Elect. Comm. in Probab. **8** (2003) 1–10

ELECTRONIC COMMUNICATIONS in PROBABILITY

A NOTE ON THE RICHNESS OF CONVEX HULLS OF VC CLASSES

GÁBOR LUGOSI[1]
Department of Economics, Pompeu Fabra University, Ramon Trias Fargas 25–27, 08005 Barcelona, Spain
email: lugosi@upf.es

SHAHAR MENDELSON[2]
RSISE, The Australian National University, Canberra 0200, Australia
email: shahar.mendelson@anu.edu.au

VLADIMIR KOLTCHINSKII[3]
Department of Mathematics and Statistics, The University of New Mexico, Albuquerque NM, 87131-1141, USA
email: vlad@math.unm.edu

Submitted 2 May 2003 , *accepted in final form* 17 November 2003

AMS 2000 Subject classification: 62G08, 68Q32
Keywords: VC dimension, convex hull, boosting

Abstract

We prove the existence of a class $\mathcal{A}$ of subsets of $\mathbb{R}^d$ of VC dimension 1 such that the symmetric convex hull $\mathcal{F}$ of the class of characteristic functions of sets in $\mathcal{A}$ is rich in the following sense. For any absolutely continuous probability measure μ on $\mathbb{R}^d$, measurable set $B \subset \mathbb{R}^d$ and $\epsilon > 0$, there exists a function $f \in \mathcal{F}$ such that the measure of the symmetric difference of B and the set where f is positive is less than ϵ. The question was motivated by the investigation of the theoretical properties of certain algorithms in machine learning.

Let $\mathcal{A}$ be a class of sets in $\mathbb{R}^d$ and define the symmetric convex hull of $\mathcal{A}$ as the class of functions

$$\text{absconv}(\mathcal{A}) = \left\{ \sum_{i=1}^k a_i \mathbb{1}_{A_i}(x) : k > 0, a_i \in \mathbb{R}, \sum_{i=1}^k |a_i| = 1, A_i \in \mathcal{A} \right\}$$

where $\mathbb{1}_A(x)$ denotes the indicator function of A. For every $f \in \text{absconv}(\mathcal{A})$, define the set $C_f = \{x \in \mathbb{R}^d : f(x) > 0\}$ and let $\mathcal{C}(\mathcal{A}) = \{C_f : f \in \text{absconv}(\mathcal{A})\}$. We say that absconv$(A)$ is *rich* with respect to the probability measure μ on $\mathbb{R}^d$ if for every $\epsilon > 0$ and measurable set $B \subset \mathbb{R}^d$ there exists a $C \in \mathcal{C}(\mathcal{A})$ such that

$$\mu(B \triangle C) < \epsilon$$

[1]SUPPORTED THE SPANISH MINISTRY OF SCIENCE AND TECHNOLOGY AND FEDER, GRANT BMF2003-03324
[2]SUPPORTED BY AN AUSTRALIAN RESEARCH COUNCIL DISCOVERY GRANT
[3]SUPPORTED BY NSA GRANT MDA904-02-1-0075 AND NSF GRANT DMS-0304861

where $B \triangle C$ denotes the symmetric difference of B and C.
Another way of measuring the richness of a class of sets (rather than the density of the class of sets) is the *Vapnik-Chervonenkis* (VC) *dimension*.

Definition 1 *Let $\mathcal{A}$ be a class of subsets of Ω. We say that $\mathcal{A}$ shatters $\{x_1, ..., x_n\} \subset \Omega$, if for every $I \subset \{1, ..., n\}$ there is a set $A_I \in \mathcal{A}$ for which $x_i \in A_I$ if $i \in I$ and $x_i \notin A_I$ if $i \notin I$. The* VC *dimension of $\mathcal{A}$ is the largest cardinality of a subset of Ω, shattered by $\mathcal{A}$.*

The problem we investigate in this note is the following. What is the smallest integer V such that there exists a class $\mathcal{A}$ of VC dimension V whose symmetric convex hull is rich with respect to a "large" collection of probability measures on $\mathbb{R}^d$? It is easy to construct classes of finite VC dimension that are rich in this sense for all probability measures. For example, the class of all linear halfspaces, which has VC dimension $d+1$, is also rich in the sense described above ([4, 6]).
The result of this note is that the minimal VC dimension guaranteeing richness of the symmetric convex hull with respect to all absolutely continuous probability measures is independent of the dimension d of the space.

Theorem 1 *For any $d \geq 1$, there exists a class $\mathcal{A}$ of measurable subsets of $\mathbb{R}^d$ of* VC *dimension equal to one such that* $\mathrm{absconv}(\mathcal{A})$ *is rich with respect to all probability measures which are absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}^d$.*

The problem discussed here is motivated by recent results in Statistical Learning Theory, where several efficient classification algorithms (e.g. "boosting" [9, 5] and "bagging" [2, 3]) form convex combinations of indicator functions of a small "base" class of sets. In order to guarantee that the resulting classifier can approximate the optimal one regardless of the distribution, the richness property described above is a necessary requirement, but the size of the estimation error is determined primarily by the VC dimension of the base class (see [7], and references therein). Therefore, it is desirable to use a base class with a VC dimension as small as possible. For a direct motivation we refer the reader to [1], where a regularized boosting algorithm is shown to have a rate of convergence faster than $O(n^{-(V+2)/4(V+1)})$ for a large class of distributions, which only depends on the richness of the convex hull.
The proof of Theorem 1 presented below is surprisingly simple. It differs from the original proof we had which was based on the existence of a space-filling curve.
The first step in the proof is the well-known Borel isomorphism Theorem (see, e.g., [8], Theorem 16, page 409) which we recall here for completeness. For a metric space X, let $\mathcal{B}(X)$ be the Borel σ-field. Recall that a mapping $\phi : (X, \mathcal{B}(X)) \to (Y, \mathcal{B}(Y))$ is a Borel equivalence if ϕ is a one-to-one and onto mapping, such that ϕ and ϕ^{-1} map Borel sets to Borel sets.

Lemma 1 *Let $(X, \mathcal{B}(X), \mu)$ be a complete, separable metric measure space, where μ is a non-atomic probability measure, and let λ be the Lebesgue measure on $[0,1]$. Then there is a mapping $\phi : [0,1] \to X$ which is a measure-preserving Borel equivalence.*

The proof of Theorem 1 follows almost immediately from the Lemma. Indeed, let $\mathcal{A} = \{[0,t] : t \in [0,1]\}$. Note that $\mathrm{VC}(\mathcal{A}) = 1$, and it is well known (see, e.g., [1]) that $\mathrm{absconv}(\mathcal{A})$ is rich. Let μ be the standard gaussian measure on $\mathbb{R}^d$ and let $\phi : ([0,1], \mathcal{B}([0,1]), \lambda) \to (\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d), \mu)$ be the Borel isomorphism guaranteed by the Lemma. Set $\mathcal{D} = \{\phi(A) : A \in \mathcal{A}\}$, and observe that since ϕ is one-to-one, we have $\mathrm{VC}(\mathcal{D}) = 1$. Moreover, $f \in \mathrm{absconv}(\mathcal{D})$ if and only if $f \circ \phi \in \mathrm{absconv}(\mathcal{A})$, and for every such f,

$$C_f = \left\{x \in \mathbb{R}^d \ : \ f(x) > 0\right\} = \phi\left(\{t \in [0,1] \ : \ f(\phi(t)) > 0\}\right),$$

implying that $\mathcal{C}(\mathcal{D}) = \{\phi(U) \ : \ U \in \mathcal{C}(\mathcal{A})\}$. The richness of $\mathcal{D}$ with respect to μ follows from the fact that $\mathcal{A}$ is rich, and that the function ϕ is one-to-one and measure preserving. The richness with respect to the Lebesgue measure follows by absolute continuity.
Note that Theorem 1 is true for much more general structures than $\mathbb{R}^d$ and measures that are absolutely continuous, because the proof relies on the existence of the Borel isomorphism.

References

[1] G. Blanchard, G. Lugosi, and N. Vayatis. On the rate of convergence of regularized boosting classifiers. *Journal of Machine Learning Research*, 4:861-894, 2003.

[2] L. Breiman. Bagging predictors. *Machine Learning*, 24:123–140, 1996.

[3] L. Breiman. Bias, variance, and arcing classifiers. Technical report, Department of Statistics, University of California at Berkeley, Report 460, 1996.

[4] G. Cybenko. Approximations by superpositions of sigmoidal functions. *Math. Control, Signals, Systems*, 2:303–314, 1989.

[5] Y. Freund. Boosting a weak learning algorithm by majority. *Information and Computation*, 121:256–285, 1995.

[6] K. Hornik, M. Stinchcombe, and H. White. Multi-layer feedforward networks are universal approximators. *Neural Networks*, 2:359–366, 1989.

[7] G. Lugosi and N. Vayatis. On the bayes-risk consistency of regularized boosting methods. *Annals of Statistics*, 2003, to appear.

[8] H.L. Royden. *Real Analysis* Third edition, Macmillan, 1988.

[9] R.E. Schapire. The strength of weak learnability. *Machine Learning*, 5:197–227, 1990.