Unsupervised Supervised Learning II: Margin-Based Classification without Labels

Krishnakumar Balasubramanian
Georgia Institute of Technology
krishnakumar3@gatech.edu

Pinar Donmez
Yahoo! Labs
pinard@yahoo-inc.com

Guy Lebanon
Georgia Institute of Technology
lebanon@cc.gatech.edu

Abstract

Many popular linear classifiers, such as logistic regression, boosting, or SVM, are trained by optimizing margin-based risk functions. Traditionally, these risk functions are computed based on a labeled dataset. We develop a novel technique for estimating such risks using only unlabeled data and knowledge of $p(y)$. We prove that the proposed risk estimator is consistent on high-dimensional datasets 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 using exclusively unlabeled data.

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 $\hat{Y} = \text{sign} \sum \theta_j X_j$ with $Y \in \{-1, +1\}$, and $X, \theta \in \mathbb{R}^d$ the margin is defined as the product

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

Training such classifiers is done by attempting to minimize the risk or expected loss

$$R(\theta) = \mathsf{E}_{p(X,Y)} L(Y, f_\theta(X)) \quad (2)$$

with the three most popular loss functions

$$L_1(Y, f_\theta(X)) = \exp\left(-Y f_\theta(X)\right) \quad (3)$$
$$L_2(Y, f_\theta(X)) = \log\left(1 + \exp\left(-Y f_\theta(X)\right)\right) \quad (4)$$
$$L_3(Y, f_\theta(X)) = \max(1 - Y f_\theta(X), 0). \quad (5)$$

being exponential loss L_1 (boosting), logloss L_2 (logistic regression) and hinge loss L_3 (SVM). Since $R(\theta)$ depends on the unknown p, it is replaced with its empirical counterpart based on a labeled training set

$$\hat{\theta}_n = \arg\min_\theta R_n(\theta). \quad (6)$$

$$R_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} L(Y^{(i)}, f_\theta(X^{(i)})) \quad (7)$$

$$(X^{(1)}, Y^{(1)}), \dots, (X^{(n)}, Y^{(n)}) \overset{\text{iid}}{\sim} p. \quad (8)$$

Note, however, that evaluating and minimizing R_n requires labeled data (8). 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

$$X^{(1)}, \dots, X^{(n)} \overset{\text{iid}}{\sim} p \quad (9)$$

instead of (8). Our estimator is based on the observations that when the data is high dimensional ($d \to \infty$)

$$f_\theta(X)|\{Y = y\}, \quad y \in \{-1, +1\} \quad (10)$$

is often 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 (10) may be estimated from unlabeled data (9) and that these distributions may be used to measure margin-based losses such as (3)-(5). 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 construct an estimator for $R(\theta)$ (2) using the unlabeled data (9) which we denote $\hat{R}_n(\theta\,;X^{(1)},\dots,X^{(n)})$ or simply $\hat{R}_n(\theta)$ (to distinguish it from R_n in (7)).

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. Furthermore, this assumption may be replaced with a weaker form in which we know the ordering of the marginal distributions e.g., $p(Y=1) > p(Y=-1)$, but without knowing the specific values of $p(Y)$.

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 $X_1,\dots,X_d$ are iid given Y. A more general CLT states that $f_\theta(X)|Y$ is asymptotically normal if $X_1,\dots,X_d|Y$ are independent (but not necessary identically distributed). Even more general CLTs state that $f_\theta(X)|Y$ is asymptotically normal if $X_1,\dots,X_d|Y$ are not independent but their dependency is limited in some way. We examine this in Section 2.1 where we also show that normality holds empirically for several standard datasets.

To derive the estimator we rewrite (2) by taking expectation with respect to Y and $\alpha = f_\theta(X)$

$$R(\theta) = \mathsf{E}_{p(f_\theta(X),Y)}\, L(Y, f_\theta(X)) \qquad (11)$$
$$= \sum_{y\in\{-1,+1\}} p(y) \int_{\mathbb{R}} p(f_\theta(X)=\alpha|y)L(y,\alpha)\,d\alpha.$$

Equation (11) involves three terms $L(y,\alpha)$, $p(y)$ and $p(f_\theta(X) = \alpha|y)$. The loss function 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 $\mu_y,\sigma_y,\ y \in \{-1,1\}$ that are estimated by maximizing the likelihood of a Gaussian mixture model. These estimated parameters are used to construct the plug-in estimator $\hat{R}_n(\theta)$:

$$\hat{R}_n(\theta) = \sum_y p(y) \int_{\mathbb{R}} p_{\hat{\mu}_y^{(n)},\hat{\sigma}_y^{(n)}}(f_\theta(X)=\alpha|y)L(y,\alpha)\,d\alpha$$

$$(\hat{\mu}^{(n)}, \hat{\sigma}^{(n)}) = \arg\max_{\mu,\sigma} \ell_n(\mu,\sigma) \qquad (12)$$

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

We make the following observations.

1. The parameters of the loglikelihood ℓ_n in (12) 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.

2. The estimation problem (12) is equivalent to computing the MLE for a 1-D 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.

3. The estimator $\hat{R}_n$ (12) is consistent in the limit of infinite unlabeled data
$$P\left(\lim_{n\to\infty} \hat{R}_n(\theta) = R(\theta)\right) = 1.$$

4. Under suitable conditions $\arg\min_\theta \hat{R}_n(\theta)$ converges to the expected risk minimizer
$$P\left(\lim_{n\to\infty} \arg\min_{\theta\in\Theta} R_n(\theta) = \arg\min_{\theta\in\Theta} R(\theta)\right) = 1.$$

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.

5. Our assumptions are: (a) known and non-uniform $p(y)$ (b) high dimensional feature vector (c) weak dependence between the features resulting in normality of the inner product $\langle w, f(X)\rangle|Y$ (d) nonsparsity of the parameter vector (e) the assumption of a margin-based linear classifier. As our analysis shows under these conditions it is possible to classify without labels in the limit of large data. And for finite data the performance is somewhat worse than for labeled data (we investigate this in our experiment section).

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

The quantity $f_\theta(X)|Y$ is essentially a sum of d random variables which for large d is likely to be normally distributed. One way to verify this is empirically, as we show in Figure 1 which contrast the histogram with a fitted normal pdf for text, digit images, and face images data. For these datasets 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 $10^2 - 10^4$.

Figure 1 shows that in these cases of text and image data $f_\theta(X)|Y$ is approximately normal for θ representing a random vector and an estimated classifier (Figure 1). The single caveat in this case is that normality may not hold when θ is sparse, as may happen for example for l_1 regularized models in the last row of Figure 1. See [1] for additional histograms.

From a theoretical standpoint normality may be argued using a central limit theorem. The original central limit theorem states that $\sum_{i=1}^{d} Z_i$ is approximately normal for large d if Z_i are iid.

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

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

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, by Lindberg, does not require the summands Z_i to be identically distributed and only requires that the data dimensions be independent. More general CLTs replace the condition that $Z_i, i \in \mathbb{N}$ be independent with the notion of limited dependence, for example, m-dependence [7], [2]. A detailed discussion about various such limit theorems may be found in [4]. The following result implies that normality holds if the dependency of the RVs if bounded or does not grow too fast in relation to d.

Definition 1. 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.

Proposition 2 ([11]). *Let $Z_1, \ldots, Z_n$ be random variables having a dependency graph with $\max_{v \in \mathcal{V}} d(v) < D$, and satisfying $|Z_i - \mathsf{E}\, Z_i| \leq B$ a.s., $\forall i$, $\mathsf{E}(\sum_{i=1}^{n} Z_i) = \lambda$ and $\mathsf{Var}(\sum_{i=1}^{n} Z_i) = \sigma^2 > 0$*

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

where Φ is the CDF corresponding to a $N(0,1)$ distribution. This result states a stronger result than convergence in distribution to a Gaussian in that it states a uniform rate of convergence of the CDF (Φ above is the $N(0,1)$ CDF). It can be shown that for bounded D, B and $\sigma = \mathsf{Var}(\sum_{i=1}^{n} Z_i) = O(n)$ we have $\sum_{i=1}^{d}(Z_i - \lambda)/\sigma \rightsquigarrow N$ (as $d \to \infty$) with an optimal rate of $n^{-1/2}$ [11].

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 Figure 1 above indicates this holds for text and image data for most values of θ, assuming it is not sparse. We refer the reader to [1] for a more detailed discussion.

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 $f_\theta(X)|Y$. Consistency of $\hat{R}_n(\theta)$, $\arg\min R_n(\theta)$ follows.

Definition 2. A parametric family $\{p_\alpha : \alpha \in A\}$ is identifiable when $p_\alpha(x) = p_{\alpha'}(x), \forall x$ implies $\alpha = \alpha'$.

Proposition 3. *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. The proof follows from the well known result that a family of Gaussian mixture model (with unknown $p(y)$) is identifiable up to a permutation of the
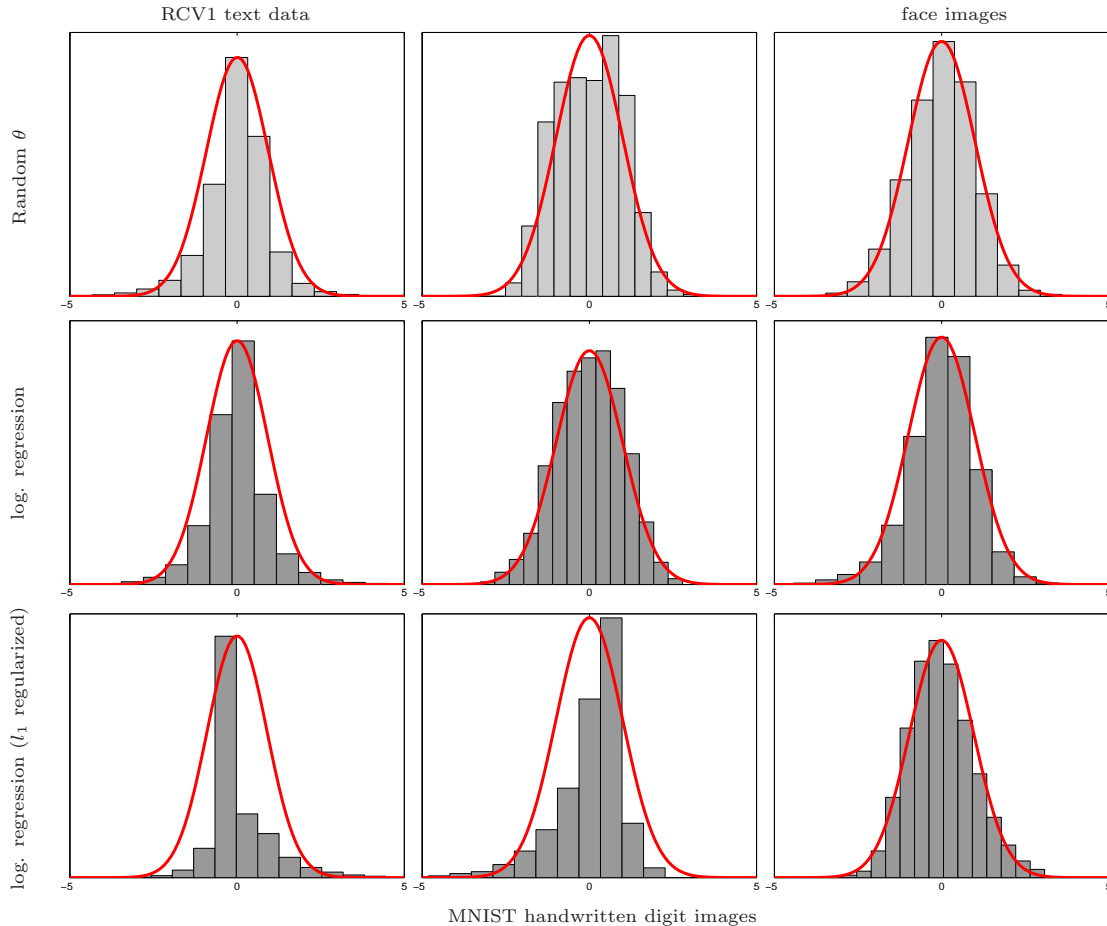
Figure 1: Centered histograms of $f_\theta(X)|\{Y = 1\}$ overlayed with the pdf of a fitted Gaussian for multiple θ vectors (three rows: random θ ($\theta_i \sim U(1/2, 1/2)$), logistic regression, and l_1 regularized logistic regression-all regularization parameters were selected by cross validation) and datasets (columns: RCV1 text data [8], MNIST digit images, and face images [9]). For uniformity we subtracted the empirical mean and divided by the empirical standard deviation. The fifteen 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).

labels y [12]. Assuming with no loss of generality that $p(y = 1) > p(y = -1)$, 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. □

Proposition 4. *Under the assumptions of Proposition 3 the MLE estimates*

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

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

are consistent i.e., $(\hat{\mu}_1^{(n)}, \hat{\mu}_{-1}^{(n)}, \hat{\sigma}_1^{(n)}, \hat{\sigma}_{-1}^{(n)})$ converge as $n \to \infty$ to the true 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 3)

in η and the available samples $z^{(i)} = f_\theta(X^{(i)})$ are iid samples from $p_\eta(z)$. Since the MLE for an identifiable parametric family is strongly consistent e.g., [6, chap. 17], the result of the proposition follows. □

Proposition 5. *Under the assumptions of Proposition 3 and assuming the loss L is given by one of (3)-(5) with a normal $f_\theta(X)|Y \sim N(\mu_y, \sigma_y^2)$, the estimate*

$$\hat{R}_n(\theta) = \sum_y p(y) \int_{\mathbb{R}} p_{\hat{\mu}_y^{(n)}, \hat{\sigma}_y^{(n)}}(f_\theta(X) = \alpha|y) L(y, \alpha) \, d\alpha.$$

is consistent, i.e., for all θ,

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

Proof. The plug-in risk estimate $\hat{R}_n$ is a continuous function (when L is given by (3), (4) or (5)) 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 4 we have

$$1 = P(\lim_n (\hat{\mu}_1^{(n)}, \hat{\mu}_{-1}^{(n)}, \hat{\sigma}_1^{(n)}, \hat{\sigma}_{-1}^{(n)}) = (\mu_1^{\text{true}}, \mu_{-1}^{\text{true}}, \sigma_1^{\text{true}}, \sigma_{-1}^{\text{true}})).$$

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 concludes the proof. $\square$

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 uniform convergence using a modification of Wald's classical MLE consistency result [6, chap. 17].

Denoting $p_\eta(z) = \sum_{y \in \{-1,+1\}} p(y) p_{\mu_y, \sigma_y}(f(X) = z|y)$, with $\eta = (\mu_1, \mu_{-1}, \sigma_1, \sigma_{-1})$, 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 6. *Let θ take values in Θ for which $\eta \in E$ for some compact set E. Assuming the conditions in Proposition 5 the convergence of the MLE $\hat{\eta}_n \to \eta_0$ is uniform in $\eta_0 \in E$ (or alternatively $\theta \in \Theta$).*

Proof. We first denote $U(z, \eta, \eta_0) = \log p_\eta(z) - \log p_{\eta_0}(z)$ and $\alpha(\eta, \eta_0) = E_{p_{\eta_0}} U(z, \eta, \eta_0) = -D(p_{\eta_0}, p_\eta) \le 0$ 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 (which is negative)

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

Invoking the uniform strong law of large numbers e.g., [6, chap. 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 $\eta = \eta_0$,

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

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 $\eta_0 \in E$. $\square$

Proposition 7. *Assuming that X, Θ are bounded in addition to the assumptions of Proposition 6 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 models (with known mixing proportions) is Lipschitz continuous in its parameters

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

which may be verified by noting that the partial derivatives of $p_\eta(z) = \sum_y p(y) p_{\mu_y, \sigma_y}(z|y)$ are bounded for a compact E. These observations, together with Proposition 6 lead to

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

uniformly over $\theta \in \Theta$. $\square$

Proposition 8. *Under the assumptions of Prop. 7*

$$P\left(\lim_{n \to \infty} \arg\min_{\theta \in \Theta} \hat{R}_n(\theta) = \arg\min_{\theta \in \Theta} R(\theta) \right) = 1.$$
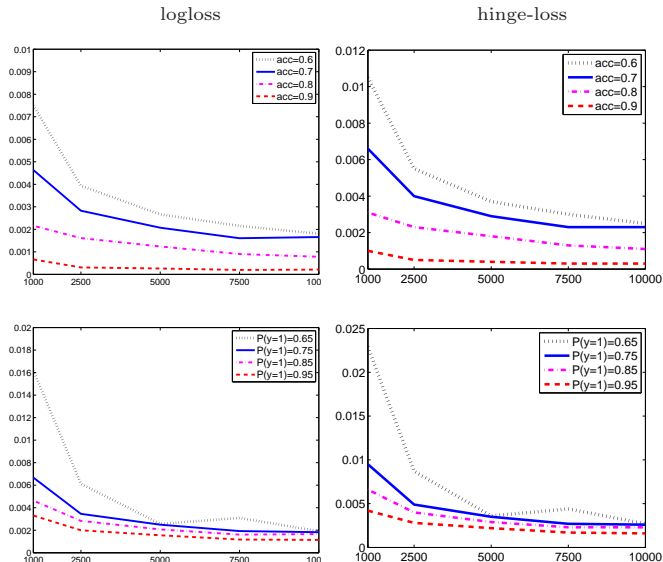
Figure 2: The relative accuracy of $\hat{R}_n$ (measured by $|\hat{R}_n(\theta) - R_n(\theta)|/R_n(\theta)$) for logloss and hinge-loss on synthetic data as a function of n (x axis), classifier accuracy (acc) and $p(Y)$. The estimation error nicely decreases with n (approaching 1% at $n = 1000$ and decaying further). It also decreases with classifier accuracy (top) and non-uniformity of $p(Y)$ (bottom).

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 $\epsilon > 0$ there exists N such that for all $n > N$, $|\hat{R}_n(\theta) - R(\theta)| < \epsilon$.

Let $S = \{\theta : \|\theta - t^*\| \geq \epsilon\}$ 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 $\theta \in S$, $\hat{R}_n(\theta) \geq R(t^*) + \epsilon$. 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 $\epsilon > 0$ there exists N such that for all $n > N$ we have $\|t_n - t^*\| \leq \epsilon$, $t_n \to t^*$ which concludes the proof. $\qquad\square$

3 Estimating Margin Based Risk

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 clas-

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

Figure 3: Error in estimating logloss for logistic regression classifiers trained on one 20-newsgroup classification task and tested on another. See text for more details.
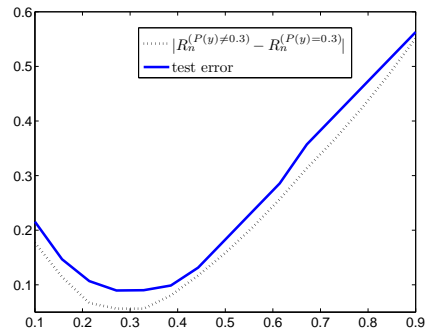

Figure 4: Performance of unsupervised classifier training on RCV1 data (top class vs. classes 2-5) for misspecified $p(Y)$. The performance of the estimated classifier (in terms of training set empirical logloss R_n (7) and test error rate measured using held-out labels) decreases with the deviation 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 truth.

sifiers 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 $X|\{Y = -1\}$ 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 2 shows that the relative error of $\hat{R}_n(\theta)$ (measured by $|\hat{R}_n(\theta) - R_n(\theta)|/R_n(\theta)$) in estimating logloss and hinge loss decreases with n achieving accuracy of greater than 99% for $n > 1000$. 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 (figure omitted). This is likely due to the exponential dependency of the loss on $Y f_\theta(X)$ which makes it very sensitive to outliers.

Figure 3 shows the accuracy of logloss estimation for a real world transfer learning experiment based on the 20-newsgroup data. Following the experimental setup of [3] we trained a classifier (logistic regression) on one 20 newsgroup classification problem and tested it on a

related problem. Specifically, we used the hierarchical category structure to generate train and testing sets with different distributions (see Figure 3 and [3] for more detail). 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 (7). The third and forth columns indicate the absolute and relative errors of $\hat{R}_n$. The unsupervised estimation of the logloss risk was very effective with relative accuracy greater than 96% and absolute error less than 0.02.

4 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 $err(\hat{\theta}_n)$. 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: a gradient descent algorithm (using numerical finite difference approximation to the gradient) and a grid search algorithm that optimizes a different dimension of $\hat{\theta}$ at each iteration. 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.

Figure 5 displays $\hat{R}_n(\hat{\theta}_n)$, $R_n(\hat{\theta}_n)$ on the training and testing sets and the test set error rate $err(\hat{\theta}_n)$ on two real world datasets: RCV1 (text documents) and MNIST (handwritten digit images) datasets using the gradient descent algorithm. Similar results were obtained using the grid search algorithm (refer [1] for more details). 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

Method	RCV1	MNIST
GMM with $\Sigma = \sigma_y^2 I$	0.3564	0.3901
GMM with Diagonal Σ	0.2083	0.3163
GMM with Regularized Σ	0.1645	0.2032
Our method	0.0923	0.1023
Supervised logistic regression	0.07	0.05

Table 1: Test error comparison of logistic regression models (USL and supervised) with Gaussian mixture model in high dimensional feature space.

data as a training set and kept the rest as a testing set.

Figure 5 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).

Table 1 shows our approach performs much better compared to Gaussian mixture model clustering in the original feature space. A likely reason is that our method works in reduced dimensionality (1 dimensional vs high dimensional) and that our 1-D normality assumption is often realistic due to the CLT (whereas assuming normality in the original high dimensional space does is much more restrictive and less realistic).

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 4 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.

5 Related Work

Related problems have been addressed in [13, 10, 5]. The work in [10] estimates labels given several datasets with different (known) label proportions. Our method
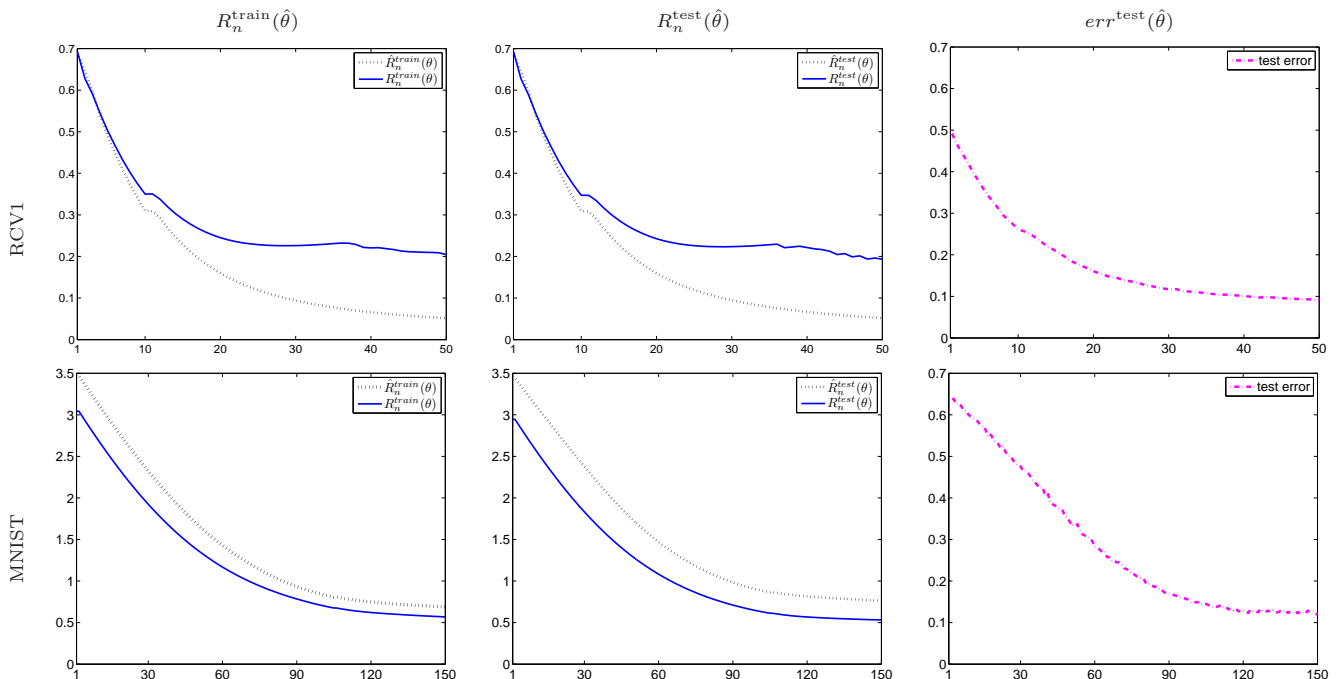
Figure 5: Performance of unsupervised logistic regression classifier $\hat{\theta}_n$ computed using gradient descent on the RCV1 (text) and MNIST (images) dataset as a function of the algorithm iteration number. The risk estimates of $\hat{\theta}_n$ were computed using the train set (left) and the test set (middle). The algorithm obtains relatively accurate classifiers whose test set error rates are 0.1 - surprisingly similar to that of supervised logistic regression for the same n (the latter has test error 0.05 for RCV1 and 0.07 for MNIST).

assumes only a single dataset with a known $p(y)$. Thus our method is applicable in many practical cases where there is single dataset with known $p(y)$. Nevertheless the two methods complement each other in interesting ways. Furthermore, as noted previously our analysis is in fact valid when only the order of label proportions is known, rather than the absolute values. Our paper is a follow up of [5] which estimates the 0-1 risk for arbitrary classifiers i.e., not necessarily linear margin based classifiers. However, [5] assumes a symmetric noise assumption which we avoid.

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 (2)). 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 shows that it performs well in practice on sev-

eral different datasets. 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 labels.

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? 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.

References

[1] K. Balasubramanian, P. Donmez, and G. Lebanon. Unsupervised supervised learning II: Training margin based classifiers without labels. *ArXiv Report 1003.0470*, 2010.

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

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

[4] J. Davidson. *Stochastic limit theory: An introduction for econometricians*. Oxford University Press, USA, 1994.

[5] 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.

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

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

[8] 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.

[9] 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.

[10] 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.

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

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

[13] L. Xu, J. Neufeld, B. Larson, and D. Schuurmans. Maximum margin clustering. In *Advances in Neural Information Processing Systems*, 2005.