in \{-1,+1\}$, and $X, \theta \in \mathbb{R}^d$ the margin is defined as the product

$$Y f_\theta(X) \quad \text{where} \quad f_\theta(X) \stackrel{\text{def}}{=} \sum_{j=1}^{d} \theta_j X_j.$$

Training such classifiers involves choosing a particular value of θ. This is done by minimizing the risk or expected loss

$$R(\theta) = \mathsf{E}_{p(X,Y)} \mathcal{L}(Y, f_\theta(X)) \tag{1}$$

with the three most popular loss functions

$$\mathcal{L}_1(Y, f_\theta(X)) = \exp(-Y f_\theta(X)), \tag{2}$$
$$\mathcal{L}_2(Y, f_\theta(X)) = \log(1+\exp(-Y f_\theta(X))) \text{ and} \tag{3}$$
$$\mathcal{L}_3(Y, f_\theta(X)) = (1-Y f_\theta(X))_+ \tag{4}$$

being exponential loss $\mathcal{L}_1$ (boosting), logloss $\mathcal{L}_2$ (logistic regression) and hinge loss $\mathcal{L}_3$ (SVM) respectively (A_+ above corresponds to A if $A > 0$ and 0 otherwise).

Since the risk $R(\theta)$ depends on the unknown distribution p, it is usually replaced during training with its empirical counterpart

$$R_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} \mathcal{L}(Y^{(i)}, f_\theta(X^{(i)})) \tag{5}$$

based on a labeled training set

$$(X^{(1)},Y^{(1)}),\ldots,(X^{(n)},Y^{(n)}) \overset{\text{iid}}{\sim} p \tag{6}$$

leading to the following estimator

$$\hat{\theta}_n = \arg\min_\theta R_n(\theta).$$

Note, however, that evaluating and minimizing R_n requires labeled data (6). While suitable in some cases, there are certainly situations in which labeled data is difficult or impossible to obtain.

In this paper we construct an estimator for $R(\theta)$ using only unlabeled data, that is using

$$X^{(1)},\ldots,X^{(n)} \overset{\text{iid}}{\sim} p \tag{7}$$

instead of (6). Our estimator is based on the assumption that when the data is high dimensional ($d \to \infty$) the quantities

$$f_\theta(X)|\{Y=y\}, \quad y \in \{-1,+1\} \tag{8}$$

are normally distributed. This phenomenon is supported by empirical evidence and may also be derived using non-iid central limit theorems. We then observe that the limit distributions of (8) may be estimated from unlabeled data (7) and that these distributions may be used to measure margin-based losses such as (2)-(4). We examine two novel unsupervised applications: (i) estimating margin-based losses in transfer learning and (ii) training margin-based classifiers. We investigate these applications theoretically and also provide empirical results on synthetic and real-world data. Our empirical evaluation shows the effectiveness of the proposed framework in risk estimation and classifier training without any labeled data.

The consequences of estimating $R(\theta)$ without labels are indeed profound. Label scarcity is a well known problem which has lead to the emergence of semisupervised learning: learning using a few labeled examples and many unlabeled ones. The techniques we develop lead to a new paradigm that goes beyond semisupervised learning in requiring no labels whatsoever.

2. Unsupervised Risk Estimation

In this section we describe in detail the proposed estimation framework and discuss its theoretical properties. Specifically, we construct an estimator for $R(\theta)$ defined in (1) using the unlabeled data (7) which we denote $\hat{R}_n(\theta; X^{(1)}, \ldots, X^{(n)})$ or simply $\hat{R}_n(\theta)$ (to distinguish it from R_n in (5)).

Our estimation is based on two assumptions. The first assumption is that the label marginals $p(Y)$ are known and that $p(Y=1) \neq p(Y=-1)$. While this assumption may seem restrictive at first, there are many cases where it holds. Examples include medical diagnosis ($p(Y)$ is the well known marginal disease frequency), handwriting recognition or OCR ($p(Y)$ is the easily computable marginal frequencies of different letters in the English language), life expectancy prediction ($p(Y)$ is based on marginal life expectancy tables). In these and other examples $p(Y)$ is known with great accuracy even if labeled data is unavailable. Our experiments show that assuming a wrong marginal $p'(Y)$ causes a graceful performance degradation in $|p(Y) - p'(Y)|$. Furthermore, the assumption of a known $p(Y)$ may be replaced with a weaker form in which we know the ordering of the marginal distributions, for example, $p(Y=1) > p(Y=-1)$, but without knowing the specific values of the marginal distributions.

The second assumption is that the quantity $f_\theta(X)|Y$ follows a normal distribution. As $f_\theta(X)|Y$ is a linear combination of random variables, it is frequently normal when X is high dimensional. From a theoretical perspective this assumption is motivated by the central limit theorem (CLT). The classical CLT states that $f_\theta(X) = \sum_{i=1}^{d} \theta_i X_i | Y$ is approximately normal for large d if the data components $X_1, \ldots, X_d$ are iid given Y. A more general CLT states that $f_\theta(X)|Y$ is asymptotically normal if $X_1, \ldots, X_d | Y$ are independent (but not necessary identically distributed). Even more general CLTs state that $f_\theta(X)|Y$ is asymptotically normal if $X_1, \ldots, X_d | Y$ are not independent but their dependency is limited in some way. We examine this issue in Section 2.1 and also show that the normality assumption holds empirically for several standard data sets.

To derive the estimator we rewrite (1) by taking expectation with respect to Y and $\alpha = f_\theta(X)$

$$R(\theta) = \mathsf{E}_{p(f_\theta(X),Y)} \mathcal{L}(Y, f_\theta(X)) = \sum_{y \in \{-1,+1\}} p(y) \int_{\mathbb{R}} p(f_\theta(X) = \alpha | y) \mathcal{L}(y, \alpha)\, d\alpha. \tag{9}$$

Equation (9) involves three terms $\mathcal{L}(y,\alpha)$, $p(y)$ and $p(f_\theta(X) = \alpha | y)$. The loss function $\mathcal{L}$ is known and poses no difficulty. The second term $p(y)$ is assumed to be known (see discussion above). The third term is assumed to be normal $f_\theta(X) \,|\, \{Y=y\} = \sum_i \theta_i X_i \,|\, \{Y=y\} \sim N(\mu_y, \sigma_y)$ with parameters μ_y, σ_y, $y \in \{-1,1\}$ that are estimated by maximizing the likelihood of a Gaussian mixture model (we denote $\mu = (\mu_1, \mu_{-1})$ and $\sigma^2 = (\sigma_1^2, \sigma_{-1}^2)$. These estimated parameters are used to construct the plug-in estimator $\hat{R}_n(\theta)$ as follows:

$$\ell_n(\mu, \sigma) = \sum_{i=1}^{n} \log \sum_{y^{(i)} \in \{-1,+1\}} p(y^{(i)}) p_{\mu_y, \sigma_y}(f_\theta(X^{(i)}) | y^{(i)}). \tag{10}$$

$$(\hat{\mu}^{(n)}, \hat{\sigma}^{(n)}) = \arg\max_{\mu, \sigma} \ell_n(\mu, \sigma). \tag{11}$$

$$\hat{R}_n(\theta) = \sum_{y \in \{-1,+1\}} p(y) \int_{\mathbb{R}} p_{\hat{\mu}_y^{(n)}, \hat{\sigma}_y^{(n)}}(f_\theta(X) = \alpha | y) \mathcal{L}(y, \alpha)\, d\alpha. \tag{12}$$

We make the following observations.

1. Although we do not denote it explicitly, μ_y and σ_y are functions of θ.

2. The loglikelihood (10) does not use labeled data (it marginalizes over the label $y^{(i)}$).

3. The parameters of the loglikelihood (10) are $\mu = (\mu_1, \mu_{-1})$ and $\sigma = (\sigma_1, \sigma_{-1})$ rather than the parameter θ associated with the margin-based classifier. We consider the latter one as a fixed constant at this point.

4. The estimation problem (11) is equivalent to the problem of maximum likelihood for means and variances of a Gaussian mixture model where the label marginals are assumed to be known. It is well known that in this case (barring the symmetric case of a uniform $p(y)$) the MLE converges to the true parameter values (Teicher, 1963).

5. The estimator $\hat{R}_n$ (12) is consistent in the limit of infinite unlabeled data

$$P\left(\lim_{n\to\infty} \hat{R}_n(\theta) = R(\theta)\right) = 1.$$

6. The two risk estimators $\hat{R}_n(\theta)$ (12) and $R_n(\theta)$ (5) approximate the expected loss $R(\theta)$. The latter uses labeled samples and is typically more accurate than the former for a fixed n.

7. Under suitable conditions $\arg\min_\theta \hat{R}_n(\theta)$ converges to the expected risk minimizer

$$P\left(\lim_{n\to\infty} \arg\min_{\theta\in\Theta} \hat{R}_n(\theta) = \arg\min_{\theta\in\Theta} R(\theta)\right) = 1.$$

   This far reaching conclusion implies that in cases where $\arg\min_\theta R(\theta)$ is the Bayes classifier (as is the case with exponential loss, log loss, and hinge loss) we can retrieve the optimal classifier without a single labeled data point.

2.1 Asymptotic Normality of $f_\theta(X)|Y$

The quantity $f_\theta(X)|Y$ is essentially a sum of d random variables which under some conditions for large d is likely to be normally distributed. One way to verify this is empirically, as we show in Figures 1-3 which contrast the histogram with a fitted normal pdf for text, digit images, and face images data. For these data sets the dimensionality d is sufficiently high to provide a nearly normal $f_\theta(X)|Y$. For example, in the case of text documents (X_i is the relative number of times word i appeared in the document) d corresponds to the vocabulary size which is typically a large number in the range $10^3 - 10^5$. Similarly, in the case of image classification (X_i denotes the brightness of the i-pixel) the dimensionality is on the order of $10^2 - 10^4$.

Figures 1-3 show that in these cases of text and image data $f_\theta(X)|Y$ is approximately normal for both randomly drawn θ vectors (Figure 1) and for θ representing estimated classifiers (Figures 2 and 3). A caveat in this case is that normality may not hold when θ is sparse, as may happen for example for L_1 regularized models (last row of Figure 2).

From a theoretical standpoint normality may be argued using a central limit theorem. We examine below several progressively more general central limit theorems and discuss whether these theorems are likely to hold in practice for high dimensional data. The original central limit theorem states that $\sum_{i=1}^d Z_i$ is approximately normal for large d if Z_i are iid.

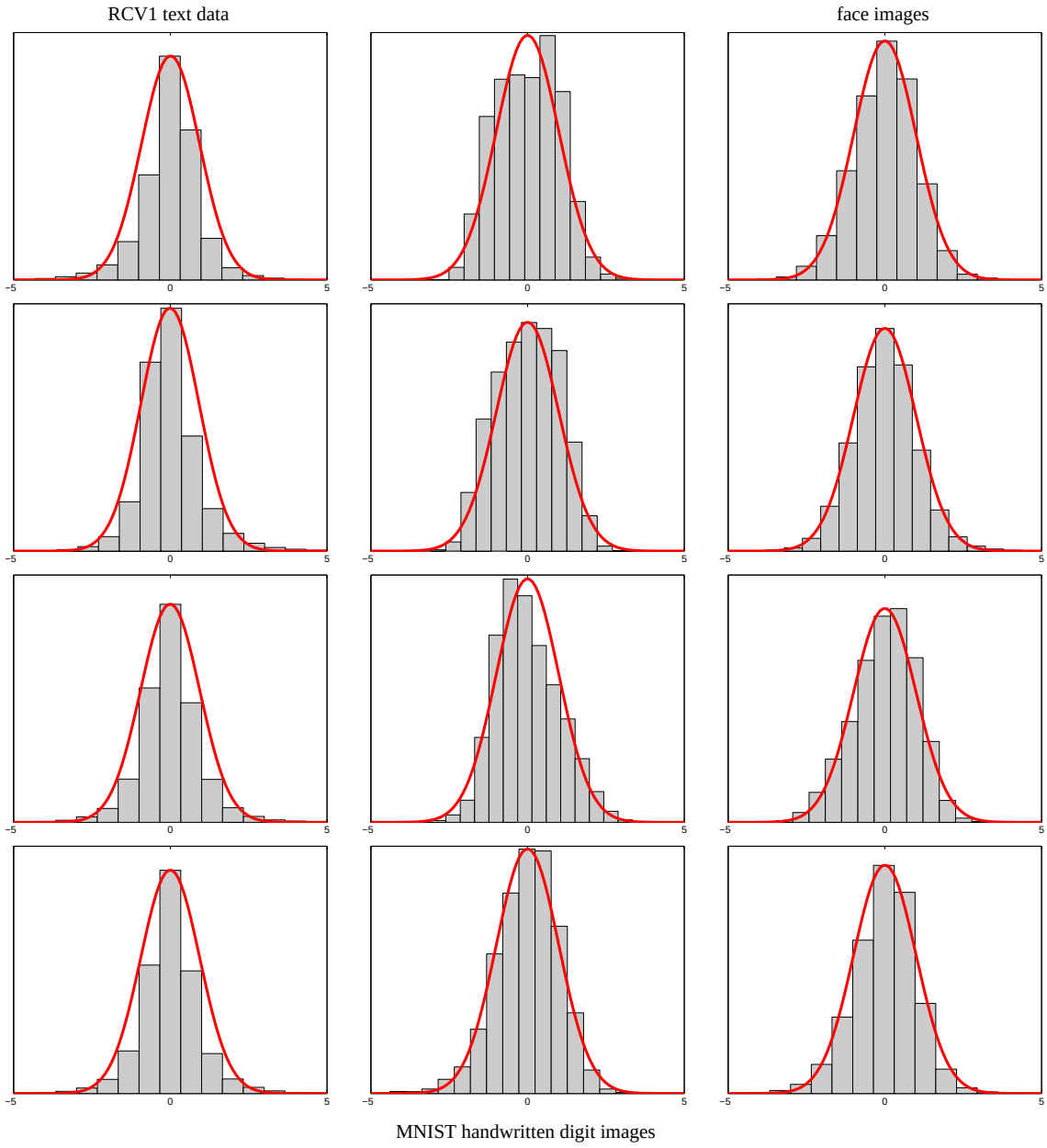

Figure 1: Centered histograms of $f_\theta(X)|\{Y=1\}$ overlayed with the pdf of a fitted Gaussian for randomly drawn θ vectors ($\theta_i \sim U(-1/2, 1/2)$). The columns represent data sets (RCV1 text data, Lewis et al., 2004, MNIST digit images, and face images, Pham et al., 2002) and the rows represent multiple random draws. For uniformity we subtracted the empirical mean and divided by the empirical standard deviation. The twelve panels show that even in moderate dimensionality (RCV1: 1000 top words, MNIST digits: 784 pixels, face images: 400 pixels) the assumption that $f_\theta(X)|Y$ is normal holds often for randomly drawn θ.

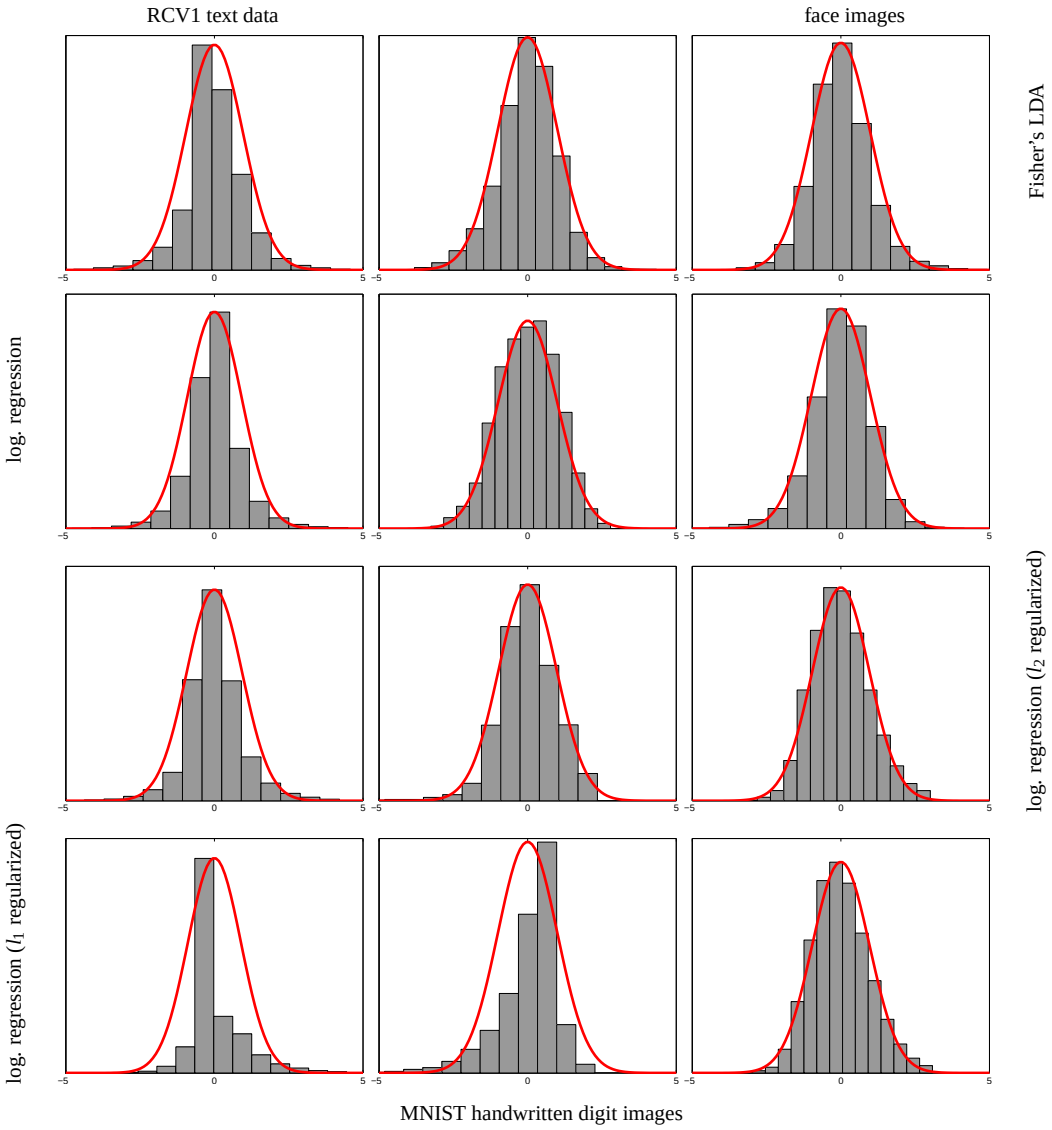

Figure 2: Centered histograms of $f_\theta(X)|\{Y=1\}$ overlayed with the pdf of a fitted Gaussian for multiple θ vectors (four rows: Fisher's LDA, logistic regression, l_2 regularized logistic regression, and l_1 regularized logistic regression-all regularization parameters were selected by cross validation) and data sets (columns: RCV1 text data, Lewis et al., 2004, MNIST digit images, and face images, Pham et al., 2002). For uniformity we subtracted the empirical mean and divided by the empirical standard deviation. The twelve panels show that even in moderate dimensionality (RCV1: 1000 top words, MNIST digits: 784 pixels, face images: 400 pixels) the assumption that $f_\theta(X)|Y$ is normal holds well for fitted θ values (except perhaps for L_1 regularization in the last row which promotes sparse θ).

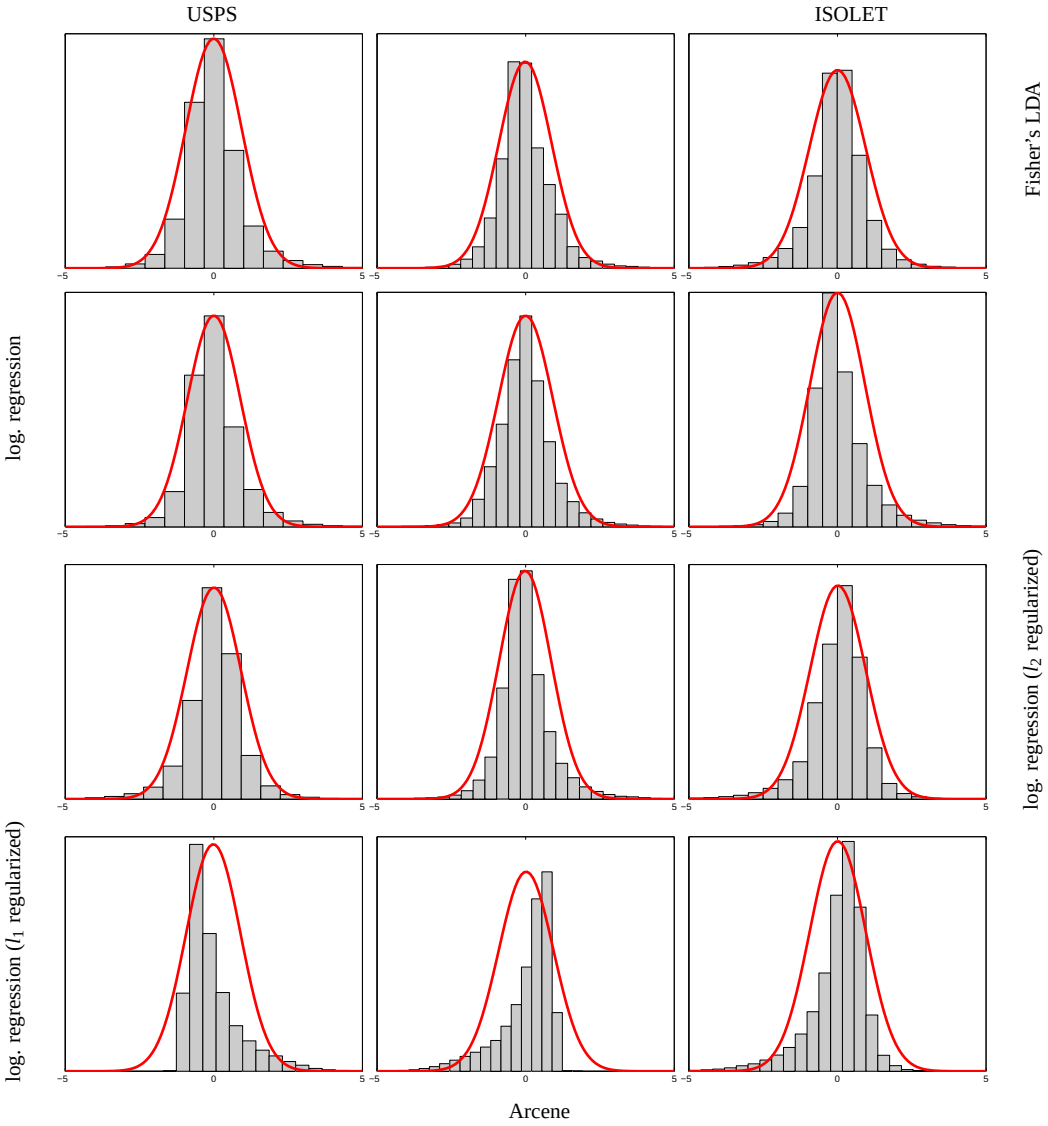

Figure 3: Centered histograms of $f_\theta(X)|\{Y=1\}$ overlayed with the pdf of a fitted Gaussian for multiple θ vectors (four rows: Fisher's LDA, logistic regression, l_2 regularized logistic regression, and l_1 regularized logistic regression-all regularization parameters were selected by cross validation) and data sets (columns: USPS Handwritten Digits, Arcene data set, and ISOLET). For uniformity we subtracted the empirical mean and divided by the empirical standard deviation. The twelve panels further confirm that the assumption that $f_\theta(X)|Y$ is normal holds well for fitted θ values (except perhaps for L_1 regularization in the last row which promotes sparse θ) for various data sets.

Proposition 1 (de-Moivre) *If $Z_i, i \in \mathbb{N}$ are iid with expectation μ and variance σ^2 and $\overline{Z}_d = d^{-1}\sum_{i=1}^d Z_i$ then we have the following convergence in distribution*

$$\sqrt{d}(\overline{Z}_d - \mu)/\sigma \rightsquigarrow N(0,1) \quad \text{as } d \to \infty.$$

As a result, the quantity $\sum_{i=1}^{d} Z_i$ (which is a linear transformation of $\sqrt{d}(\bar{Z}_d - \mu)/\sigma$) is approximately normal for large d. This relatively restricted theorem is unlikely to hold in most practical cases as the data dimensions are often not iid.

A more general CLT does not require the summands Z_i to be identically distributed.

Proposition 2 (Lindberg) *For $Z_i, i \in \mathbb{N}$ independent with expectation μ_i and variance σ_i^2, and denoting $s_d^2 = \sum_{i=1}^{d} \sigma_i^2$, we have the following convergence in distribution as $d \to \infty$*

$$s_d^{-1} \sum_{i=1}^{d} (Z_i - \mu_i) \rightsquigarrow N(0,1)$$

if the following condition holds for every $\varepsilon > 0$

$$\lim_{d\to\infty} s_d^{-2} \sum_{i=1}^{d} E\,(Z_i - \mu_i)^2 1_{\{|X_i - \mu_i| > \varepsilon s_d\}} = 0. \tag{13}$$

This CLT is more general as it only requires that the data dimensions be independent. The condition (13) is relatively mild and specifies that contributions of each of the Z_i to the variance s_d should not dominate it. Nevertheless, the Lindberg CLT is still inapplicable for dependent data dimensions.

More general CLTs replace the condition that $Z_i, i \in \mathbb{N}$ be independent with the notion of $m(k)$-dependence.

Definition 3 *The random variables $Z_i, i \in \mathbb{N}$ are said to be $m(k)$-dependent if whenever $s - r > m(k)$ the two sets $\{Z_1, \ldots, Z_r\}$, $\{Z_s, \ldots, Z_k\}$ are independent.*

An early CLT for $m(k)$-dependent RVs was provided by Hoeffding and Robbins (1948). Below is a slightly weakened version of the CLT, as proved in Berk (1973).

Proposition 4 (Berk) *For each $k \in \mathbb{N}$ let $d(k)$ and $m(k)$ be increasing sequences and suppose that $Z_1^{(k)}, \ldots, Z_{d(k)}^{(k)}$ is an $m(k)$-dependent sequence of random variables. If*

1. *$E|Z_i^{(k)}|^2 \le M$ for all i and k,*
2. *$Var(Z_{i+1}^{(k)} + \ldots + Z_j^{(k)}) \le (j-i)K$ for all i, j, k,*
3. *$\lim_{k\to\infty} Var(Z_1^{(k)} + \ldots + Z_{d(k)}^{(k)})/d(k)$ exists and is non-zero, and*
4. *$\lim_{k\to\infty} m^2(k)/d(k) = 0$*

then $\frac{\sum_{i=1}^{d(k)} Z_i^{(k)}}{\sqrt{d(k)}}$ is asymptotically normal as $k \to \infty$.

Proposition 4 states that under mild conditions the sum of $m(k)$-dependent RVs is asymptotically normal. If $m(k)$ is a constant, that is, $m(k) = m$, $m(k)$-dependence implies that a Z_i may only depend on its neighboring dimensions (in the sense of Definition 3). Intuitively, dimensions whose indices are far removed from each other are independent. The full power of Proposition 4 is invoked when $m(k)$ grows with k relaxing the independence restriction as the dimensionality grows. Intuitively, the dependency of the summands is not fixed to a certain order, but it cannot grow too rapidly.

A more realistic variation of $m(k)$ dependence where the dependency of each variable is specified using a dependency graph (rather than each dimension depends on neighboring dimensions) is advocated in a number of papers, including the following recent result by Rinott (1994).

Definition 5 *A graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ indexing random variables is called a dependency graph if for any pair of disjoint subsets of $\mathcal{V}$, A_1 and A_2 such that no edge in $\mathcal{E}$ has one endpoint in A_1 and the other in A_2, we have independence between $\{Z_i : i \in A_1\}$ and $\{Z_i : i \in A_2\}$. The degree $d(v)$ of a vertex is the number of edges connected to it and the maximal degree is* $\max_{v \in \mathcal{V}} d(v)$.

Proposition 6 (Rinott) *Let $Z_1, \ldots, Z_n$ be random variables having a dependency graph whose maximal degree is strictly less than D, satisfying $|Z_i - \mathsf{E} Z_i| \leq B$ a.s., $\forall i$, $\mathsf{E}(\sum_{i=1}^n Z_i) = \lambda$ and $\mathsf{Var}(\sum_{i=1}^n Z_i) = \sigma^2 > 0$, Then for any $w \in \mathbb{R}$,*

$$\left| P\left(\frac{\sum_{i=1}^n Z_i - \lambda}{\sigma} \leq w \right) - \Phi(w) \right| \leq \frac{1}{\sigma} \left(\frac{1}{\sqrt{2\pi}} DB + 16 \left(\frac{n}{\sigma^2} \right)^{1/2} D^{3/2} B^2 + 10 \left(\frac{n}{\sigma^2} \right) D^2 B^3 \right)$$

where $\Phi(w)$ is the CDF corresponding to a N(0,1) distribution.

The above theorem states a stronger result than convergence in distribution to a Gaussian in that it states a uniform rate of convergence of the CDF. Such results are known in the literature as Berry Essen bounds (Davidson, 1994). When D and B are bounded and $\mathsf{Var}(\sum_{i=1}^n Z_i) = O(n)$ it yields a CLT with an optimal convergence rate of $n^{-1/2}$.

The question of whether the above CLTs apply in practice is a delicate one. For text one can argue that the appearance of a word depends on some words but is independent of other words. Similarly for images it is plausible to say that the brightness of a pixel is independent of pixels that are spatially far removed from it. In practice one needs to verify the normality assumption empirically, which is simple to do by comparing the empirical histogram of $f_\theta(X)$ with that of a fitted mixture of Gaussians. As the figures above indicate this holds for text and image data for some values of θ, assuming it is not sparse. Also, it is worth mentioning that one dimensional CLTs kick in relatively early perhaps at 50 or 100 dimensions. Even when the high dimensional data lie on a lower dimensional manifold whose dimensionality is on the order of 100 dimensions, the CLT still applies to some extent (see histogram plots).

2.2 Unsupervised Consistency of $\hat{R}_n(\theta)$

We start with proving identifiability of the maximum likelihood estimator (MLE) for a mixture of two Gaussians with known ordering of mixture proportions. Invoking classical consistency results in conjunction with identifiability we show consistency of the MLE estimator for (μ, σ) parameterizing the distribution of $f_\theta(X)|Y$. As a result consistency of the estimator $\hat{R}_n(\theta)$ follows.

Definition 7 *A parametric family $\{p_\alpha : \alpha \in A\}$ is identifiable when $p_\alpha(x) = p_{\alpha'}(x), \forall x$ implies $\alpha = \alpha'$.*

Proposition 8 *Assuming known label marginals with $p(Y = 1) \neq p(Y = -1)$ the Gaussian mixture family*

$$p_{\mu,\sigma}(x) = p(y = 1) N(x; \mu_1, \sigma_1^2) + p(y = -1) N(x; \mu_{-1}, \sigma_{-1}^2)$$

is identifiable.

Proof It can be shown that the family of Gaussian mixture model with no apriori information about label marginals is identifiable up to a permutation of the labels y (Teicher, 1963). We proceed by

assuming with no loss of generality that $p(y=1) > p(y=-1)$. The alternative case $p(y=1) < p(y=-1)$ may be handled in the same manner. Using the result of Teicher (1963) we have that if $p_{\mu,\sigma}(x) = p_{\mu',\sigma'}(x)$ for all x, then $(p(y),\mu,\sigma) = (p(y),\mu',\sigma')$ up to a permutation of the labels. Since permuting the labels violates our assumption $p(y=1) > p(y=-1)$ we establish $(\mu,\sigma) = (\mu',\sigma')$ proving identifiability. ■

The assumption that $p(y)$ is known is not entirely crucial. It may be relaxed by assuming that it is known whether $p(Y=1) > p(Y=-1)$ or $p(Y=1) < p(Y=-1)$. Proving Proposition 8 under this much weaker assumption follows identical lines.

Proposition 9 *Under the assumptions of Proposition 8 the MLE estimates for* $(\mu,\sigma) = (\mu_1,\mu_{-1},\sigma_1,\sigma_{-1})$

$$(\hat{\mu}^{(n)},\hat{\sigma}^{(n)}) = \arg\max_{\mu,\sigma} \ell_n(\mu,\sigma),$$

$$\ell_n(\mu,\sigma) = \sum_{i=1}^{n} \log \sum_{y^{(i)}\in\{-1,+1\}} p(y^{(i)}) p_{\mu_y,\sigma_y}(f_\theta(X^{(i)})|y^{(i)}).$$

are consistent, that is, $(\hat{\mu}_1^{(n)},\hat{\mu}_{-1}^{(n)},\hat{\sigma}_1^{(n)},\hat{\sigma}_{-1}^{(n)})$ *converge as* $n\to\infty$ *to the true parameter values with probability 1.*

Proof Denoting $p_\eta(z) = \sum_y p(y) p_{\mu_y,\sigma_y}(z|y)$ with $\eta = (\mu,\sigma)$ we note that p_η is identifiable (see Proposition 8) in η and the available samples $z^{(i)} = f_\theta(X^{(i)})$ are iid samples from $p_\eta(z)$. We therefore use standard statistics theory which indicates that the MLE for identifiable parametric model is strongly consistent (Ferguson, 1996, Chapter 17). ■

Proposition 10 *Under the assumptions of Proposition 8 and assuming the loss* $\mathcal{L}$ *is given by one of* (2)-(4) *with a normal* $f_\theta(X)|Y \sim N(\mu_y,\sigma_y^2)$*, the plug-in risk estimate*

$$\hat{R}_n(\theta) = \sum_{y\in\{-1,+1\}} p(y) \int_{\mathbb{R}} p_{\hat{\mu}_y^{(n)},\hat{\sigma}_y^{(n)}}(f_\theta(X)=\alpha|y)\mathcal{L}(y,\alpha)\,d\alpha. \tag{14}$$

is consistent, that is, for all θ*,*

$$P\left(\lim_n \hat{R}_n(\theta) = R(\theta)\right) = 1.$$

Proof The plug-in risk estimate $\hat{R}_n$ in (14) is a continuous function (when L is given by (2), (3) or (4)) of $\hat{\mu}_1^{(n)},\hat{\mu}_{-1}^{(n)},\hat{\sigma}_1^{(n)},\hat{\sigma}_{-1}^{(n)}$ (note that μ_y and σ_y are functions of θ), which we denote $\hat{R}_n(\theta) = h(\hat{\mu}_1^{(n)},\hat{\mu}_{-1}^{(n)},\hat{\sigma}_1^{(n)},\hat{\sigma}_{-1}^{(n)})$.

Using Proposition 9 we have that

$$\lim_{n\to\infty}(\hat{\mu}_1^{(n)},\hat{\mu}_{-1}^{(n)},\hat{\sigma}_1^{(n)},\hat{\sigma}_{-1}^{(n)}) = (\mu_1^{\text{true}},\mu_{-1}^{\text{true}},\sigma_1^{\text{true}},\sigma_{-1}^{\text{true}})$$

with probability 1. Since continuous functions preserve limits we have

$$\lim_{n\to\infty} h(\hat{\mu}_1^{(n)},\hat{\mu}_{-1}^{(n)},\hat{\sigma}_1^{(n)},\hat{\sigma}_{-1}^{(n)}) = h(\mu_1^{\text{true}},\mu_{-1}^{\text{true}},\sigma_1^{\text{true}},\sigma_{-1}^{\text{true}})$$

with probability 1 which implies convergence $\lim_{n\to\infty}\hat{R}_n(\theta) = R(\theta)$ with probability 1. ■

2.3 Unsupervised Consistency of $\arg\min \hat{R}_n(\theta)$

The convergence above $\hat{R}_n(\theta) \to R(\theta)$ is pointwise in θ. If the stronger concept of uniform convergence is assumed over $\theta \in \Theta$ we obtain consistency of $\arg\min_\theta \hat{R}_n(\theta)$. This surprising result indicates that in some cases it is possible to retrieve the expected risk minimizer (and therefore the Bayes classifier in the case of the hinge loss, log-loss and exp-loss) using only unlabeled data. We show this uniform convergence using a modification of Wald's classical MLE consistency result (Ferguson, 1996, Chapter 17).

Denoting

$$p_\eta(z) = \sum_{y\in\{-1,+1\}} p(y) p_{\mu_y,\sigma_y}(f(X) = z|y), \quad \eta = (\mu_1, \mu_{-1}, \sigma_1, \sigma_{-1})$$

we first show that the MLE converges to the true parameter value $\hat{\eta}_n \to \eta_0$ uniformly. Uniform convergence of the risk estimator $\hat{R}_n(\theta)$ follows. Since changing $\theta \in \Theta$ results in a different $\eta \in E$ we can state the uniform convergence in $\theta \in \Theta$ or alternatively in $\eta \in E$.

Proposition 11 *Let θ take values in Θ for which $\eta \in E$ for some compact set E. Then assuming the conditions in Proposition 10 the convergence of the MLE to the true value $\hat{\eta}_n \to \eta_0$ is uniform in $\eta_0 \in E$ (or alternatively $\theta \in \Theta$).*

Proof We start by making the following notation

$$\begin{aligned} U(z,\eta,\eta_0) &= \log p_\eta(z) - \log p_{\eta_0}(z), \\ \alpha(\eta,\eta_0) &= E_{p_{\eta_0}} U(z,\eta,\eta_0) = -D(p_{\eta_0}, p_\eta) \le 0 \end{aligned}$$

with the latter quantity being non-positive and 0 iff $\eta = \eta_0$ (due to Shannon's inequality and identifiability of p_η).

For $\rho > 0$ we define the compact set $S_{\eta_0,\rho} = \{\eta \in E : \|\eta - \eta_0\| \ge \rho\}$. Since $\alpha(\eta,\eta_0)$ is continuous it achieves its maximum (with respect to η) on $S_{\eta_0,\rho}$ denoted by $\delta_\rho(\eta_0) = \max_{\eta\in S_{\eta_0,\rho}} \alpha(\eta,\eta_0) < 0$ which is negative since $\alpha(\eta,\eta_0) = 0$ iff $\eta = \eta_0$. Furthermore, note that $\delta_\rho(\eta_0)$ is itself continuous in $\eta_0 \in E$ and since E is compact it achieves its maximum

$$\delta = \max_{\eta_0\in E} \delta_\rho(\eta_0) = \max_{\eta_0\in E} \max_{\eta\in S_{\eta_0,\rho}} \alpha(\eta,\eta_0) < 0$$

which is negative for the same reason.

Invoking the uniform strong law of large numbers (Ferguson, 1996, Chapter 16) we have $n^{-1}\sum_{i=1}^n U(z^{(i)},\eta,\eta_0) \to \alpha(\eta,\eta_0)$ uniformly over $(\eta,\eta_0) \in E^2$. Consequentially, there exists N such that for $n > N$ (with probability 1)

$$\sup_{\eta_0\in E} \sup_{\eta\in S_{\eta_0,\rho}} \frac{1}{n}\sum_{i=1}^n U(z^{(i)},\eta,\eta_0) < \delta/2 < 0.$$

But since $n^{-1}\sum_{i=1}^n U(z^{(i)},\eta,\eta_0) \to 0$ for $\eta = \eta_0$ it follows that the MLE

$$\hat{\eta}_n = \max_{\eta\in E} \frac{1}{n}\sum_{i=1}^n U(z^{(i)},\eta,\eta_0)$$

is outside $S_{\eta_0,\rho}$ (for $n > N$ uniformly in $\eta_0 \in E$) which implies $\|\hat{\eta}_n - \eta_0\| \le \rho$. Since $\rho > 0$ is arbitrarily and N does not depend on η_0 we have $\hat{\eta}_n \to \eta_0$ uniformly over $\eta_0 \in E$. ■

Proposition 12 *Assuming that X, Θ are bounded in addition to the assumptions of Proposition 11 the convergence $\hat{R}_n(\theta) \to R(\theta)$ is uniform in $\theta \in \Theta$.*

Proof Since X, Θ are bounded the margin value $f_\theta(X)$ is bounded with probability 1. As a result the loss function is bounded in absolute value by a constant C. We also note that a mixture of two Gaussian model (with known mixing proportions) is Lipschitz continuous in its parameters

$$\left| \sum_{y\in\{-1,+1\}} p(y) p_{\hat{\mu}_y^{(n)},\hat{\sigma}_y^{(n)}}(z) - \sum_{y\in\{-1,+1\}} p(y) p_{\mu_y^{true},\sigma_y^{true}}(z) \right| \leq t(z) \cdot \left|\left| (\hat{\mu}_1^{(n)}, \hat{\mu}_{-1}^{(n)}, \hat{\sigma}_1^{(n)}, \hat{\sigma}_{-1}^{(n)}) - (\mu_1^{\text{true}}, \mu_{-1}^{\text{true}}, \sigma_1^{\text{true}}, \sigma_{-1}^{\text{true}}) \right|\right|$$

which may be verified by noting that the partial derivatives of $p_\eta(z) = \sum_y p(y) p_{\mu_y,\sigma_y}(z|y)$

$$\frac{\partial p_\eta(z)}{\partial \hat{\mu}_1^{(n)}} = \frac{p(y=1)(z-\hat{\mu}_1^{(n)})}{(2\pi)^{1/2}\hat{\sigma}_1^{(n)^3}} e^{-\frac{(z-\hat{\mu}_1^{(n)})^2}{2\hat{\sigma}_1^{(n)^3}}},$$

$$\frac{\partial p_\eta(z)}{\partial \hat{\mu}_{-1}^{(n)}} = \frac{p(y=-1)(z-\hat{\mu}_{-1}^{(n)})}{(2\pi)^{1/2}\hat{\sigma}_{-1}^{(n)^3}} e^{-\frac{(z-\hat{\mu}_{-1}^{(n)})^2}{2\hat{\sigma}_{-1}^{(n)^3}}},$$

$$\frac{\partial p_\eta(z)}{\partial \hat{\sigma}_1^{(n)}} = -\frac{p(y=1)(z-\hat{\mu}_1^{(n)})^2}{(2\pi)^{3/2}\hat{\sigma}_1^{(n)^6}} e^{-\frac{(z-\hat{\mu}_1^{(n)})^2}{2\hat{\sigma}_1^{(n)^2}}},$$

$$\frac{\partial p_\eta(z)}{\partial \hat{\sigma}_{-1}^{(n)}} = -\frac{p(y=-1)(z-\hat{\mu}_{-1}^{(n)})^2}{(2\pi)^{3/2}\hat{\sigma}_{-1}^{(n)^6}} e^{-\frac{(z-\hat{\mu}_{-1}^{(n)})^2}{2\hat{\sigma}_{-1}^{(n)^2}}}$$

are bounded for a compact E. These observations, together with Proposition 11 lead to

$$\begin{aligned}
|\hat{R}_n(\theta) - R(\theta)| &\leq \sum_{y\in\{-1,+1\}} p(y) \int \left| p_{\hat{\mu}_y^{(n)},\hat{\sigma}_y^{(n)}}(f_\theta(X)=\alpha) - p_{\mu_y^{\text{true}},\sigma_y^{\text{true}}}(f_\theta(X)=\alpha) \right| |\mathcal{L}(y,\alpha)| d\alpha \\
&\leq C \int \left| \sum_{y\in\{-1,+1\}} p(y) p_{\hat{\mu}_y^{(n)},\hat{\sigma}_y^{(n)}}(\alpha) - \sum_{y\in\{-1,+1\}} p(y) p_{\mu_y^{\text{true}},\sigma_y^{\text{true}}}(\alpha) \right| d\alpha \\
&\leq C \left\| (\hat{\mu}_1^{(n)}, \hat{\mu}_{-1}^{(n)}, \hat{\sigma}_1^{(n)}, \hat{\sigma}_{-1}^{(n)}) - (\mu_1^{\text{true}}, \mu_{-1}^{\text{true}}, \sigma_1^{\text{true}}, \sigma_{-1}^{\text{true}}) \right\| \int_a^b t(z) dz \\
&\leq C' \left\| (\hat{\mu}_1^{(n)}, \hat{\mu}_{-1}^{(n)}, \hat{\sigma}_1^{(n)}, \hat{\sigma}_{-1}^{(n)}) - (\mu_1^{\text{true}}, \mu_{-1}^{\text{true}}, \sigma_1^{\text{true}}, \sigma_{-1}^{\text{true}}) \right\| \to 0
\end{aligned}$$

uniformly over $\theta \in \Theta$. ■

Proposition 13 *Under the assumptions of Proposition 12*

$$P\left(\lim_{n\to\infty} \arg\min_{\theta\in\Theta} \hat{R}_n(\theta) = \arg\min_{\theta\in\Theta} R(\theta) \right) = 1.$$

Proof We denote $t^* = \arg\min R(\theta)$, $t_n = \arg\min \hat{R}_n(\theta)$. Since $\hat{R}_n(\theta) \to R(\theta)$ uniformly, for each $\varepsilon > 0$ there exists N such that for all $n > N$, $|\hat{R}_n(\theta) - R(\theta)| < \varepsilon$.

Let $S = \{\theta : \|\theta - t^*\| \geq \varepsilon\}$ and $\min_{\theta \in S} R(\theta) > R(t^*)$ (S is compact and thus R achieves its minimum on it). There exists N' such that for all $n > N'$ and $\theta \in S$, $\hat{R}_n(\theta) \geq R(t^*) + \varepsilon$. On the other hand, $\hat{R}_n(t^*) \to R(t^*)$ which together with the previous statement implies that there exists N'' such that for $n > N''$, $\hat{R}_n(t^*) < \hat{R}_n(\theta)$ for all $\theta \in S$. We thus conclude that for $n > N''$, $t_n \notin S$. Since we showed that for each $\varepsilon > 0$ there exists N such that for all $n > N$ we have $\|t_n - t^*\| \leq \varepsilon$, $t_n \to t^*$ which concludes the proof. ■

2.4 Asymptotic Variance

In addition to consistency, it is useful to characterize the accuracy of our estimator $\hat{R}_n(\theta)$ as a function of $p(y), \mu, \sigma$. We do so by computing the asymptotic variance of the estimator which equals the inverse Fisher information

$$\sqrt{n}(\hat{\eta}_n^{\text{mle}} - \eta_0) \rightsquigarrow N(0, I^{-1}(\eta^{\text{true}}))$$

and analyzing its dependency on the model parameters. We first derive the asymptotic variance of MLE for mixture of Gaussians (we denote below $\eta = (\eta_1, \eta_2), \eta_i = (\mu_i, \sigma_i)$)

$$\begin{aligned} p_\eta(z) &= p(Y=1)N(z;\mu_1,\sigma_1^2) + p(Y=-1)N(z;\mu_{-1},\sigma_{-1}^2) \\ &= p_1 p_{\eta_1}(z) + p_{-1} p_{\eta_{-1}}(z). \end{aligned}$$

The elements of 4×4 information matrix $I(\eta)$

$$I(\eta_i, \eta_j) = \mathsf{E}\left(\frac{\partial \log p_\eta(z)}{\partial \eta_i} \frac{\partial \log p_\eta(z)}{\partial \eta_j}\right)$$

may be computed using the following derivatives

$$\begin{aligned} \frac{\partial \log p_\eta(z)}{\partial \mu_i} &= \frac{p_i}{\sigma_i}\left(\frac{z-\mu_i}{\sigma_i}\right)\frac{p_{\eta_i}(z)}{p_\eta(z)}, \\ \frac{\partial \log p_\eta(z)}{\partial \sigma_i^2} &= \frac{p_i}{2\sigma_i}\left(\left(\frac{z-\mu_i}{\sigma_i}\right)^2 - 1\right)\frac{p_{\eta_i}(z)}{p_\eta(z)} \end{aligned}$$

for $i = 1, -1$. Using the method of Behboodian (1972) we obtain

$$\begin{aligned} I(\mu_i,\mu_j) &= \frac{p_i p_j}{\sigma_i \sigma_j} M_{11}\Big(p_{\eta_i}(z), p_{\eta_i}(z)\Big), \\ I(\mu_1,\sigma_i^2) &= \frac{p_1 p_i}{2\sigma_1\sigma_i^2}\Big[M_{12}\Big(p_{\eta_i}(z), p_{\eta_i}(z)\Big) - M_{10}\Big(p_{\eta_1}(z), p_{\eta_i}(z)\Big)\Big], \\ I(\mu_{-1},\sigma_i^2) &= \frac{p_{-1} p_i}{2\sigma_{-1}\sigma_i^2}\Big[M_{21}\Big(p_{\eta_i}(z), p_{\eta_{-1}}(z)\Big) - M_{01}\Big(p_{\eta_i}(z), p_{\eta_{-1}}(z)\Big)\Big], \\ I(\sigma_i^2,\sigma_i^2) &= \frac{p_i^4}{4\sigma_i^4}\Big[M_{00}\Big(p_{\eta_i}(z), p_{\eta_i}(z)\Big) - 2M_{11}\Big(p_{\eta_i}(z), p_{\eta_i}(z)\Big) + M_{22}\Big(p_{\eta_i}(z), p_{\eta_i}(z)\Big)\Big], \\ I(\sigma_1^2,\sigma_{-1}^2) &= \frac{p_1 p_{-1}}{4\sigma_1^2\sigma_{-1}^2}\Big[M_{00}\Big(p_{\eta_1}(z), p_{\eta_{-1}}(z)\Big) - M_{20}\Big(p_{\eta_1}(z), p_{\eta_{-1}}(z)\Big) \\ &\qquad - M_{02}\Big(p_{\eta_1}(z), p_{\eta_{-1}}(z)\Big) + M_{22}\Big(p_{\eta_1}(z), p_{\eta_{-1}}(z)\Big)\Big] \end{aligned}$$

where

$$M_{m,n}\Big(p_{\eta_i}(z), p_{\eta_j}(z)\Big) = \int_{-\infty}^{\infty} \left(\frac{z-\mu_i}{\sigma_i}\right)^m \left(\frac{z-\mu_j}{\sigma_j}\right)^n \frac{p_{\eta_i}(z)p_{\eta_j}(z)}{p_\eta(z)}\,dx.$$

In some cases it is more instructive to consider the asymptotic variance of the risk estimator $\hat{R}_n(\theta)$ rather than that of the parameter estimate for $\eta = (\mu, \sigma)$. This could be computed using the delta method and the above Fisher information matrix

$$\sqrt{n}(\hat{R}_n(\theta) - R(\theta)) \rightsquigarrow N(0, \nabla h(\eta^{\text{true}})^T I^{-1}(\eta^{true})\nabla h(\eta^{\text{true}}))$$

where ∇h is the gradient vector of the mapping $R(\theta) = h(\eta)$. For example, in the case of the exponential loss (2) we get

$$\begin{aligned}
h(\eta) &= p(Y=1)\sigma_1\sqrt{2}\exp\Big(\frac{(\mu_1-1)^2}{2} - \frac{\mu_1^2}{2\sigma_1^2}\Big) + p(Y=-1)\sigma_{-1}\sqrt{2}\exp\Big(\frac{(\mu_{-1}-1)^2}{2} - \frac{\mu_{-1}^2}{2\sigma_{-1}^2}\Big),\\
\frac{\partial h(\eta)}{\partial \mu_1} &= \frac{\sqrt{2}P(Y=1)(\mu_1(\sigma_1^2-1)-\sigma_1^2)}{\sigma_1}\exp\left(\frac{(\mu_1-1)^2}{2} - \frac{\mu_1^2}{2\sigma_1^2}\right),\\
\frac{\partial h(\eta)}{\partial \mu_{-1}} &= \frac{\sqrt{2}P(Y=-1)(\mu_{-1}(\sigma_{-1}^2-1)+\sigma_{-1}^2)}{\sigma_{-1}}\exp\left(\frac{(\mu_{-1}+1)^2}{2} - \frac{\mu_{-1}^2}{2\sigma_{-1}^2}\right),\\
\frac{\partial h(\eta)}{\partial \sigma_1^2} &= \frac{P(Y=1)(\mu_1^2+\sigma_1^2)}{\sqrt{2\sigma_1}}\Big(\frac{(\mu_1-1)^2}{2} - \frac{\mu_1^2}{2\sigma_1^2}\Big),\\
\frac{\partial h(\eta)}{\partial \sigma_{-1}^2} &= \frac{P(Y=-1)(\mu_{-1}^2+\sigma_{-1}^2)}{\sqrt{2\sigma_{-1}}}\Big(\frac{(\mu_{-1}+1)^2}{2} - \frac{\mu_{-1}^2}{2\sigma_{-1}^2}\Big).
\end{aligned}$$

Figure 4 plots the asymptotic accuracy of $\hat{R}_n(\theta)$ for log-loss. The left panel shows that the accuracy of $\hat{R}_n$ increases with the imbalance of the marginal distribution $p(Y)$. The right panel shows that the accuracy of $\hat{R}_n$ increases with the difference between the means $|\mu_1 - \mu_{-1}|$ and the variances σ_1/σ_2.

2.5 Multiclass Classification

Thus far, we have considered unsupervised risk estimation in binary classification. In this section we describe a multiclass extension based on standard extensions of the margin concept to multiclass classification. In this case the margin vector associated with the multiclass classifier

$$\hat{Y} = \underset{k=1,\ldots,K}{\arg\max}\, f_{\theta^k}(X), \qquad X, \theta^k \in \mathbb{R}^d$$

is $f_\theta(X) = (f_{\theta^1}(X), \ldots, f_{\theta^K}(X))$. Following our discussion of the binary case, $f_{\theta^k}(X)|Y, k = 1, \ldots, K$ is assumed to be normally distributed with parameters that are estimated by maximizing the likelihood of a Gaussian mixture model. We thus have K Gaussian mixture models, each one with K mixture components. The estimated parameters are plugged-in as before into the multiclass risk

$$R(\boldsymbol{\theta}) = E_{p(f_{\boldsymbol{\theta}}(X),Y)}\mathcal{L}(Y, f_{\boldsymbol{\theta}}(X))$$

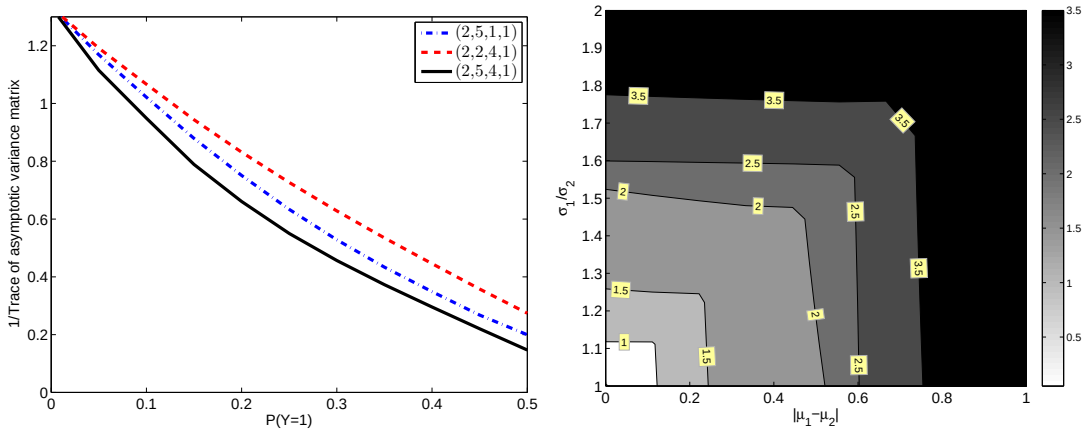

Figure 4: Left panel: asymptotic accuracy (inverse of trace of asymptotic variance) of $\hat{R}_n(\theta)$ for logloss as a function of the imbalance of the class marginal $p(Y)$. The accuracy increases with the class imbalance as it is easier to separate the two mixture components. Right panel: asymptotic accuracy (inverse of trace of asymptotic variance) as a function of the difference between the means $|\mu_1 - \mu_{-1}|$ and the variances σ_1/σ_2. See text for more information.

where $\mathcal{L}$ is a multiclass margin based loss function such as

$$\mathcal{L}(Y, f_\theta(X)) = \sum_{k \neq Y} \log(1 + \exp(-f_{\theta^k}(X))), \tag{15}$$

$$\mathcal{L}(Y, f_\theta(X)) = \sum_{k \neq Y} (1 + f_{\theta^k}(X))_+. \tag{16}$$

Care should be taken when defining the loss function for the multi-class case, as a straight-forward extension from the binary case might render the framework inconsistent. We use the specific extension which is proved to be consistent for various loss functions (including hinge-loss) by Tewari and Bartlett (2007). Since the MLE for a Gaussian mixture model with K components is consistent (assuming $P(Y)$ is known and all probabilities $P(Y = k), k = 1, \ldots, K$ are distinct) the MLE estimator for $f_{\theta^k}(X)|Y = k'$ are consistent. Furthermore, if the loss $\mathcal{L}$ is a continuous function of these parameters (as is the case for (15)-(16)) the risk estimator $\hat{R}_n(\theta)$ is consistent as well.

3. Application 1: Estimating Risk in Transfer Learning

We consider applying our estimation framework in two ways. The first application, which we describe in this section, is estimating margin-based risks in transfer learning where classifiers are trained on one domain but tested on a somewhat different domain. The transfer learning assumption that labeled data exists for the training domain but not for the test domain motivates the use of our unsupervised risk estimation. The second application, which we describe in the next section, is more ambitious. It is concerned with training classifiers without labeled data whatsoever.

In evaluating our framework we consider both synthetic and real-world data. In the synthetic experiments we generate high dimensional data from two uniform distributions $X|\{Y = 1\}$ and

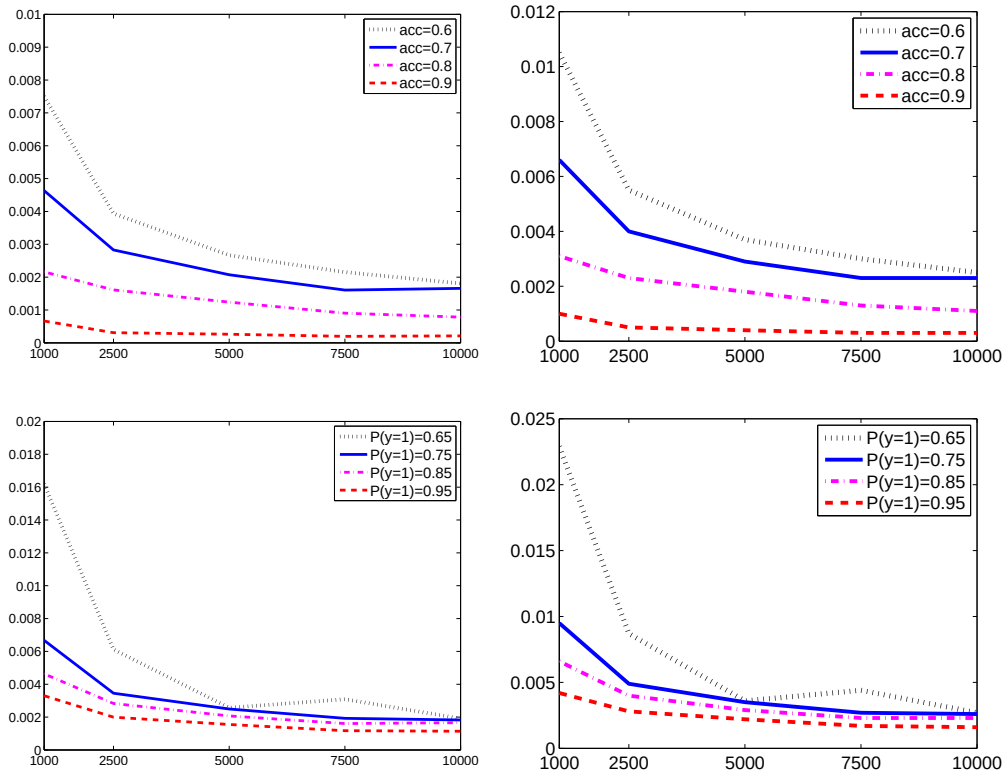

Figure 5: The relative accuracy of $\hat{R}_n$ (measured by $|\hat{R}_n(\theta) - R_n(\theta)|/R_n(\theta)$) as a function of n, classifier accuracy (acc) and the label marginal $p(Y)$ (left: logloss, right: hinge-loss). The estimation error nicely decreases with n (approaching 1% at $n = 1000$ and decaying further). It also decreases with the accuracy of the classifier (top) and non-uniformity of $p(Y)$ (bottom) in accordance with the theory of Section 2.4.

$X|\{Y = -1\}$ with independent dimensions and prescribed $p(Y)$ and classification accuracy. This controlled setting allows us to examine the accuracy of the risk estimator as a function of n, $p(Y)$, and the classifier accuracy.

Figure 5 shows that the relative error of $\hat{R}_n(\theta)$ (measured by $|\hat{R}_n(\theta) - R_n(\theta)|/R_n(\theta)$) in estimating the logloss (left) and hinge loss (right). The curves decrease with n and achieve accuracy of greater than 99% for $n > 1000$. In accordance with the theoretical results in Section 2.4 the figure shows that the estimation error decreases as the classifiers become more accurate and as $p(Y)$ becomes less uniform. We found these trends to hold in other experiments as well. In the case of exponential loss, however, the estimator performed substantially worse across the board, in some cases with an absolute error of as high as 10. This is likely due to the exponential dependency of the loss on $Y f_\theta(X)$ which makes it very sensitive to outliers.

Table 1 shows the accuracy of logloss estimation for a real world transfer learning experiment based on the 20-newsgroup data. We followed the experimental setup of used by Dai et al. (2007) in order to have different distributions for training and test sets. More specifically, 20-newsgroup

Data	R_n	$\lvert R_n - \hat{R}_n \rvert$	$\lvert R_n - \hat{R}_n \rvert / R_n$	n	$p(Y=1)$
sci vs. comp	0.7088	0.0093	0.013	3590	0.8257
sci vs. rec	0.641	0.0141	0.022	3958	0.7484
talk vs. rec	0.5933	0.0159	0.026	3476	0.7126
talk vs. comp	0.4678	0.0119	0.025	3459	0.7161
talk vs. sci	0.5442	0.0241	0.044	3464	0.7151
comp vs. rec	0.4851	0.0049	0.010	4927	0.7972

Table 1: Error in estimating logloss for logistic regression classifiers trained on one 20-newsgroup classification task and tested on another. We followed the transfer learning setup described by Dai et al. (2007) which may be referred to for more detail. The train and testing sets contained samples from two top categories in the topic hierarchy but with different subcategory proportions. The first column indicates the top category classification task and the second indicates the empirical log-loss R_n calculated using the true labels of the testing set (5). The third and forth columns indicate the absolute and relative errors of $\hat{R}_n$. The fifth and sixth columns indicate the train set size and the label marginal distribution.

data has a hierarchical class taxonomy and the transfer learning problem is defined at the top-level categories. We split the data based on subcategories such that the training and test sets contain data sampled from different subcategories within the same top-level category. Hence, the training and test distributions differ. We trained a logistic regression classifier on the training set and estimate its risk on the test set of a different distribution. Our unsupervised risk estimator was quite effective in estimating the risk with relative accuracy greater than 96% and absolute error less than 0.02.

4. Application 2: Unsupervised Learning of Classifiers

Our second application is a very ambitious one: training classifiers using unlabeled data by minimizing the unsupervised risk estimate $\hat{\theta}_n = \arg\min \hat{R}_n(\theta)$. We evaluate the performance of the learned classifier $\hat{\theta}_n$ based on three quantities: (i) the unsupervised risk estimate $\hat{R}_n(\hat{\theta}_n)$, (ii) the supervised risk estimate $R_n(\hat{\theta}_n)$, and (iii) its classification error rate. We also compare the performance of $\hat{\theta}_n = \arg\min \hat{R}_n(\theta)$ with that of its supervised analog $\arg\min R_n(\theta)$.

We compute $\hat{\theta}_n = \arg\min \hat{R}_n(\theta)$ using two algorithms (see Algorithms 1-2) that start with an initial $\theta^{(0)}$ and iteratively construct a sequence of classifiers $\theta^{(1)}, \ldots, \theta^{(T)}$ which steadily decrease $\hat{R}_n$. Algorithm 1 adopts a gradient descent-based optimization. At each iteration t, it approximates the gradient vector $\nabla \hat{R}_n(\theta^{(t)})$ numerically using a finite difference approximation (17). We compute the integral in the loss function estimator using numeric integration. Since the integral is one dimensional a variety of numeric methods may be used with high accuracy and fast computation. Algorithm 2 proceeds by constructing a grid search along every dimension of $\theta^{(t)}$ and set $[\theta^{(t)}]_i$ to the grid value that minimizes $\hat{R}_n$ (iteratively optimize one dimension at a time). This amounts to greedy search converging to local maxima. The same might hold for Algorithm 1, but we observe that Algorithm 1 works slightly better in practice, leading to lower test error with less number of training iterations.

Although we focus on unsupervised training of logistic regression (minimizing unsupervised logloss estimate), the same techniques may be generalized to train other margin-based classifiers such as SVM by minimizing the unsupervised hinge-loss estimate.

Algorithm 1 Unsupervised Gradient Descent

Input: $X^{(1)},\ldots,X^{(n)} \in \mathbb{R}^d$, $p(Y)$, step size α
Initialize $t=0$, $\theta^{(t)} = \theta^0 \in \mathbb{R}^d$
repeat
  Compute $f_{\theta^{(t)}}(X^{(j)}) = \langle \theta^{(t)}, X^{(j)} \rangle \; \forall j = 1,\ldots,n$
  Estimate $(\hat{\mu}_1, \hat{\mu}_{-1}, \hat{\sigma}_1, \hat{\sigma}_{-1})$ by maximizing (11)
  for $i = 1$ **to** d **do**
    Plug-in the estimates into (14) to approximate

$$\frac{\partial \hat{R}_n(\theta^{(t)})}{\partial \theta_i} = \frac{\hat{R}_n(\theta^{(t)} + h_i e_i) - \hat{R}_n(\theta^{(t)} - h_i e_i)}{2h_i}$$
$$(e_i \text{ is an all zero vector except for } [e_i]_i = 1) \qquad (17)$$

  end for
  Form $\nabla \hat{R}_n(\theta^{(t)}) = (\frac{\partial \hat{R}_n(\theta^{(t)})}{\partial \theta_1^{(t)}}, \ldots, \frac{\partial \hat{R}_n(\theta^{(t)})}{\partial \theta_d^{(t)}})$
  Update $\theta^{(t+1)} = \theta^{(t)} - \alpha \nabla \hat{R}_n(\theta^{(t)})$, $t = t+1$
until convergence
Output: linear classifier $\theta^{\text{final}} = \theta^{(t)}$

Algorithm 2 Unsupervised Grid Search

Input: $X^{(1)},\ldots,X^{(n)} \in \mathbb{R}^d$, $p(Y)$, grid-size τ
Initialize $\theta_i \sim \text{Uniform}(-2,2)$ for all i
repeat
  for $i = 1$ **to** d **do**
    Construct τ points grid in the range $[\theta_i - 4\tau, \theta_i + 4\tau]$
    Compute the risk estimate (14) where all dimensions of $\theta^{(t)}$ are fixed except for $[\theta^{(t)}]_i$ which is evaluated at each grid point.
    Set $[\theta^{(t+1)}]_i$ to the grid value that minimized (14)
  end for
until convergence
Output: linear classifier $\theta^{\text{final}} = \theta$

Figures 6-7 display $\hat{R}_n(\hat{\theta}_n)$, $R_n(\hat{\theta}_n)$ and error-rate$(\hat{\theta}_n)$ on the training and testing sets as on two real world data sets: RCV1 (text documents) and MNIST (handwritten digit images) data sets. In the case of RCV1 we discarded all but the most frequent 504 words (after stop-word removal) and represented documents using their tfidf scores. We experimented on the binary classification task of distinguishing the top category (positive) from the next 4 top categories (negative) which resulted in $p(y=1) = 0.3$ and $n = 199328$. 70% of the data was chosen as a (unlabeled) training set and the rest was held-out as a test-set. In the case of MNIST data, we normalized each of the $28 \times 28 = 784$

pixels to have 0 mean and unit variance. Our classification task was to distinguish images of the digit one (positive) from the digit 2 (negative) resulting in 14867 samples and $p(Y=1)=0.53$. We randomly choose 70% of the data as a training set and kept the rest as a testing set.

Figures 6-7 indicate that minimizing the unsupervised logloss estimate is quite effective in learning an accurate classifier without labels. Both the unsupervised and supervised risk estimates $\hat{R}_n(\hat{\theta}_n)$, $R_n(\hat{\theta}_n)$ decay nicely when computed over the train set as well as the test set. Also interesting is the decay of the error rate. For comparison purposes supervised logistic regression with the same n achieved only slightly better test set error rate: 0.05 on RCV1 (instead of 0.1) and 0.07 or MNIST (instead of 0.1).

In another experiment we examined the proposed approach on several different data sets and compared the classification performance with a supervised baseline (logistic regression) and Gaussian mixture modeling (GMM) clustering with known label proportions in the original data space (Table 2). The comparison was made under the same experimental setting (n, $p(Y)$) for all three approaches. We used data sets from UCI machine learning repository (Frank and Asuncion, 2010) and from previously cited sources, unless otherwise noted. The following tasks were considered for each data set.

- RCV1: top category versus next 4 categories
- MNIST: Digit 1 versus Digit 2
- 20 newsgroups: Comp category versus Recreation category
- USPS[1]: Digit 2 versus Digit 5
- Umist[1]: Male face (16 subjects) versus Female faces (4 subjects) with image resolution reduced to 40×40
- Arcene: Cancer versus Normal
- Isolet: Vowels versus Consonants
- Dexter: Documents about corporate acquisitions versus rest
- Secom: Semiconductor manufacturing defects versus good items
- Pham faces: Face versus Non-face images
- CMU pie face[2]: male (30 subjects) vs female (17 subjects)
- Madelon[3]: It consists of data points (artificially generated) grouped in 32 clusters placed on the vertices of a five dimensional hypercube and randomly labeled +1 or -1, corrupted with features that are not useful for classification.

1. Data set can be found at `http://www.cs.nyu.edu/~roweis/data.html`.
2. Data set can be found at `http://www.zjucadcg.cn/dengcai/Data/FaceData.html`.
3. Data set can be found at `http://archive.ics.uci.edu/ml/machine-learning-databases/madelon/Dataset.pdf`.

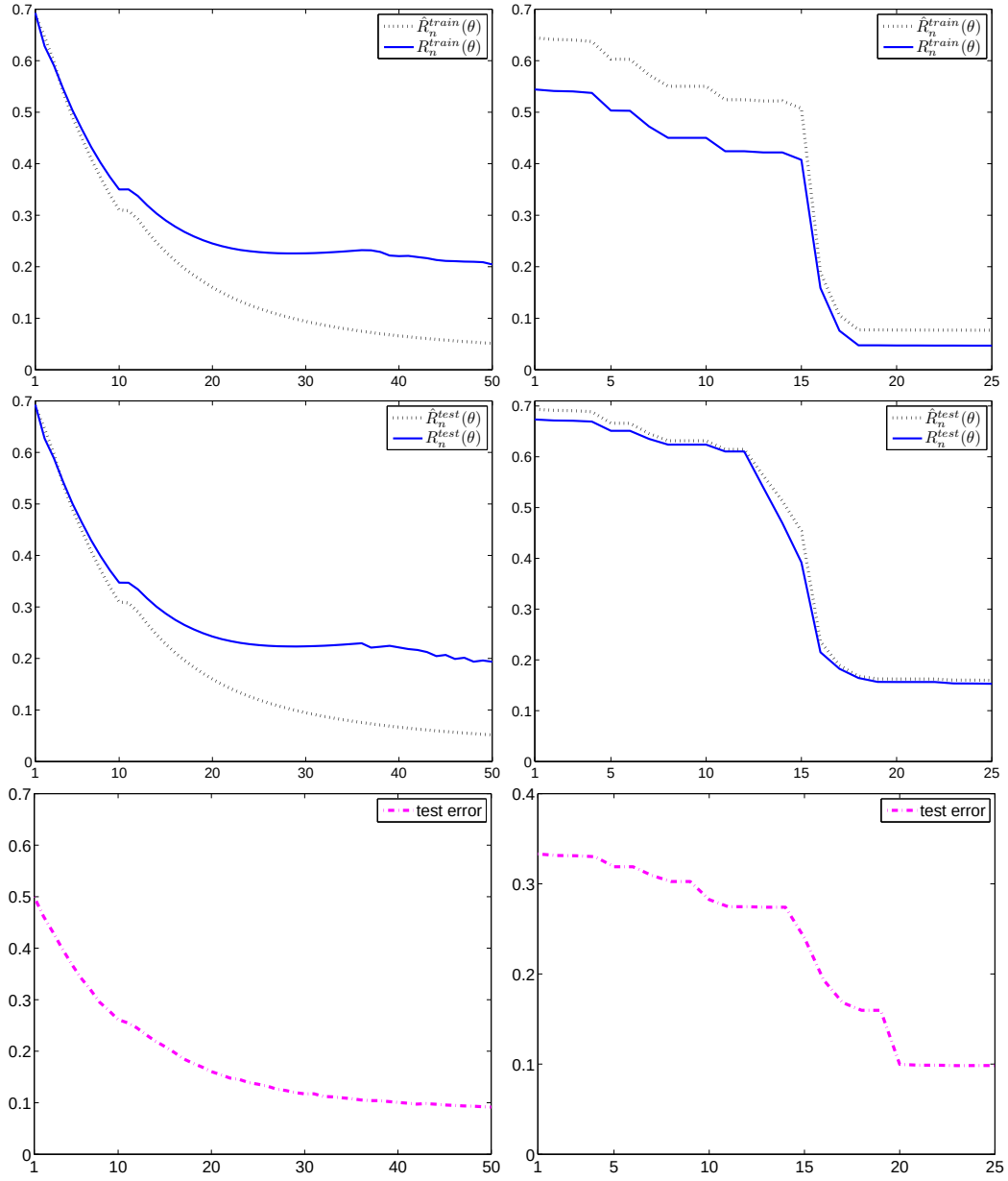

Figure 6: Performance of unsupervised logistic regression classifier $\hat{\theta}_n$ computed using Algorithm 1 (left) and Algorithm 2 (right) on the RCV1 data set. The top two rows show the decay of the two risk estimates $\hat{R}_n(\hat{\theta}_n)$, $R_n(\hat{\theta}_n)$ as a function of the algorithm iterations. The risk estimates of $\hat{\theta}_n$ were computed using the train set (top) and the test set (middle). The bottom row displays the decay of the test set error rate of $\hat{\theta}_n$ as a function of the algorithm iterations. The figure shows that the algorithm obtains a relatively accurate classifier (testing set error rate 0.1, and $\hat{R}_n$ decaying similarly to R_n) without the use of a single labeled example. For comparison, the test error rate for supervised logistic regression with the same n is 0.07.

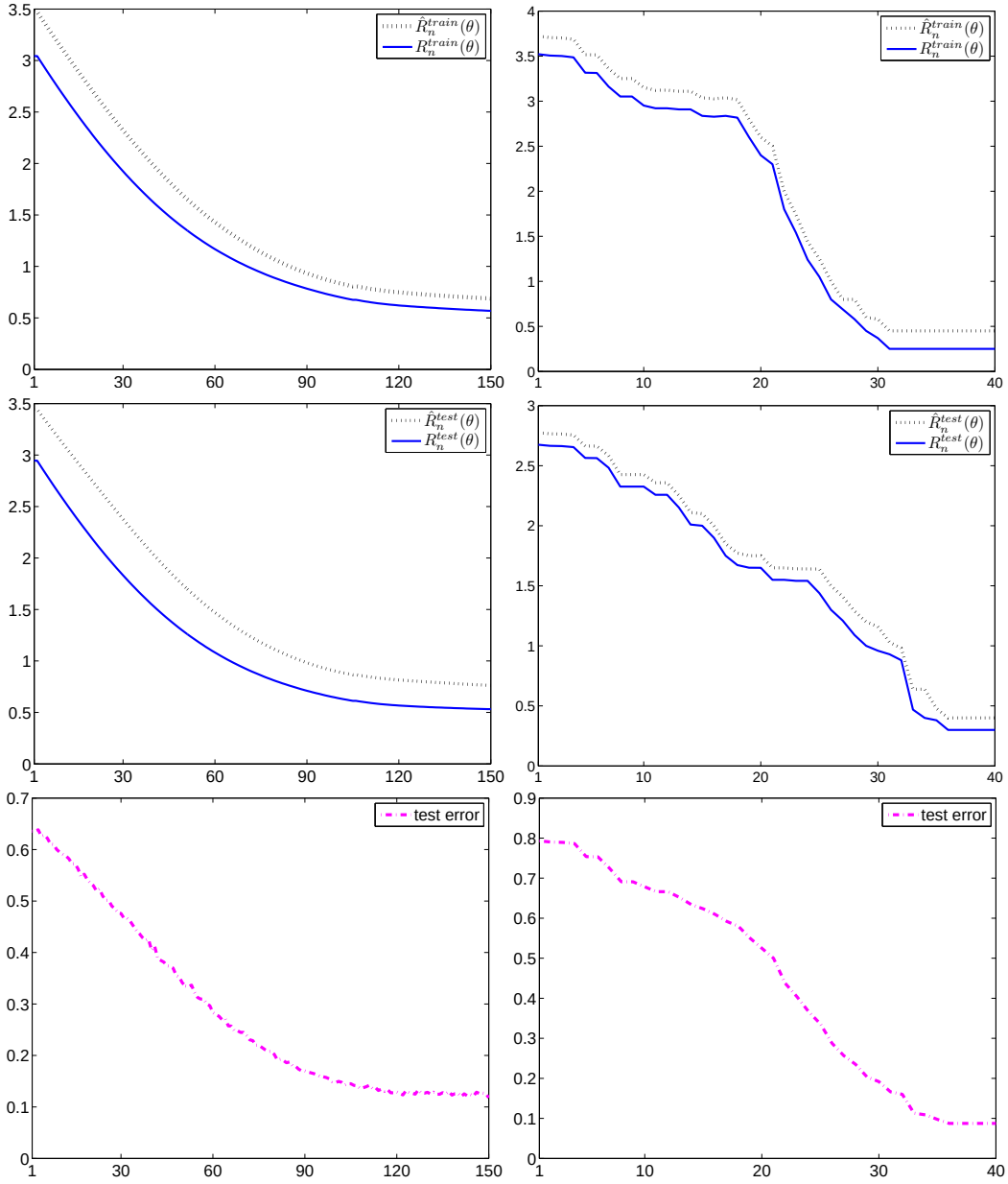

Figure 7: Performance of unsupervised logistic regression classifier $\hat{\theta}_n$ computed using Algorithm 1 (left) and Algorithm 2 (right) on the MNIST data set. The top two rows show the decay of the two risk estimates $\hat{R}_n(\hat{\theta}_n)$, $R_n(\hat{\theta}_n)$ as a function of the algorithm iterations. The risk estimates of $\hat{\theta}_n$ were computed using the train set (top) and the test set (middle). The bottom row displays the decay of the test set error rate of $\hat{\theta}_n$ as a function of the algorithm iterations. The figure shows that the algorithm obtains a relatively accurate classifier (testing set error rate 0.1, and $\hat{R}_n$ decaying similarly to R_n) without the use of a single labeled example. For comparison, the test error rate for supervised logistic regression with the same n is 0.05.

Data set	Dimensions	Supervised log-reg	USL-2	GMM
RCV1	top 504 words	0.0500	0.0923	0.2083
Mnist	784	0.0700	0.1023	0.3163
20 news group	top 750 words	0.0652	0.0864	0.1234
USPS	256	0.0348	0.0545	0.1038
Umist	400 PCA components	0.1223	0.1955	0.2569
Arcene	1000 PCA components	0.1593	0.1877	0.3843*
Isolet	617	0.0462	0.0568	0.1332
Dexter	top-700 words	0.0564	0.1865	0.2715
Secom	591	0.1246	0.1532	0.2674
Pham faces	400	0.1157	0.1669	0.2324
CMU pie face	1024	0.0983	0.1386	0.2682*
Madelon	500	0.0803	0.1023	0.1120

Table 2: Comparison (test set error rate) between supervised logistic regression, Unsupervised logistic regression and Gaussian mixture modeling in original data space. The unsupervised classifier performs better than the GMM clustering on the original space and compares well with its supervised counterpart on most data sets. See text for more details. The stars represent GMM with covariance $\sigma^2 I$ due to the high dimensionality. In all other cases we used a diagonal covariance matrix. Non-diagonal covariance matrix was impractical due to the high dimensionality.

Table 2 displays the test set error for the three methods on each data set. We note that our unsupervised approach achieves test set errors comparable to the supervised logistic regression in several data sets. The poor performance of the unsupervised technique on the Dexter data set is due to the fact that the data contains many irrelevant features. In fact it was engineered for a feature selection competition and has a sparse solution vector. In general our method significantly outperforms Gaussian mixture model clustering in the original feature space. A likely explanation is that (i) $f_\theta(X)|Y$ is more likely to be normal than $X|Y$ and (ii) it is easier to estimate in one dimensional space rather than in a high dimensional space.

4.1 Inaccurate Specification of $p(Y)$

Our estimation framework assumes that the marginal $p(Y)$ is known. In some cases we may only have an inaccurate estimate of $p(Y)$. It is instructive to consider how the performance of the learned classifier degrades with the inaccuracy of the assumed $p(Y)$.

Figure 8 displays the performance of the learned classifier for RCV1 data as a function of the assumed value of $p(Y=1)$ (correct value is $p(Y=1)=0.3$). We conclude that knowledge of $p(Y)$ is an important component in our framework but precise knowledge is not crucial. Small deviations of the assumed $p(Y)$ from the true $p(Y)$ result in a small degradation of logloss estimation quality and testing set error rate. Naturally, large deviation of the assumed $p(Y)$ from the true $p(Y)$ renders the framework ineffective.

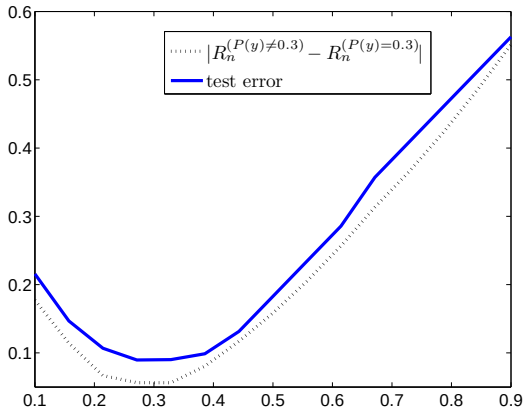

Figure 8: Performance of unsupervised classifier training on RCV1 data (top class vs. classes 2-5) for misspecified $p(Y)$. The performance of the estimated classifier (in terms of training set empirical logloss R_n (5) and test error rate measured using held-out labels) decreases with the deviation between the assumed and true $p(Y=1)$ (true $p(Y=1)=0.3$)). The classifier performance is very good when the assumed $p(Y)$ is close to the truth and degrades gracefully when the assumed $p(Y)$ is not too far from the truth.

4.2 Effect of Regularization and Dimensionality Reduction

In Figure 9 we examine the effect of regularization on the performance of the unsupervised classifier. In this experiment we use the L_1 regularization software available at `http://www.cs.ubc.ca/~schmidtm/Software/L1General.html`. Clearly, regularization helps in the supervised case. It appears that in the USL case weak regularization may improve performance but not as drastically as in the supervised case. Furthermore, the positive effect of L_1 regularization in the USL case appears to be weaker than L_2 regularization (compare the left and right panels of Figure 9). One possible reason is that the sparsity promoting nature of L_1 conflicts with the CLT assumption.

In Figure 10 we examine the effect of reducing the data dimensionality via PCA prior to training the unsupervised classifier. Specifically, the 256 dimensions USPS image data set was embedded in an increasingly lower dimensional space via PCA. For the original dimensionality of 256 or a slightly lower dimensionality the classification performance of the unsupervised classifier is comparable to the supervised. Once the dimensions are reduced to less than 150 a significant performance gap appears. This is consistent with our observation above that for lower dimensions the CLT approximation is less accurate. The supervised classifier also degrades in performance as less dimensions are used but not as fast as the unsupervised classifier.

5. Related Work

Semi-supervised approaches: Semisupervised learning is closely related to our work in that unsupervised classification may be viewed as a limiting case. One of the first attempts at studying the sample complexity of classification with unlabeled and labeled data was by Castelli and Cover (1995). They consider a setting when data is generated by mixture distributions and show that with infinite unlabeled data, the probability of error decays exponentially faster in the labeled data to the

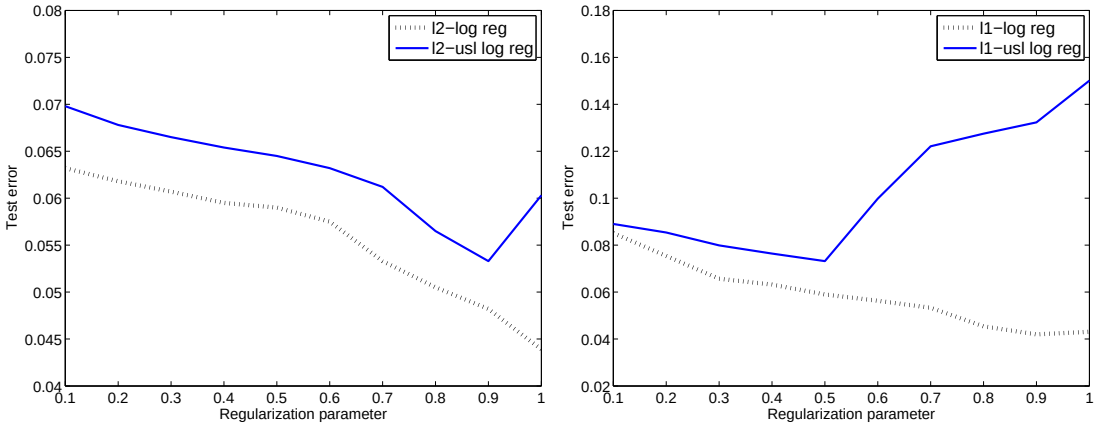

Figure 9: Test set error rate versus regularization parameter (L_2 on the left panel and L_1 on the right panel) for supervised and unsupervised logistic regression on RCV1 data set.

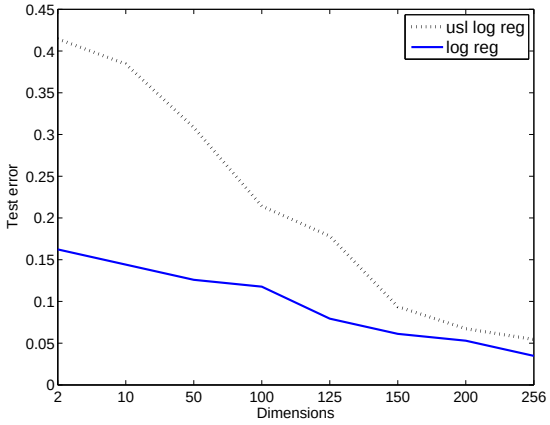

Figure 10: Test set error rate versus the amount of dimensions used (extracted via PCA) for supervised and unsupervised logistic regression on USPS data set. The original dimensionality was 256.

Bayes risk. They also analyze the case when there are only finite labeled and unlabeled data samples, with known class conditional densities but unknown mixing proportions (Castelli and Cover, 1996). A variant of the same scenario with known parametric forms for the class conditionals (specifically *n*-dimensional Gaussians) but unknown parameters and mixing proportions is also analyzed by J. Ratsaby and Venkatesh (1995). Some of the more recent work in the area concentrated on analyzing semisupervised learning under the cluster assumption or the manifold assumption. We refer the reader to a recent survey by Zhu and Goldberg (2009) for a discussion of recent approaches. However, none of the prior work consider mixture modeling in the projected 1-d space along with a CLT assumption which we exploit. In addition, assuming known mixing proportions, we propose

a framework for training a classifier with no labeled samples, while approaches above still need labeled samples for classification.

Unsupervised approaches: The most recent related research approaches are by Quadrianto et al. (2009), Gomes et al. (2010), and Donmez et al. (2010). The work by Quadrianto et al. (2009) aims to estimate the labels of an unlabeled testing set using known label proportions of several sets of unlabeled observations. The key difference between their approach and ours is that they require separate training sets from different sampling distributions with different and known label marginals (one for each label). Our method assumes only a single data set with a known label marginal but on the other hand assumed the CLT approximation. Furthermore, as noted previously (see comment after Proposition 8), our analysis is in fact valid when only the order of label proportions is known, rather than the absolute values.

A different attempt at solving this problem is provided by Gomes et al. (2010) which focuses on discriminative clustering. This approach attempts to estimate a conditional probabilistic model in an unsupervised way by maximizing mutual information between the empirical input distribution and the label distribution. A key difference is the focus on probabilistic classifiers and in particular logistic regression whereas our approach is based on empirical risk minimization which also includes SVM. Another key difference is that the work by Gomes et al. (2010) lacks consistency results which characterize when it works from a theoretical perspective. The approach by Donmez et al. (2010) focuses on estimating the error rate of a given stochastic classifier (not necessarily linear) without labels. It is similar in that it estimates the 0/1 risk rather than the margin based risk. However, it uses a different strategy and it replaces the CLT assumption with a symmetric noise assumption.

An important distinction between our work and the references above is that our work provides an estimate for the margin-based risk and therefore leads naturally to unsupervised versions of logistic regression and support vector machines. We also provide asymptotic analysis showing convergence of the resulting classifier to the optimal classifier (minimizer of (1)). Experimental results show that in practice the accuracy of the unsupervised classifier is on the same order (but slightly lower naturally) as its supervised analog.

6. Discussion

In this paper we developed a novel framework for estimating margin-based risks using only unlabeled data. We show that it performs well in practice on several different data sets. We derived a theoretical basis by casting it as a maximum likelihood problem for Gaussian mixture model followed by plug-in estimation.

Remarkably, the theory states that assuming normality of $f_\theta(X)$ and a known $p(Y)$ we are able to estimate the risk $R(\theta)$ without a single labeled example. That is the risk estimate converges to the true risk as the number of unlabeled data increase. Moreover, using uniform convergence arguments it is possible to show that the proposed training algorithm converges to the optimal classifier as $n \to \infty$ without any labeled data. The results in the paper are applicable only to linear classifiers, which are an extremely important class of classifiers especially in the high dimensional case. In the non-linear classification scenario, it is worth examining if the CLT assumptions on the mapped high-dimensional feature space could be used for building non-linear classifiers via the kernel trick.

On a more philosophical level, our approach points at novel questions that go beyond supervised and semi-supervised learning. What benefit do labels provide over unsupervised training? Can

our framework be extended to semi-supervised learning where a few labels do exist? Can it be extended to non-classification scenarios such as margin based regression or margin based structured prediction? When are the assumptions likely to hold and how can we make our framework even more resistant to deviations from them? These questions and others form new and exciting open research directions.

Acknowledgments

The authors thank the action editor and anonymous reviewers for their constructive comments, that not only helped at a conceptual level but also helped improve the presentation. In addition, we thank John Lafferty and Vladimir Koltchinskii for discussions and several insightful comments. This work was funded in part by NSF grant IIS-0906550.

References

J. Behboodian. Information matrix for a mixture of two normal distributions. *Journal of Statistical Computation and Simulation*, 1(4):295–314, 1972.

K. N. Berk. A central limit theorem for *m*-dependent random variables with unbounded *m*. *The Annals of Probability*, 1(2):352–354, 1973.

V. Castelli and T. M. Cover. On the exponential value of labeled samples. *Pattern Recognition Letters*, 16(1):105–111, 1995.

V. Castelli and T. M. Cover. The relative value of labeled and unlabeled samples in pattern recognition with an unknown mixing parameter. *IEEE Transactions on Information Theory*, 42(6): 2102–2117, 1996.

W. Dai, Q. Yang, G.-R. Xue, and Y. Yu. Boosting for transfer learning. In *Proc. of International Conference on Machine Learning*, 2007.

J. Davidson. *Stochastic Limit Theory: An Introduction for Econometricians*. Oxford University Press, USA, 1994.

P. Donmez, G. Lebanon, and K. Balasubramanian. Unsupervised supervised learning I: Estimating classification and regression error rates without labels. *Journal of Machine Learning Research*, 11(April):1323–1351, 2010.

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

A. Frank and A. Asuncion. UCI machine learning repository. *University of California, School of Information and Computer Science, Irvine, CA. Available at http://archive.ics.uci.edu/ml/*, 2010.

R. Gomes, A. Krause, and P. Perona. Discriminative clustering by regularized information maximization. In *Advances in Neural Information Processing Systems 24*, 2010.

W. Hoeffding and H. Robbins. The central limit theorem for dependent random variables. *Duke Mathematical Journal*, 15:773–780, 1948.

J. J. Ratsaby and S. S. Venkatesh. Learning from a mixture of labeled and unlabeled examples with parametric side information. In *Annual conference on Computational learning theory*, 1995.

D. Lewis, Y. Yang, T. Rose, and F. Li. RCV1: A new benchmark collection for text categorization research. *Journal of Machine Learning Research*, 5:361–397, 2004.

T. V. Pham, M. Worring, and A. W. M. Smeulders. Face detection by aggregated bayesian network classifiers. *Pattern Recognition Letters*, 23(4):451–461, February 2002.

N. Quadrianto, A. J. Smola, T. S. Caetano, and Q. V. Le. Estimating labels from label proportions. *Journal of Machine Learning Research*, 10:2349–2374, 2009.

Y. Rinott. On normal approximation rates for certain sums of dependent random variables. *Journal of Computational and Applied Mathematics*, 55(2):135–143, 1994.

H. Teicher. Identifiability of finite mixtures. *The Annals of Mathematical Statistics*, 34(4):1265–1269, 1963.

A. Tewari and P. L. Bartlett. On the consistency of multiclass classification methods. *Journal of Machine Learning Research*, pages 1007–1025, 2007.

X. Zhu and A. B. Goldberg. *Introduction to Semi-supervised Learning*. Morgan & Claypool Publishers, 2009.