Learning Decision Rules for Pattern Classification Under a Family of Probability Measures*

S.R. Kulkarni[1] M. Vidyasagar[2]

[1]Department of Electrical Engineering, Princeton University
Princeton, NJ 08544, email: kulkarni@ee.princeton.edu

[2]Centre for Artificial Intelligence and Robotics, Raj Bhavan Circle
High Grounds, Bangalore 560 001, India, email: sagar@cair.ernet.in

Abstract — **In this paper, the PAC learnability of decision rules for pattern classification under a family of probability measures is investigated. It is shown that uniform boundedness of the metric entropy of the class of decision rules is both necessary and sufficient for learnability if the family of probability measures is either *compact*, or *contains an interior point*, with respect to total variation metric. Then it is shown that learnability is preserved under finite unions of families of probability measures, and also that learnability with respect to each of a finite number of measures implies learnability with respect to the convex hull of the families of "commensurate" probability measures.**

I. Summary

Consider the following standard pattern classification problem. We wish to classify an observed feature vector $x \in X$ to either class 0 or class 1, where x is drawn according to some distribution $P(x)$ and given x the two classes have conditional probabilities $P(0|x)$, $P(1|x)$. We will assume that $P(x)$ is known to belong to a family of distributions $\mathcal{P}$, but is otherwise unknown, and that $P(0|x)$ and $P(1|x)$ are completely unknown. In addition to the observed feature vector x, we also have labeled observations $(x_1, y_1), \ldots (x_m, y_m)$ drawn i.i.d. from the (unknown) distributions.

Often (e.g., in neural networks) the classifier is restricted to belong to some class, $\mathcal{D}$, of decision rules. Each decision rule $D \in \mathcal{D}$ is a map $D : X \to \{0, 1\}$. Conditioned on $x \in X$, the probability of error of a decision rule $D \in \mathcal{D}$ is given by $R(D|x) = 1 - P(D(x)|x)$, and the average performance of D is $R(D) = ER(D|x)$ where the expectation is with respect to $P(x)$. The optimal performance over all $D \in \mathcal{D}$ is given by $R^* = \inf_{D \in \mathcal{D}} R(D)$. An **algorithm** is a family of maps $A_m : [X \times \{0, 1\}]^m \to \mathcal{D}$, for each integer $m \geq 1$, so that $H_m = A_m[(x_1, y_1), \ldots, (x_m, y_m)]$ is the decision rule generated after m observations. The class $\mathcal{D}$ is said to be **PAC-learnable** (e.g., see [4]) with respect to $\mathcal{P}$ if there exists an algorithm $\{A_m\}$ such that for every $\epsilon, \delta \in (0, 1)$ there exists an integer $m = m(\epsilon, \delta)$ such that $\Pr\{(x_1, y_1), \ldots, (x_m, y_m) : R(H_m) \leq R^* + \epsilon\} \geq 1 - \delta$ for all $P \in \mathcal{P}$ and all $P(y|x)$ for $y \in \{0, 1\}$.

Almost all work has considered either a singleton set $\mathcal{P}$ or $\mathcal{P} = \mathcal{P}^*$ the set of all distributions on a fixed σ-algebra on X. In the former case (pp. 149-151 of [5], [1]), $\mathcal{D}$ is learnable if and only if it has a finite ϵ-cover for each positive ϵ with respect to the pseudometric $d_P(A, B) = E_P|A(x) - B(x)|$. In the latter case [5, 2], $\mathcal{D}$ is learnable if and only if it has finite Vapnik-Chervonenkis (VC) dimension. Thus, the two extreme situations are well-understood, but little is known about learnability w.r.t. general $\mathcal{P}$. Clearly a *necessary* condition for learnability for arbitrary $\mathcal{P}$ is that the class of decision rules should have uniformly bounded covering numbers w.r.t. $\mathcal{P}$, or, equivalently, uniformly bounded metric entropy (which we refer to as UBME). However, it has been recently shown that this condition is *not* sufficient in general [3]. Thus the problem of deriving necessary and sufficient conditions for learnability under a general family of measures is still open. The present paper derives a few results in this direction.

Theorem 1 *Let $\mathcal{D}$ be a class of decision rules, and suppose the family $\mathcal{P}$ has a nonempty interior w.r.t. total variation metric (i.e., $\rho(P, Q) = \sup_{A \in \mathcal{S}} |P(A) - Q(A)|$). Then the following are equivalent: (i) $\mathcal{D}$ is learnable with respect to $\mathcal{P}$, (ii) $\mathcal{D}$ has UBME w.r.t $\mathcal{P}$, and (iii) $\mathcal{D}$ has finite VC dimension.*

Theorem 2 *Suppose $\mathcal{D}$ is a class of decision rules and $\mathcal{P}$ is a precompact family of probability measures. Then the following are equivalent: (i) $\mathcal{D}$ is learnable w.r.t $\mathcal{P}$, (ii) $\mathcal{D}$ has UBME with respect to $\mathcal{P}$, and (iii) $\mathcal{D}$ has a finite ϵ-cover with respect to the pseudometric $\bar{d}_{\mathcal{P}} = \sup_{P \in \mathcal{P}} d_P(A, B)$ for each $\epsilon > 0$.*

Hence, Theorems 1 and 2 are essentially like the distribution-free and fixed distribution cases, respectively.

Theorem 3 *Let $\mathcal{D}$ be a class of decision rules. Let $\mathcal{P}_1, \ldots, \mathcal{P}_n$ be n families of measures. If $\mathcal{D}$ is learnable w.r.t. $\mathcal{P}_i$ for $i = 1, \ldots, n$ then $\mathcal{D}$ is learnable w.r.t. $\cup_{i=1}^n \mathcal{P}_i$.*

Let P_0 be a fixed probability measure and let $b \geq 1$. Define $\mathcal{M}(b, P_0)$ to be the set of all probability measures P on $\mathcal{S}$ such that $P(S) \leq b\, P_0(S)$, $\forall\, S \in \mathcal{S}$.

Theorem 4 *Suppose $\mathcal{D}$ is a class of decision rules, $P_1, \ldots, P_n$ are probability measures, and $b_1, \ldots, b_n \geq 1$. Then $\mathcal{D}$ is learnable w.r.t. to the closed convex hull of $\cup_{i=1}^n \mathcal{M}(b_i, P_i)$ iff $\mathcal{D}$ is learnable w.r.t. each P_i.*

*The work of the first author was supported in part by by the National Science Foundation under grant IRI-9209577 and by the Army Research Office under grant DAAL03-92-G-0320.

References

[1] G.M. Benedek and A. Itai, "Learnability with respect to fixed distributions," *Theoretical Comp. Sci.*, 86, pp. 377-390, 1991.

[2] A. Blumer, A. Ehrenfeucht, D. Haussler and M. Warmuth, "Learnability and the Vapnik-Chervonenkis dimension," *J. ACM*, 36(4), pp. 929-965, 1990.

[3] R.M. Dudley, S.R. Kulkarni, T.J. Richardson and O. Zeitouni, "A metric entropy bound is not sufficient for learnability," to appear *IEEE Trans. Info. Theory*.

[4] D. Haussler, "Decision theoretic generalizations of the PAC model for neural net and other learning applications," *Information and Computation*, vol. 100, pp. 78-150, 1992.

[5] V.N. Vapnik, *Estimation of Dependences Based on Empirical Data*, Springer-Verlag, 1982.

0 - 7803-2015-8/94/$4.00