THE COMPLEXITY OF LEARNING FROM A MIXTURE OF LABELED AND UNLABELED EXAMPLES*†

Joel Ratsaby Santosh S. Venkatesh

April 1994

The learning of a pattern classification rule rests on acquiring information to constitute a decision rule that is close to the optimal Bayes rule. Among the various ways of conveying information, showing the learner examples from the different classes is an obvious approach and ubiquitous in the pattern recognition field. Basically there are two types of examples: labeled in which the learner is provided with the correct classification of the example and unlabeled in which this classification is missing. Driven by the reality that often unlabeled examples are plentiful whereas labeled examples are difficult or expensive to acquire we explore the tradeoff between labeled and unlabeled sample complexities (the number of examples required to learn to within a specified error), specifically getting a quantitative measure of the reduction in the labeled sample complexity as a result of introducing unlabeled examples. This problem was posed in this form by T. M. Cover and may be succinctly, if inexactly, stated as follows: How many unlabeled examples is one labeled example worth?

The direction taken in this dissertation focuses on the archetypal problem of learning a classification problem with two pattern classes that are typified by fea ture vectors, i.e., examples drawn from class conditional Gaussian distributions and where the learning approaches are parametric and nonparametric. Denoting the dimensionality of the example-space as N, and the number of labeled and unlabeled examples as m and n respectively, then for specific algorithms, it is shown that under a nonparametric scenario the classification error probability decreases roughly as $O\left(\left(c_0 n^{-2/N}\right)^{\log N}\right) + O\left(e^{-c_1 m}\right)$, and in the parametric scenario the error decreases roughly as $O\left(N^{3/5} n^{-1/5}\right) + O\left(e^{-c_1 m}\right)$. where c_0, $c_1 > 0$ are constants with respect to N, m and n. This shows that in both the parametric and nonparametric cases it takes roughly exponentially more unlabeled examples than labeled examples for the same reduction in error. When considering the effect of the dimensionality N, roughly speaking, a labeled ex-

*Abstract of `www.ariel.ac.il/sites/ratsaby/Publications/PDF/Ratsaby_Phd.pdf`, April 1994, University of Pennsylvania.

†*Keywords*: Semi-supervised learning, Labeled examples, Unlabeled examples

ample is worth exponentially more in the nonparametric than in the parametric scenario.

The parametric approach uses the Maximum Likelihood technique with labeled and unlabeled samples to construct a decision rule estimate. In this scenario the learner knows the parametric form of the pattern class densities. Sufficient finite sample complexities are established by which the value of one labeled example in terms of the number of unlabeled examples is determined to be polynomial in the dimensionality N. The analysis may provide the details for broadening the results to other non Gaussian parametric based families of problems. An extension to the case of different a priori class probabilities is investigated under this parametric scenario, and for the non-unit covariance Gaussian problem it is conjectured that the value of a labeled example is still polynomial in N.

In the nonparametric scenario the primary focus is on an algorithm which is based on Kernel Density Estimation. It uses a mixed sample to construct a decision rule where now the learner has significantly less side information about the class den sities. The finite sample complexities for learning the Gaussian based problem are established by which the value of one labeled example is determined to be exponential in the dimensionality N. An extension to a larger family of nonparametric classification problems is provided where the same tradeoff applies. A variant of this approach is investigated in which only a finite number of functional values of the underlying mixture density are estimated. This yields a smaller tradeoff but is still exponential in N. The mixed sample complexities for the classical k-means clustering procedure are also determined.

An experimental investigation using neural networks examines the value of a labeled example when learning a classification problem based on a Gaussian mixture. For other classification problems, the cost of learning measured by the labeled sample size as a function of the dimensionality N, is shown to be lower for a two-layer network than with the regular single layer Kohonen network. This is attributed to the better discrimination ability of the partition of the classifier.