ISCAS 2000 - IEEE International Symposium on Circuits and Systems, May 28-31, 2000, Geneva, Switzerland

On Partially Blind Learning Complexity

*Joel Ratsaby and Santosh S. Venkatesh**

Abstract

We call a learning environment partially blind when there is an admixture of supervised and unsupervised (or blind) learning. Such situations typically arise in practice when supervised training data labelled by a teacher are scarce or expensive and are supplemented by inexpensive unlabelled (or blind) data available in relative profusion. Vapnik-Červonenkis theory can be deployed in such settings to quantify the relative worth of supervision (and the lack thereof) in learning. We illustrate the nature of the tradeo s possible in a simple setting of hyperplane decision functions and make explicit the role of dimensionality and side-information in these tradeo s in the context of d-variate Gaussian mixtures.

1 Introduction

Blind (or unsupervised) learning in the absence of (supervised) training examples is gaining importance in divers areas in communications, signal processing, and pattern recognition. In these settings, we call a learning environment partially blind if the learner has additional access to a small amount of labelled training data. Such situations typically occur in practice when unlabelled (or blind) data exist in profusion in nature (or are inexpensively acquired) while their labelled counterparts are few in number (or are expensive to acquire). For example, in a two-class tree recognition problem "Douglas Fir" and "Spruce," unlabelled examples of these trees may exist in profusion in a forest but labelling them requires the services of a human expert who charges by the hour. Or, in an (idealised) cancer diagnosis scenario, obtaining X-rays of cells may be relatively cheap compared to the cost of the expert labelling them. A mixed sample learning system utilising both labelled and unlabelled examples may hence be of interest where one trades o expensive labelled examples for (very many) unlabelled examples.

In such settings, one would like to be able to quantify the relative worth of supervision (and the lack thereof) in learning. This question was first posed by T. M. Cover [1, 2, 3] in the following succinct but evocative form: How many unlabelled examples is each labelled example worth? We are interested, in particular, in the impact of dimensionality on the tradeo .

In the sequel we focus on a simple setting of hyperplane decision functions to illustrate the precise tradeo s that may be determined.

Consider the problem of learning a d-variate decision function of the form

$$h^*(x) = \begin{cases} 1 & \text{if } (w^*, x) \geq w_0^*, \\ 0 & \text{if } (w^*, x) < w_0^*, \end{cases}$$

where $w^* = (w_1^*, \ldots, w_d^*) \in \mathbb{R}^d$ is a d-dimensional parameter vector, $w_0^* \in \mathbb{R}$ is a real threshold, and $(w^*, x) = \sum_{i=1}^d w_i^* x_i$ is the usual inner product. The decision function h engenders a hyperplane in d-dimensions which dichotomises $\mathbb{R}^d$ into two disjoint sets: the half-space of positive examples for which $h(x) = 1$ and the remaining half-space of negative examples for which $h(x) = 0$. Let $\theta_1^* \in \mathbb{R}^d$ and $\theta_0^* \in \mathbb{R}^d$ be exemplars of positive and negative examples, respectively, on either side of the hyperplane and equidistant from and orthogonal to it. Equivalently, the separating hyperplane is orthogonal to the vector $\theta_1^* - \theta_0^*$ and passes through the point $(\theta_1^* + \theta_0^*)/2$.

The learner is provided with an i.i.d. sequence of noisy observations of the form $x = \theta + z$ where θ is chosen with equal probability $1/2$ from the exemplars $\{\theta_0^*, \theta_1^*\}$ and z is a zero-mean, d-variate Gaussian vector with unit covariance matrix. For $i = 0, 1$, let $\varphi(x \mid \theta_i^*)$ denote a d-variate Gaussian density with mean θ_i^* and unit covariance matrix. Then the noisy observations conform to a mixture Gaussian density $f(x \mid \theta) = \frac{1}{2}\varphi(x \mid \theta_0^*) + \frac{1}{2}\varphi(x \mid \theta_1^*)$, where $\theta = [\theta_0^*, \theta_1^*]$ is the $2d$-dimensional parameter vector of means.

The data are of two types: (1) labelled (supervised) data $\{(\xi_j, y_j), 1 \leq j \leq m\}$ where, for each j, $y_j \in \{0, 1\}$ is the label of the chosen exemplar and $\xi_j \sim \varphi(x \mid \theta_{y_j}^*)$ is drawn from the

Corresponding author: Santosh S. Venkatesh. Address: Department of Electrical Engineering, University of Pennsylvania, Philadelphia, PA 19104, USA. Email: venkatesh@ee.upenn.edu.

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

0-7803-5482-6/00/$10.00©2000IEEE

distribution of exemplars of class y_j; and (2) unlabelled (blind) data $\{x_i, 1 \le i \le n\}$ where each x_i is drawn by independent sampling from the mixture density $f(x \mid \theta^*)$. We generate a random hypothesis decision function $\hat{h}(x) = \hat{h}_{m,n}(x)$ as a function of the data. Our goal is to estimate the error probability $P_{m,n} = Pr\{\hat{h} \neq h^*\}$, i.e., the probability that the hypothesis and the target decision function disagree on a randomly chosen $x \sim f(\cdot \mid \theta^*)$. More specifically, for a given choice of algorithm, we wish to determine sample complexities $m = m(\epsilon, \delta, d)$ and $n = n(\epsilon, \delta, d)$ for which

$$Pr\{P_{m,n} \ge \epsilon\} \le \delta$$

for sufficiently small ϵ and δ. If the above holds, we say that $P_{mn} < \epsilon$ with confidence at least $1 - \delta$.

It should be clear that the setting here can be quickly generalised to other decision functions, and classification and hypothesis testing settings.

2 Empirical Mean Estimate

Given a sufficiency of labelled examples $\{(\xi_j, y_j) \in \mathbb{R}^d \times \{0,1\} : 1 \le j \le m\}$ obtained by independent sampling from the marginal distribution on pairs (x, y), it is easy to see that moment estimation yields optimal results.

Algorithm E *For $i = 0, 1$, let*

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

be the subsequence of examples generated by exemplar θ^*_i. (Clearly, $m_1 + m_2 = m$.)

E1. *Estimate the unknown mean vector θ^*_i by*

$$\hat{\theta}_i = \frac{1}{m_i} \sum_{k=1}^{m_i} \xi^i_{j_k} \qquad (i = 0, 1).$$

E2. *Select as separating surface the hyperplane which is orthogonal to the vector $\hat{\theta}_1 - \hat{\theta}_0$ and passes through the point $(\hat{\theta}_1 + \hat{\theta}_0)/2$.*

E3. *For each $i = 0, 1$, let $\mathcal{R}_i$ denote the half-space containing the mean estimate $\hat{\theta}_i$. Set*

$$\hat{h}(x) = \begin{cases} 1 & \text{if } x \in \mathcal{R}_1, \\ 0 & \text{if } x \in \mathcal{R}_0. \end{cases}$$

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

Theorem 1 *For sufficiently small $\epsilon > 0$ and arbitrary $\delta > 0$, given*

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

labelled examples (and $n = 0$ unlabelled examples), with confidence at least $1 - \delta$, Algorithm E results in a decision rule with error bounded by $P_{m,0} < c_0\epsilon$, where $c_0 > 0$ is a constant depending only on the distance between the means.

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

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

3 Maximum Likelihood Estimate

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

Suppose that it is known that the data are engendered by a Gaussian mixture of the form $f(x \mid \theta) = \frac{1}{2}\varphi(x \mid \theta_0) + \frac{1}{2}\varphi(x \mid \theta_1)$ but that the true parameter vector $\theta^* = [\theta^*_0, \theta^*_1]$ is unknown. We assume in addition that θ^* lies in a compact set $\Theta \subset \mathbb{R}^{2d}$. Our approach will be to utilise the unlabelled sample to obtain an estimate $\hat{\theta}$ of the unknown parameter vector θ^*, from which we can then obtain an estimate of the separating hyperplane. The labelled sample will then be utilised to provide a labelling of the two disjoint regions in $\mathbb{R}^d$ induced by the separating hyperplane.

Algorithm M *Let $\{x_i, 1 \le i \le n\}$ be the unlabelled n-sample, and let $\{(\xi_j, y_j), 1 \le j \le m\}$ denote the labelled m-sample.*

M1. *Estimate the unknown parameter vector θ^* by the maximum likelihood estimate*

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

M2. *Select as separating surface the hyperplane that passes through the point $(\hat{\theta}_1 + \hat{\theta}_0)/2$ and is orthogonal to the vector $\hat{\theta}_1 - \hat{\theta}_0$.*

M3. *Label each of the two decision regions separated by the hyperplane according to the majority of the labelled examples in the region.*

The following is indicative of the kind of tradeo *between labelled and unlabelled examples that is possible in this setting.*

Theorem 2 *For su ciently small* $\epsilon > 0$ *and arbitrary* $\delta > 0$, *given*

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

labelled examples and

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

unlabelled examples, with confidence in excess of $1-\delta$, *Algorithm M determines a decision rule with error bounded by* $P_{m,n} < c_3\epsilon$. *In the above,* c_1, c_2, *and* c_3 *are positive constants which are independent of* ϵ, δ, d, *and the choice of* $\theta^* \in \Theta$.

The main technical tool needed in the proof is a rate of convergence in a uniform strong law for bounded functions (cf. Pollard [5, Theorem II.6.37] and Haussler [4]). 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 Kernel Density Estimate

Under weaker side information, the sample complexity demands increase dramatically. Consider, for instance, a rich family of distributions (which include Gaussian mixtures) for which the separating hyperplane is identified solely by the modes of the mixture. The approach here uses kernel estimation combined with a mode-estimation algorithm with which the Bayes decision rule is estimated. We briefly sketch the algorithm. (For the technical details pertaining to the problem class and the algorithm, see [6].)

Algorithm K *Let* $\{x_i, 1 \le i \le n\}$ *be the unlabelled* n*-sample and let* $\{(\xi_j, y_j), 1 \le j \le m\}$ *denote the labelled* m*-sample.*

K1. *Use kernel estimation to obtain the mixture density estimate* f_n *from the unlabelled sample. (It is not necessary for* f_n *to be a bona fide density.)*

K2. *Determine a set of modes of* f_n *which are consistent estimators of the modes of the underlying mixture* f.

K3. *Let the decision border be a hyperplane which passes through the average of the mode estimates and is perpendicular to their least-square line. (In the Gaussian case there are two mode estimates.)*

K4. *Label the two decision regions induced by the hyperplane by the label of the majority of the examples in each region.*

The following sample complexity estimates for Algorithm K also arise by careful application of uniform strong laws.

Theorem 3 *For su ciently small* $\epsilon > 0$ *and arbitrary* $\delta > 0$, *given*

$$n = \frac{c_4 13^{d \log d + d \log 5}}{\epsilon^{\frac{d}{2 \log d}}} \log\left(\epsilon^{-1}\delta^{-1}\right)$$

unlabelled examples and

$$m = c_5 \log \delta^{-1}$$

labelled examples, Algorithm K determines a decision rule with error $P_{m,n} < c_6\epsilon$ *with confidence at least* $1 - \delta$, *where* c_4, c_5 *and* c_6 *are absolute positive constants independent of the choice of,* δ, *and* $\theta^* \in \Theta$.

Remark: *Based on Algorithm K, a more general statement can in fact be made pertaining to every problem in a regular family of mixtures whose modes identify the Bayes border. Indeed, the sample complexities specified in Theorem 3 will hold uniformly for all problems in this class.*

5 Conclusions

With parametric side-information, Algorithm M yields a hypothesis with error $P_{m,n} < c_3\epsilon$ *with confidence in excess of* $1 - \delta$ *given a random mixed sample of*

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

unlabelled examples and

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

labelled examples. Compare this with the purely-labelled sample scenario of Section 2 where a learner using Algorithm E requires

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

labelled examples to achieve the same learning performance. Clearly, the introduction of unlabelled examples has reduced the demands on

the number of labelled examples. As a rough measure one may take the ratio

$$\frac{n_M}{(m_E - m_M)} = \frac{c_7 d^2}{\epsilon^3 \delta \log d}$$

as indicative of the number of unlabelled examples needed to compensate for the removal of each labelled example. In this sense, the worth of supervision is polynomial in d, $\frac{1}{\epsilon}$, *and* $\frac{1}{\delta}$. *Given detailed parametric side-information, labelled examples are worth a polynomial number of unlabelled examples.*

Inverting the sample complexity estimates for the mixed sample case shows that the error rate contribution from the labelled examples decreases exponentially fast in m, *while the error rate contribution from the unlabelled examples decreases only as an inverse polynomial in* n. *Note also that the e cacy of the unlabelled examples in the algorithmic scenarios discussed here diminishes as the dimensionality* d *increases. More explicitly, the hypothesis error probability under Algorithm M is* $\mathcal{O}\left(d^{3/5} n^{-1/5}\right) + \mathcal{O}\left(e^{-cm}\right)$.

The analysis works for other parametric families as long as the tails of the distributions are not too heavy. In particular, the analysis works in totofor other members of the exponential family, albeit with di erent constants and rates.

In nonparametric cases, the tradeo between labelled and unlabelled examples is much more pronounced as seen in Section 4. For instance, if Kernel density estimation is used to infer the modes, hence the separating hyperplane, each labelled example is worth roughly

$$\frac{c_8 13^{d \log(5 + \log d)}}{d \log d \epsilon^{\frac{d}{2 \log d}}}$$

unlabelled examples whence the hypothesis error probability under Algorithm K is $\mathcal{O}\left(d^{c_9 \log d} n^{-2 \log d / d}\right) + \mathcal{O}\left(e^{-cm}\right)$. *Labelled examples are worth an exponential number of unlabelled examples when little is known about the problem.*

References

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

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

[3] *T. M. Cover, "Learning and generalisation," 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.*

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

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

[6] *J. Ratsaby, The Complexity of Learning from a Mixture of Labelled and Unlabelled Examples, Ph.D. Thesis, University of Pennsylvania, 1994.*