LEARNING FROM A MIXTURE OF LABELED AND UNLABELED EXAMPLES WITH PARAMETRIC SIDE INFORMATION

Joel Ratsaby*and Santosh S. Venkatesh†
Department of Electrical Engineering
University of Pennsylvania
Philadelphia, PA 19104
Email: jer@ee.technion.ac.il; venkatesh@ee.upenn.edu

Abstract

We investigate the tradeoff between labeled and unlabeled sample complexities in learning a classification rule for a parametric two-class problem. In the problem considered, a sample of m labeled examples and n unlabeled examples generated from a two-class, N-variate Gaussian mixture is provided together with side information specifying the parametric form of the probability densities. The class means and *a priori* class probabilities are, however, unknown parameters. In this framework we use the maximum likelihood estimation method to estimate the unknown parameters and utilize rates of convergence of uniform strong laws to determine the tradeoff between error rate and sample complexity. In particular, we show that for the algorithm used, the misclassification probability deviates from the minimal Bayes error rate by $\mathcal{O}(N^{3/5}n^{-1/5}) + \mathcal{O}(e^{-cm})$ where N is the dimension of the feature space, m is the number of labeled examples, n is the number of unlabeled examples, and c is a positive constant.

1 INTRODUCTION

The classical problem of *learning* a classification rule can be stated as follows: patterns from classes "1" and "2" (or "states of nature") appear with probabilities $p_1 = p$ and $p_2 = 1 - p$, respectively; the pattern classes are represented by feature vectors x in a common N-dimensional Euclidean space $\mathbb{R}^N$, the patterns of class "i" distributed according to the class-conditional probability density $f_i(x)$ $(i = 1, 2)$. Labeled pairs $(x, y) \in \mathbb{R}^N \times \{1, 2\}$ are assumed generated according to the following mechanism: a pattern class (or "label") $y \in \{1, 2\}$ is first drawn randomly according to the distribution of classes $\{p_1, p_2\}$; a corresponding random feature vector $x \in \mathbb{R}^N$ is then drawn according to the class-conditional density f_y. In the supervised learning scenario, a *labeled* m-sample $\{(x_j, y_j), 1 \leq j \leq m\}$ is acquired by independent sampling from the distribution of pairs (x, y). Using the sample, the objective is to construct a decision rule which when presented with a random pattern x drawn from the mixture density

$$f(x) = p_1 f_1(x) + p_2 f_2(x)$$

produces a label which disagrees with the true class of origin by a probability P_{error} close to the minimal P_{Bayes} error rate. Formally this learning problem can be formulated in the framework of the Probably Approximately Correct (PAC) learning model (cf. [9, 10]) as follows: Given $\epsilon > 0$, $\delta > 0$, and a labeled sample of size $m = m(\epsilon, \delta)$, construct a classification rule which with confidence in excess of $1 - \delta$ has an error probability[1]

$$P_{error} = P_{Bayes}(1 + c_0\epsilon).$$

As is well known in the unsupervised learning literature (cf. [1]), an *unlabeled* teaching sample $\{x_i \in \mathbb{R}^N : 1 \leq i \leq n\}$ drawn by independent sampling from the

*Currently with the Department of Electrical Engineering, Technion, Israel.

†This research was supported by the Air Force Office of Scientific Research under grant F49620-93-1-0120.

COLT''95 Santa Cruz, CA USA© 1995 ACM 0-89723-5/95/0007..$3.50

[1]In the sequel, $c_0, c_1, c_2, \ldots$ represent positive constants independent of the error and confidence parameters ϵ and δ and the dimension of the feature space N.

mixture density $f(x)$, can also be used in the process of learning classification. If the mixture probability density function

$$f(x \mid \theta) = p_1 f_1(x \mid \theta_1) + p_2 f_2(x \mid \theta_2)$$

is parametric (with unknown parameters $\theta = [\theta_1, \theta_2]$ and p) and is identifiable (cf. Teicher [8]) then the unlabeled sample $\{x_i\}$ can be used to generate an identifiable density function $f(x \mid \hat{\theta})$ as an estimate of $f(x \mid \theta)$. Identifiability ensures that the Bayes optimal decision border $\{x : p_1 f_1(x \mid \theta_1) = p_2 f_2(x \mid \theta_2)\}$ can be deduced if $f(x \mid \theta)$ is known and therefore one can construct an estimate of the Bayes border by using $f(x \mid \hat{\theta})$ instead of $f(x \mid \theta)$ (such plug-in rules are asymptotically optimal cf. Glick [11]). Once the decision border is estimated, a *small* labeled sample suffices to learn (with high confidence and small error) the appropriate class labels $l_1, l_2 \in \{1, 2\}$ associated with the two disjoint decision regions in $\mathbb{R}^N$ generated by the estimate of the Bayes decision border.

Clearly, unlabeled examples, albeit of less worth than labeled examples, can still carry information of value to the learner. Furthermore, in many cases of practical interest, unlabeled examples exist in profusion in nature (or are inexpensively acquired) while their labeled counterparts are few in number (or are expensive to acquire). For example, in a two-class tree recognition problem "Douglas Fir" and "Spruce," unlabeled examples of these trees may exist in profusion in a forest but labeling them requires the services of a human expert who charges by the hour. Or, in an (idealized) cancer diagnosis scenario, obtaining X-rays of cells may be relatively cheap compared to the cost of the expert labeling them. A mixed sample learning system utilizing both labeled and unlabeled examples may hence be of interest where one trades off expensive labeled examples for (very many) unlabeled examples. In such a scenario, the learner would like to determine the exact tradeoff between labeled and unlabeled examples. This question was first posed by T. M. Cover [4] in the following succinct but evocative form: *How many unlabeled examples is each labeled example worth?* In recent work Castelli and Cover [2] have shown that if an infinity of unlabeled examples is available then for identifiable classes with p, f_1 and f_2 unknown, the probability of error decreases exponentially fast in the number of labeled examples to the Bayes risk; in related work [3] they have also shown that if only a finite number of labeled and unlabeled examples are available but the class-conditional densities f_1 and f_2 are known with the only unknown being the mixing parameter p, then under suitable regularity conditions labeled samples are exponentially more valuable than unlabeled samples.

In this paper we consider a parametric situation where the class-conditional densities $f_i(x \mid \theta_i)$ $(i = 1, 2)$ are specified by a parameter (vector) $\theta_i \in \mathbb{R}^N$. We assume that the learner is provided with a finite mixed sample of m labeled examples and n unlabeled examples together with side-information specifying the parametric *form* of the class-conditional densities (but not the densities themselves); the parameter vector $\theta = [\theta_1, \theta_2]$ as well as the mixing parameter p are, however, unknown to the learner. Our goal is to determine how the error rate depends on the sample sizes m and n and on the dimensionality N. The results are not immediate even for the archtypal problem of two N-dimensional Gaussian distributed pattern classes and for definiteness we will focus on the case of Gaussian mixtures in this paper. As we will see in the proof of the main theorem, however, the analysis techniques pertain to other parametric families as well though the rates and constants depend upon the family under consideration.

For conciseness, we limit our discussion in the main body of the paper to a mixture of multi-dimensional Gaussians with unit covariance matrices

$$f_i(x \mid \theta_{0i}) = (2\pi)^{-N/2} e^{-\frac{1}{2}(x-\theta_{0i})^T (x-\theta_{0i})}, \qquad i = 1, 2.$$

To keep things simple we sketch the analysis for the case where the mixing parameter p is identically $1/2$ and is known to the learner. The mixture density is then just

$$f(x \mid \theta_0) = \tfrac{1}{2} f_1(x \mid \theta_{01}) + \tfrac{1}{2} f_2(x \mid \theta_{02}).$$

We will suppose that the learner knows the form of f but does not know the true parameter vector $\theta_0 = [\theta_{01}, \theta_{02}] \in \mathbb{R}^{2N}$. The results for the general case where the mixing parameter $p \in [0, 1]$ is also unknown to the learner are summarized at the end of the paper.

We begin with the purely labeled case where all the examples are labeled: $m < \infty$, $n = 0$. The results for this case provide a useful point of comparison for the mixed sample case.

2 PURELY LABELED SAMPLE

Given a sample $\{(\xi_j, y_j) \in \mathbb{R}^N \times \{1, 2\} : 1 \le j \le m\}$ of m labeled examples obtained by independent sampling from the marginal distribution on pairs (x, y), it is easy to see that in the lack of any further information, moment estimation yields optimal results.

Algorithm E For $i = 1, 2$, let

$$\{\xi^i_{j_k} : y^i_{j_k} = i, 1 \le k \le m_i\}$$

be the subsequence of examples generated by class "i." (Clearly, $m_1 + m_2 = m$.) Estimate the unknown mean vector θ_{0i} of class "i" by

$$\hat{\theta}_i = \frac{1}{m_i} \sum_{k=1}^{m_i} \xi^i_{j_k} \qquad (i = 1, 2),$$

and consider the dichotomy of $\mathbb{R}^N$ induced by the hyperplane which is orthogonal to the vector $\hat{\theta}_2 - \hat{\theta}_1$ and passes through the point $(\hat{\theta}_1 + \hat{\theta}_2)/2$. (This separating hyperplane is just the empirical estimate of the Bayes decision border.) Label points in the resulting two regions R_1 and R_2 partitioning $\mathbb{R}^N$ by the subscript "i" of the mean estimate $\hat{\theta}_i$ which they contain.

Simple Chernoff bounds now suffice to prove the following assertion. The proof is omitted.

Theorem 1 *Suppose two equiprobable pattern classes are distributed according to N-variate Gaussian probability densities $f_1(x \mid \theta_{01})$, $f_2(x \mid \theta_{02})$, both with unit covariance matrices and unknown means θ_{01}, $\theta_{02} \in \mathbb{R}^N$. Then, for sufficiently small $\epsilon > 0$ and arbitrary $\delta > 0$, given*

$$m = \frac{4N}{\epsilon} \log\left(\frac{8N}{\delta}\right)$$

labeled examples and $n = 0$ unlabeled examples, with confidence at least $1 - \delta$, Algorithm E results in a decision rule with classification error bounded by

$$P_{\text{error}} \leq P_{\text{Bayes}}(1 + c_0\epsilon),$$

where $c_0 > 0$ is a constant depending only on the distance between the means.

For a given labeled sample size, the minimal error rate varies (roughly) as the inverse of the sample size; the deleterious effects of dimensionality are evidenced in the $N \log N$ term.

We now turn to the mixed sample case and our main result.

3 MIXED SAMPLE

Let us now consider the effect of introducing n unlabeled examples; the reduction in the size m of the labeled sample needed to obtain a given error rate will yield a rough measure of the worth of each labeled example in terms of the number of unlabeled examples that are needed to compensate for it.

Recall that the class conditional densities $f_i(x \mid \theta_{0i})$ $(i = 1, 2)$ are N-dimensional Gaussians with unit covariance matrices and unknown mean vectors $\theta_{0i} \in \mathbb{R}^N$. We assume in addition that the unknown parameter vector $\theta_0 = [\theta_{01}, \theta_{02}]$ lies in a compact set $\Theta \subset \mathbb{R}^{2N}$. Our approach will be to utilize the unlabeled sample to estimate the unknown parameter vector θ_0, hence also the mixture density $f(x \mid \theta_0) = \frac{1}{2}f_1(x \mid \theta_{01}) + \frac{1}{2}f_1(x \mid \theta_{01})$, from which we can then obtain an estimate of the hyperplane which constitutes the Bayes decision border in this case. The labeled sample will then be utilized to provide a labeling of the two disjoint regions in $\mathbb{R}^N$ induced by the separating hyperplane.

Algorithm M Let $\{x_i, 1 \leq i \leq n\}$ be the unlabeled n-sample obtained by independent sampling from the mixture density $f(x \mid \theta_0)$, and let $\{(\xi_j, y_j), 1 \leq j \leq m\}$ denote the labeled m-sample obtained by independent sampling from the distribution of pairs (x, y). The algorithm proceeds in two steps: (1) Estimate the unknown parameter vector θ_0 by the maximum likelihood estimate

$$\hat{\theta} = \arg\sup_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^{n} \log f(x_i \mid \theta),$$

and select as a separating surface the hyperplane that passes through the point $(\hat{\theta}_1 + \hat{\theta}_2)/2$ and is orthogonal to the vector $\hat{\theta}_2 - \hat{\theta}_1$. (2) Label each of the two decision regions separated by the hyperplane according to the majority of the labeled examples in the region.

The following assertion is our main result.

Theorem 2 *Suppose two equiprobable pattern classes are distributed according to N-variate Gaussian probability densities $f_1(x \mid \theta_{01})$, $f_2(x \mid \theta_{02})$, both with unit covariance matrices and unknown means θ_{01}, $\theta_{02} \in \mathbb{R}^N$, and suppose that the unknown parameter vector $\theta_0 = [\theta_{01}, \theta_{02}]$ lies in a compact subset Θ of $\mathbb{R}^{2N}$. Then for sufficiently small $\epsilon > 0$ and arbitrary $\delta > 0$, given*

$$m = c_1 \log \delta^{-1}$$

labeled examples and

$$n = \frac{c_2 N^2}{\epsilon^3 \delta}\left(N \log \epsilon^{-1} + \log \delta^{-1}\right)$$

unlabeled examples, with confidence in excess of $1 - \delta$, Algorithm M determines a decision rule with classification error bounded by

$$P_{\text{error}} \leq P_{\text{Bayes}}(1 + c_3\epsilon).$$

In the above, c_1, c_2, and c_3 are positive constants which are independent of ϵ, δ, N, and the choice of $\theta_0 \in \Theta$.

The main technical tool needed in the proof is a rate of convergence in a uniform strong law for bounded functions (cf. Pollard [6, Theorem II.6.37] and Haussler [5]). The messy technical considerations involve truncation arguments to exploit the exponential decay of the Gaussian to massage the problem into a form where the uniform strong law is applicable.

4 SKETCH OF PROOF

In the following, all expectations are taken with respect to $f(x \mid \theta_0)$. We omit technical details in the proofs for brevity.

We first use the uniform strong law to show that using the n unlabeled examples it is possible to estimate the

true parameter θ_0 uniformly over the parameter space Θ by its maximum likelihood estimate $\hat{\theta}$ to within small deviation provided n is large enough. This implies that the empirical decision boundary is close to the Bayes optimal separating hyperplane. We then calculate the number m of labeled examples needed to guarantee the correct labeling of the decision regions with high confidence.

4.1 The Likelihood Function

Define the usual likelihood function

$$L(\theta) = \sum_{i=1}^{n} \log f(x_i|\theta)$$

where x_i are the unlabeled examples. The learner calculates the value of θ which achieves the global maximum of $L(\theta)$; call it $\hat{\theta}$. We now proceed to show that $\hat{\theta}$ is ϵ-close to θ_0 for a sufficiently large choice of n. We first determine a value of n that suffices to guarantee the existence of a (possibly relative) maximum of $L(\theta)$ inside the closed ϵ-ball at θ_0 (denoted by $B(\theta_0, \epsilon)$). In other words, for such a choice of n we show that there exists some $\theta_a \in B(\theta_0, \epsilon)$ such that $L(\theta_a) \geq L(\theta_0)$. We then show that for all θ outside this ball, $L(\theta) < L(\theta_0)$ with high confidence. This will imply that by picking the global maximum of $L(\theta)$ the learner chooses a hypothesis $\hat{\theta}$ which is ϵ-close to θ_0.[2]

As $L(\cdot)$ is continuous in θ, it will have a (relative) maximum inside $B(\theta_0, \epsilon)$ if $L(\theta_\epsilon) < L(\theta_0)$ for all points $\theta_\epsilon \in \partial B(\theta_0, \epsilon)$, the surface of the ball of radius ϵ at θ_0. It will transpire that this is all we need to show that in fact $L(\theta)$ achieves its global maximum in the ball $B(\theta_0, \epsilon)$. As a first step we hence show that

$$\sum_{i=1}^{n} \log \frac{f(x_i \mid \theta_\epsilon)}{f(x_i \mid \theta_0)} < 0, \qquad \theta_\epsilon \in \partial B(\theta_0, \epsilon). \tag{1}$$

We take the usual tack of comparing the empirical average

$$\frac{1}{n} \sum_{i=1}^{n} \log \frac{f(x_i \mid \theta_\epsilon)}{f(x_i \mid \theta_0)}$$

with the ensemble average

$$\mathbb{E} \log \frac{f(x \mid \theta_\epsilon)}{f(x \mid \theta_0)}.$$

[2]More accurately, $\hat{\theta}$ will be close to either $\theta_0 = [\theta_{01}, \theta_{02}]$ or its permutation $[\theta_{02}, \theta_{01}]$. This assertion follows from Theorem 2 and Proposition 1 of Teicher [8] and Proposition 2 of Yakowitz [12] which together imply that mixtures of N-dimensional Gaussians are identifiable, whence θ_0 is uniquely determined up to permutations of θ_{01} and θ_{02}. Such permutations create no particular difficulties here as the decision border is unaffected by permutations of the class labels. We will keep the notation simple by ignoring the nuisance permutation.

Note that the latter quantity is well-defined and strictly less than zero as $\theta_\epsilon \neq \theta_0$.

4.2 Extremum of the Likelihood Function

Suppose we can show that for a suitable choice of $\alpha > 0$ (possibly depending on ϵ), the inequality

$$\sup_{\theta \in B(\theta_0, \epsilon)} \left| \frac{1}{n} \sum_{i=1}^{n} \log f(x_i \mid \theta) - \mathbb{E} \log f(x \mid \theta) \right| \leq \frac{\alpha}{2} \tag{2}$$

holds with high confidence for a sufficiently large choice of n. As special cases, (2) implies the pair of inequalities:

$$\sup_{\theta_\epsilon \in \partial B(\theta_0, \epsilon)} \left| \frac{1}{n} \sum_{i=1}^{n} \log f(x_i \mid \theta_\epsilon) - \mathbb{E} \log f(x \mid \theta_\epsilon) \right| \leq \frac{\alpha}{2},$$

$$\left| \frac{1}{n} \sum_{i=1}^{n} \log f(x_i \mid \theta_0) - \mathbb{E} \log f(x \mid \theta_0) \right| \leq \frac{\alpha}{2}.$$

These inequalities together imply that

$$\sup_{\theta_\epsilon} \left| \frac{1}{n} \sum_{i=1}^{n} \log \frac{f(x_i \mid \theta_\epsilon)}{f(x_i \mid \theta_0)} - \mathbb{E} \log \frac{f(x \mid \theta_\epsilon)}{f(x \mid \theta_0)} \right| \leq \alpha.$$

For (1) to hold it hence suffices if the inequality (2) holds for a choice

$$\alpha(\epsilon) = \inf_{\theta_\epsilon} \left| \mathbb{E} \log \frac{f(x \mid \theta_\epsilon)}{f(x \mid \theta_0)} \right|.$$

As a first step we estimate the dependence of $\alpha(\epsilon)$ on ϵ and in particular show that $\alpha(\epsilon)$ is not identically zero for any fixed $\epsilon > 0$. We then use a uniform strong law of large numbers in conjunction with truncation arguments to show that (2) holds.

For small $\epsilon > 0$ expand $\mathbb{E} \log \frac{f(x|\theta_\epsilon)}{f(x|\theta_0)}$ in a 2N-dimensional Taylor series around θ_0: first write

$$\mathbb{E} \log \frac{f(x \mid \theta_\epsilon)}{f(x \mid \theta_0)} = \int f(x \mid \theta_0) \log f(x \mid \theta_\epsilon)\, dx - \int f(x \mid \theta_0) \log f(x \mid \theta_0)\, dx.$$

Adopt the nonce notations $D_i \triangleq \frac{\partial}{\partial \theta_{0i}}$, $D_{ij} \triangleq \frac{\partial^2}{\partial \theta_{0i} \partial \theta_{0j}}$ and $D_{ijk} \triangleq \frac{\partial^3}{\partial \theta_{0i} \partial \theta_{0j} \partial \theta_{0k}}$. Writing $\theta_\epsilon - \theta_0 = \underline{\epsilon} = (\epsilon_1, \ldots, \epsilon_N)^T$, where $\underline{\epsilon}$ has length ϵ, the first term becomes

$$\int f(x \mid \theta_0) \log f(x \mid \theta_\epsilon)\, dx = \int f(x \mid \theta_0) \log f(x \mid \theta_0)\, dx + \sum_{i=1}^{2N} \epsilon_i \int f(x \mid \theta_0) D_i \log f(x \mid \theta_0)\, dx$$

$$+\frac{1}{2}\sum_{i,j=1}^{2N}\epsilon_i\epsilon_j\int f(x\mid\theta_0)D_{ij}\log f(x\mid\theta_0)\,dx$$
$$+\frac{1}{6}\sum_{i,j,k=1}^{2N}\epsilon_i\epsilon_j\epsilon_k\int f(x\mid\theta_0)D_{ijk}\log f(x\mid\theta_{\epsilon'})\,dx, \tag{3}$$

where $\underline{\epsilon}'$ is on the line joining 0 and $\underline{\epsilon}$.

The various terms in (3) can be evaluated to show that $\alpha(\epsilon) = \Omega(\epsilon^2)$ $(\epsilon \to 0)$. More precisely, there exist positive constants c_4 and c_5 independent of N and ϵ such that

$$\alpha(\epsilon) = \inf_{\theta_\epsilon}\left|\mathbb{E}\log\frac{f(x\mid\theta_\epsilon)}{f(x\mid\theta_0)}\right| \geq c_4\epsilon^2 + c_5\epsilon^3.$$

We now estimate the unlabeled sample size n needed to guarantee (2). Define a function class $\mathcal{G}$ as follows: let $D \subset \mathbb{R}^N$ be a compact subset of the probability one support of $f(x)$ and denote its complement by D^c. Let

$$\mathcal{G} \triangleq \{\log f(x\mid\theta)1_D(x) : \theta \in B(\theta_0,\epsilon)\},$$

where 1_D is the indicator function for the set D. We now have

$$\begin{aligned}
&\mathbb{P}\left(\sup_{\theta\in B(\theta_0,\epsilon)}\left|\mathbb{E}\log f(x\mid\theta) - \frac{1}{n}\sum_{i=1}^{n}\log f(x_i\mid\theta)\right| > \frac{\alpha}{2}\right)\\
&\quad= \mathbb{P}\Bigg(\sup_{\theta\in B(\theta_0,\epsilon)}\Bigg|\int_D f(x\mid\theta_0)\log f(x\mid\theta)\,dx\\
&\qquad\qquad - \frac{1}{n}\sum_{i=1}^{n}\log f(x_i\mid\theta)1_D(x_i)\\
&\qquad\qquad + \int_{D^c} f(x\mid\theta_0)\log f(x\mid\theta)\,dx\\
&\qquad\qquad - \frac{1}{n}\sum_{i=1}^{n}\log f(x_i\mid\theta)1_{D^c}(x_i)\Bigg| > \frac{\alpha}{2}\Bigg)\\
&\quad\leq \mathbb{P}\Bigg(\sup_{\theta\in B(\theta_0,\epsilon)}\Bigg|\int_D f(x\mid\theta_0)\log f(x\mid\theta)\,dx\\
&\qquad\qquad - \frac{1}{n}\sum_{i=1}^{n}\log f(x_i\mid\theta)1_D(x_i)\Bigg| > \frac{\alpha}{4}\Bigg)\\
&\qquad + \mathbb{P}\Bigg(\sup_{\theta\in B(\theta_0,\epsilon)}\Bigg|\int_{D^c} f(x\mid\theta_0)\log f(x\mid\theta)\,dx\\
&\qquad\qquad - \frac{1}{n}\sum_{i=1}^{n}\log f(x_i\mid\theta)1_{D^c}(x_i)\Bigg| > \frac{\alpha}{4}\Bigg).
\end{aligned} \tag{4}$$

The first term on the right-hand side is the probability of uniform convergence for functions in $\mathcal{G}$. As the functions in $\mathcal{G}$ are bounded, the uniform strong law [6, Theorem II.6.37] can be used here to make the first term arbitrary small by choosing n sufficiently large once suitable covering numbers have been evaluated. The uniform strong law cannot be directly applied to the second term since $\log f(x\mid\theta)$ is unbounded over D^c. However this term can be made arbitrary small for a proper choice of region D by utilizing the rapid decay of the Gaussian outside a large enough sphere centered at the mean. Omitting the messy details, we find that the right-hand side of (4) is bounded above by $\delta/2$ provided

$$n \geq \frac{c_6N^2}{\epsilon^6\delta}\left(N\log\epsilon^{-1} + \log\delta^{-1}\right). \tag{5}$$

Consequently, for such a choice of n, there exists a (relative) maximum of $L(\theta)$ inside the ball $B(\theta_0,\epsilon)$ with probability at least $1-\delta/2$.

The rapid decay of the Gaussian mixture will allow us to conclude by arguments similar to (if slightly messier) than those above that, in fact, the maximum likelihood estimate $\hat{\theta}$ is ϵ-close to the true parameter vector θ_0 (again, up to permutations). Chernoff bounding arguments now readily show that the classification error of the hypothesis (under optimal labeling of the decision regions) is bounded above by $P_{error} \leq P_{Bayes}(1+c_7\epsilon^2)$. Replacing ϵ^2 by ϵ here and in (5), we obtain the unlabeled sample complexity estimate of the theorem. It only remains to determine the sample complexity m of the labeled examples requisite to label the decision regions optimally with high confidence.

4.3 Labeling the Partition

We have two unlabeled regions separated by the hyperplane between $\hat{\theta}_1$ and $\hat{\theta}_2$ where both are ϵ-close to their respective true parameters θ_{01} and θ_{02}. The probability that the majority label of the labeled examples lying in any one of these regions does not agree with the optimal labeling of the region decreases exponentially fast in n (the binomial tail). The labeled sample estimate in the theorem follows readily. This concludes the sketch of the proof of the theorem.

5 CONCLUSIONS

For the mixture Gaussian problem, a classification error probability of $P_{Bayes}(1+\mathcal{O}(\epsilon))$ can be obtained with confidence in excess of $1-\delta$ given a random mixed sample of

$$n_M = \frac{c_8N^2}{\epsilon^3\delta}\left(N\log\epsilon^{-1} + \log\delta^{-1}\right)$$

unlabeled examples and

$$m_M = c_9\log\delta^{-1}$$

labeled examples. Compare this with the purely-labeled sample scenario of Section 2 where a learner using Al-

gorithm E requires

$$m_E = \frac{4N}{\epsilon} \log\left(\frac{8N}{\delta}\right)$$

labeled examples to achieve the same learning performance. Clearly, the introduction of unlabeled examples has reduced the demands on the number of labeled examples. As a rough measure one may take the ratio

$$\frac{n_M}{(m_E - m_M)} = \frac{c_{10}N^2}{\epsilon^3 \delta \log N}$$

as indicative of the number of unlabeled examples needed to compensate for the removal of each labeled example. In this sense, the tradeoff is polynomial in N, $\frac{1}{\epsilon}$, and $\frac{1}{\delta}$. Note that, as claimed in the abstract, inverting the sample complexity estimates for the mixed sample case shows that the error rate contribution from the labeled examples decreases exponentially fast in m, while the error rate contribution from the unlabeled examples decreases only as an inverse polynomial in n. Note also that the efficacy of the unlabeled examples in the algorithmic scenarios discussed here diminishes as the dimensionality N increases. More explicitly, the classification error probability under Algorithm M deviates from the Bayes optimal error rate by $\mathcal{O}(N^{3/5}n^{-1/5}) + \mathcal{O}(e^{-cm})$.

Note that the analysis will work for other parametric families as long as the tails of the distributions are not too heavy. In particular, the analysis works *in toto* for other members of the exponential family,, albeit with different constants.

The case of a general unknown mixing parameter p is also easily handled in this framework. Two modifications to the approaches are indicated. In the first approach, the labeled examples can be used to directly estimate p according to the relative frequency of class "1;" Chernoff bounds show that very good estimates of p can be obtained from a small labeled sample. In the second approach, p is treated as an additional unknown parameter and maximum likelihood estimation is used to estimate p as well as the parameter vector θ_0. If p is bounded away from 0 and 1, a similar analysis shows that each labeled example has to be replaced by of the order of

$$\frac{c_{11}N^2 \log \epsilon^{-1}}{\epsilon^4 p^{11} \log N}$$

unlabeled examples to achieve the same performance.

In nonparametric cases, the tradeoff between labeled and unlabeled examples is much more pronounced. For instance, if Kernel density estimation is used to infer the mixture density of an identifiable family of bimodal mixture densities (including Gaussian mixtures), each labeled example is worth roughly

$$\frac{c_{12}13^{N \log(5+\log N)}}{N \log N \epsilon^{\frac{N}{2 \log N}}}$$

unlabeled examples [7].

References

[1] R. O. Duda and P. E. Hart, *Pattern Classification and Scene Analysis*, John Wiley & Sons, New York, 1973.

[2] V. Castelli and T. M. Cover, "Classification rules in the unknown mixture parameter case: relative value of labeled and unlabeled examples," *Proc. 1994 IEEE Int. Symp. Inform. Theory*, p. 111, Trondheim, Norway, 1994.

[3] V. Castelli and T. M. Cover, "On the exponential value of labeled samples," to appear *Pattern Recognition Letters.*

[4] T. M. Cover, "Learning and generalization," in *Proc. 4th Annual Workshop on Computational Learning Theory*, (eds. L. G. Valiant and M. K. Warmuth), p. 3, Morgan Kaufmann, San Mateo, California, 1991.

[5] D. Haussler, "Decision theoretic generalizations of the PAC model for neural net and other learning applications," *Technical Report: University of California, Santa Cruz*, UCSC–CRL–91–02, 1989.

[6] D. Pollard, *Convergence of Stochastic Processes*, Springer Verlag, New York, 1984.

[7] J. Ratsaby, *The Complexity of Learning from a Mixture of Labeled and Unlabeled Examples*, Ph.D. Thesis, University of Pennsylvania, 1994.

[8] H. Teicher, "Identifiability of finite mixtures," *Annals of Mathematical Statistics*, vol. 34, pp. 1265–1269, 1963.

[9] L. G. Valiant, "A Theory of the learnable," *Comm. ACM*, vol. 27, no. 11, pp. 1134–1142, 1984.

[10] A. Blumer, A. Ehrenfeucht, D. Haussler, and M. Warmuth, "Learnability and the Vapnik-Chervonenkis dimension," *JACM*, vol. 36, no. 4, pp. 929–965, 1989.

[11] N. Glick, "Sample-based classification procedures derived from density estimators," *J. American Statistical Association*, vol. 67, 1972.

[12] S. J. Yakowitz and J. D. Spragins, "On identifiability of finite mixtures," *Annals of Mathematical Statistics*, vol. 39, pp. 209–214, 1968.