On Multiple-Instance Learning of Halfspaces[*]

D. I. Diochnos[1], R. H. Sloan[1], and Gy. Turán[1,2]

[1]University of Illinois at Chicago ddioch2|sloan|gyt @uic.edu
[2]Research Group on AI, Hungarian Acad. Sciences & U. Szeged, Hungary

June 1, 2012

Abstract

In multiple-instance learning the learner receives bags, i.e., sets of instances. A bag is labeled positive if it contains a positive example of the target. An $\Omega(d \log r)$ lower bound is given for the VC-dimension of bags of size r for d-dimensional halfspaces and it is shown that the same lower bound holds for halfspaces over any large point set in general position. This lower bound improves an $\Omega(\log r)$ lower bound of Sabato and Tishby, and it is sharp in order of magnitude. We also show that the hypothesis finding problem is NP-complete and formulate several open problems.

1 Introduction

Multiple-instance or multi-instance learning (MIL) is a variant of the standard PAC model of concept learning where, instead of receiving labeled instances as examples, the learner receives labeled bags, i.e., labeled sets of instances. A bag is labeled positive if it contains at least one positive example, and it is labeled negative otherwise. There are different probability models for the distribution of bags; one possible model, which we will call the *independent* model, assumes that instances in a bag are independent and identically distributed. The multi-instance setting, introduced by Dietterich *et al.* [6], is natural for several learning applications, for example, in drug design and image classification. In drug design, a bag may consist of several shapes of a molecule and it is labeled positive if some shape binds to a specific binding site. In image classification, a bag may be a photo containing several objects and it is labeled positive if it contains some object of interest.

[*]This work was supported by the National Science Foundation under Grant No. CCF-0916708.

Blum and Kalai [2] showed that every learning problem that is efficiently learnable with statistical queries is also efficiently learnable in the independent MIL model, and, more generally, the same holds for problems efficiently learnable with one-sided random classification noise. Every problem known to be efficiently PAC-learnable is also known to be efficiently learnable with one-sided random classification noise, although no formal relationship is proven so far (see Simon [15] for further discussion of the one-sided random classification noise model). Thus [2] implies the efficient independent MIL-PAC-learnability of all known efficiently PAC-learnable classes.

A detailed study of sample sizes in the MIL model was initiated Sabato and Tishby [12]. They proved a general upper bound for the VC-dimension of bags, and a lower bound for the concept class of halfspaces. Kundakcioglu *et al.* [10] considered margin maximization for bags of halfspaces and gave NP-completeness and experimental results.

In this note we continue the study of multi-instance learning of halfspaces. We improve the VC-dimension lower bound of [12] from $\Omega(\log r)$ to $\Omega(d \log r)$, where d is the dimension and r is the bag size, which is optimal up to order of magnitude. A similar result was given independently by Sabato and Tishby [13]. We also show that the same lower bound holds for bags over every sufficiently large point set in general position. Thus the situation is somewhat analogous to standard halfspaces, where every simplex forms a maximum shattered set. The proofs are based on cyclic polytopes. We also show that hypothesis finding for bags of halfspaces is NP-complete, using a variant of the construction of [10]. These two results, in view of the well-known relationship between PAC-learnability, VC-dimension and hypothesis finding, indicate differences between the PAC and the independent MIL-PAC models.

There are several open problems related to the multi-instance learning of halfspaces. Some of these are discussed in the concluding section of the paper.

2 Preliminaries

A halfspace in $\mathbf{R}^d$ is given as $H = \{x \in \mathbf{R}^d \,:\, w \cdot x \geq t\}$, for weight vector $w \in \mathbf{R}^d$ and threshold $t \in \mathbf{R}$. A bag of size r, or an r-bag, is an r-element multiset $B = \{x_1, \ldots, x_r\}$ in $\mathbf{R}^d$. An r-bag B is positive for H if $B \cap H \neq \emptyset$, and B is negative for H otherwise. A set of bags $\mathcal{B} = \{B_1, \ldots B_s\}$ is shattered by halfspaces if for every $\pm$ labeling of the bags there is a halfspace that assigns the same labels to the bags in $\mathcal{B}$. The VC-dimension of r-bags for d-dimensional halfspaces is the largest s such that there are s shattered bags. For $r = 1$ one gets the usual notion of VC-dimension of halfspaces and it is a basic fact that this equals $d + 1$.

3 The VC-dimension of r-bags for d-dimensional halfspaces

We first formulate a general upper bound of Sabato and Tishby [12], and then we give the matching lower bound for halfspaces. The lower bound is based on properties of cyclic polytopes. The discussion is essentially self-contained as we include a brief overview of the background material (details not given here can be found in Matoušek [11]).

3.1 A general upper bound

Sabato and Tishby [12] showed that the VC-dimension of r-bags for any concept class is essentially at most a $\log r$ factor larger than the VC-dimension of the concept class. We formulate their result in a slightly different form.

Theorem 1 ([12])**.** *For any concept class of VC-dimension $\tilde{d}$, the VC-dimension of r-bags is $O(\tilde{d}\log r)$.*

Proof. Let $\mathcal{B} = \{B_1, \ldots B_s\}$ be a shattered set of r-bags. Then $\mathcal{B}$ contains at most rs instances, and by Sauer's lemma, those can be classified by concepts in the class in at most $((ers)/\tilde{d})^{\tilde{d}}$ many ways. The classification of the instances in the bag determines the classification of the bags. Thus

$$2^s \leq \left(\frac{ers}{\tilde{d}}\right)^{\tilde{d}}.$$

Writing $x = s/\tilde{d}$ this becomes $2^x/x \leq er$. The function $2^x/x$ is monotone if $x \geq 1/\ln 2$. Thus it is sufficient to show that $2^x/x > er$ for $x = \log r + 2\log\log r$, if r is sufficiently large, which follows directly. □

3.2 Lower bound for halfspaces

Sabato and Tishby showed that the VC-dimension of r-bags of halfpaces in the plane is at least $\lfloor \log r \rfloor + 1$, which implies the same bound for higher dimensions. We now prove a lower bound by adding the 'missing' factor d, which is optimal in order of magnitude by Theorem 1.

The d-dimensional moment curve is given parametrically as $x(t) = (t, t^2, \ldots, t^d)$. The convex hull of points $x(t_1), \ldots, x(t_n)$ on the moment curve, for $t_1 < \cdots < t_n$, with $n \geq d+1$, is called a *cyclic polytope*. For any $I \subseteq [n]$, $|I| \leq \lfloor d/2 \rfloor$, the polynomial

$$\prod_{i \in I}(t - t_i)^2 = \sum_{j=0}^{d} w_j t^j$$

is 0 at every $t_i, i \in I$ and positive at every $t_i, i \notin I$. Thus the halfspace $-\sum_{j=1}^{d} w_j x_j \geq w_0$ contains every point $x(t_i), i \in I$, and none of the points $x(t_i), i \notin I$. Hence every set of at most $\lfloor d/2 \rfloor$ vertices forms a face of a cyclic polytope.

Theorem 2. *The VC-dimension of d-dimensional halfspaces over bags of size r is at least $\lfloor d/2 \rfloor (\lfloor \log r \rfloor + 1)$.*

Proof. Let ℓ be an integer,

$$s = \left\lfloor \frac{d}{2} \right\rfloor (\ell + 1), \quad r = 2^{\ell}, \quad n = \left\lfloor \frac{d}{2} \right\rfloor \cdot 2^{\ell+1}.$$

Let $t_1 < \cdots < t_n$ be arbitrary and consider the set of n instances $X = \{x(t_1), \ldots, x(t_n)\}$. Divide X into $\lfloor d/2 \rfloor$ blocks of size $2^{\ell+1}$ each, i.e., let

$$X_i = \{x(t_j) \, : \, (i-1) \cdot 2^{\ell+1} < j \leq i \cdot 2^{\ell+1}\}, \; i = 1, \ldots, \lfloor d/2 \rfloor .$$

Let f_i be a bijection between X_i and the subsets of integers in the interval $[(i-1) \cdot (\ell+1) + 1, i \cdot (\ell + 1)]$ and let

$$B_k = \{x(t_j) \, : \, k \in f_i(x(t_j))\}$$

for every k such that $(i-1) \cdot (\ell+1) < k \leq i \cdot (\ell+1)$. We claim that $\{B_1, \ldots, B_s\}$ is a family of bags of size r shattered by d-dimensional halfspaces. Each bag is of size r as it contains a half of a block. For any subset $S \subseteq [s]$ let $S_i = S \cap [(i-1) \cdot (\ell+1) + 1, i \cdot (\ell+1)]$ and let $x(t_{j(i)})$ be the point such that $f_i(x(t_{j(i)})) = S_i$, for $i = 1, \ldots, \lfloor d/2 \rfloor$. Then the set $\{x(t_{j(i)}) \, : \, i = 1, \ldots, \lfloor d/2 \rfloor\}$ can be separated from the rest of X by a halfspace, and that halfspace classifies precisely those bags B_k as positive for which $k \in S$. Thus the family of bags is indeed shattered by halfspaces. The VC-dimension bound follows directly from the definition of s and r. □

Now we prove a strengthening of Theorem 2. A finite subset of $\mathbf{R}^d$ is in *general position* if all its $(d+1)$-subsets are affinely independent, i.e., have no linear combination equal to 0, with coefficients adding up to 0. Halfspaces in $\mathbf{R}^d$ shatter *every* simplex, i.e., every set of $(d+1)$ points in general position. In analogy to this fact, we prove a VC-dimension lower bound similar to Theorem 2 for bags of halfspaces when the instances are restricted to *any* sufficiently large subset in general position.

The proof uses some further properties of cyclic polytopes. Given a convex polytope P, its *face lattice* is the family of its faces partially ordered by set inclusion. Two convex polytopes are *combinatorially equivalent* if their face lattices are isomorphic. Combinatorial equivalence follows from the existence of a bijection between the vertex sets of the two polytopes which form a bijection between their facets (i.e., $(d-1)$-dimensional faces). The facets of cyclic polytopes are described by *Gale's evenness condition*: for $t_{i_1} < \cdots < t_{i_d}$ the vertices $x(t_{i_1}), \ldots, x(t_{i_d})$ form a facet if and only if for any two other vertices $x(t_u)$ and $x(t_v)$ there are an even number of values t_{i_j} between t_u and t_v. This is proven by considering the hyperplane $\sum_{j=1}^{d} w_j x_j = -w_0$, where the coefficients are defined by

$$\prod_{j=1}^{d}(t - t_{i_j}) = \sum_{j=0}^{d} w_j t^j.$$

The condition follows by counting the number of sign changes between t_u and t_v.

For $a \in \mathbf{R}^d$ let a' be the vector obtained from a by adding 1 as a first component. Then for any $t_{i_0} < \cdots < t_{i_d}$, the matrix with columns $x(t_0)', \ldots, x(t_d)'$ is a Vandermonde matrix and thus its determinant is positive.

According to Ramsey's theorem (see [8]), there is a function $R(u, v)$ such that if the u-subsets of a set of size at least $R(u, v)$ are two-colored then there is a subset of size v with all its u-subsets colored the same.

The following lemma is referred to as "unpublished 'folklore' " and proven in an oriented matroid version by Cordovil and Duchet [4] [1]. It is also given as an exercise in Matoušek [11], and it is proven here for completeness.

Lemma 3 (See [4, 11])**.** *Every set $A \subseteq \mathbf{R}^d$ of $R(d + 1, n)$ points in general position contains n points such that their convex hull is combinatorially equivalent to d-dimensional cyclic polytopes on n vertices.*

Proof. Consider a set A of $R(d + 1, n)$ points in general position and fix an arbitrary ordering $<$ of the elements of A. Color each $(d + 1)$-subset of A with the sign of the determinant of the matrix formed by the column vectors of the points in the subset, written in increasing order according to the fixed ordering, with an additional first row of ones added. Then there is a subset $A' = \{a_1, \ldots, a_n\}$ of A with $a_1 < \cdots < a_n$ such that determinants associated with each $(d + 1)$-subset all have the same sign.

Consider an arbitrary ordered d-subset $S = \{a_{i_1} < \cdots < a_{i_d}\}$ of A'. Denote by H the hyperplane determined by S. For any point $a_j \in A' \setminus S$, the sign of $\det(a'_j, a_{i_1}, \ldots, a_{i_d})$ determines which side of H contains a_j. The sequence $j, i_1, \ldots, i_d$ is not necessarily increasing; j can be brought into its proper place in the sequence by a sequence of transpositions. Each transposition corresponds to a column exchange which changes the sign of the determinant.

The set S forms a facet if and only if the sign of $\det(a_j, a_{i_1}, \ldots, a_{i_d})$ is the same for every vertex $a_j \in A' \setminus S$. This happens iff the parity of transpositions needed to bring j into its proper place is the same for every j such that $a_j \in A' \setminus S$. Hence S is a facet iff for every j, k such that $a_j, a_k \in A' \setminus S$ there is an even number of points in S between j and k. This also implies directly that every point belongs to a facet and thus the points of A' form a convex polytope. Thus the points in A' form a convex polytope whose facets, using the ordering $<$ on A', are described by Gale's evenness condition. Therefore the polytope is combinatorially equivalent to a cyclic polytope. □

[1] The paper is an updated version of an unpublished, but circulated, manuscript from 1986/87.

If A is any subset of $\mathbf{R}^d$ then a *halfspace over A* is a set $H \cap A$ for some d-dimensional halfspace H. A set of r-bags over A (i.e., a set of r-element multisets of A) is shattered by halfspaces over A if for every ± labeling of the bags there is a halfspace over A that assigns those labels to the bags. The VC-dimension of halfspaces over bags of size r from A is the largest s such that there are s many r-bags over A that are shattered by halfspaces over A.

Now we formulate the strengthening of Theorem 2.

Theorem 4. *There is a function $g(d,r)$ such that for every set A of $m \geq g(d,r)$ points in general position in $\mathbf{R}^d$, halfspaces over bags of size r from A have VC-dimension at least $\lfloor d/2 \rfloor (\log r + 1)$.*

Proof. The result follows by combining the construction of Theorem 2 with Lemma 3, setting $g(d,r) = R(d+1, dr)$. The set A contains a subset A' of size dr which determines a convex polytope combinatorially equivalent to a cyclic polytope with dr vertices. Every $\lfloor d/2 \rfloor$-subset of this polytope forms a face, thus we can repeat the construction of Theorem 2 to get $\lfloor d/2 \rfloor (\log r + 1)$ bags of size r over A', and thus over A as well, that are shattered by halfspaces. □

4 NP-completeness of hypothesis finding

The hypothesis-finding problem for r-bags for d-dimensional halfspaces is the following: given a set of labeled r-bags in $\mathbf{R}^d$, is there a halfspace that assigns these labels to the bags? The reduction below is a variant of a reduction in Kundakciouglu *et al.* [10].

Theorem 5. *The hypothesis finding problem for r-bags of d-dimensional halfspaces is NP-complete for every fixed $r \geq 3$.*

Proof. We give a reduction from 3-SAT (containment in NP is trivial). Let $C_1, \ldots, C_m$ be an instance of 3-SAT over variables $x_1, \ldots, x_d$. Let e_i be the i'th unit vector in $\mathbf{R}^d$. For $j = 1, \ldots, m$ let B_j be a positive bag containing e_i if x_i is in C_j, and $-e_i$ if $\neg x_i$ is in C_j. For $i = 1, \ldots, d$ let B'_i be a positive bag containing e_i and $-e_i$. Finally, let B^* be a negative bag containing the origin. We claim that the original formula is satisfiable iff there is a consistent hypothesis for the set of bags described.

Let $(a_1, \ldots, a_d)$ be a satisfying truth assignment. Then the halfspace $w_1 u_1 + \ldots + w_d u_d \geq 1$ is consistent, where $w_i = 1$ if $a_i = 1$ and $w_i = -1$ otherwise, for $i = 1, \ldots, d$.

In the other direction, let $w_1 u_1 + \cdots + w_d u_d \geq t$ be a consistent hypothesis. Then $t > 0$ as B^* is negative. Also, $w_i \neq 0$, as B'_i is positive. It follows directly that the truth assignment defined by $a_i = sign(w_i)$ satisfies the formula. □

5 Further remarks and open problems

We showed that the VC-dimension of r-bags of d-dimensional halfspaces is $\Theta(d \log r)$ over every sufficiently large point set in general position, and that hypothesis finding for r-bags of d-dimensional halfspaces is NP-complete. The latter implies that, unlike in the case of learning halfspaces, one does not get an efficient independent MIL-PAC learning algorithm by drawing $O(d \log r)$ random bags and finding a consistent hypothesis. On the other hand, the result of Blum and Kalai [2] *does* provide an efficient algorithm with sample size polynomial in r and d, but larger than the VC-dimension.

This raises two open questions concerning learning d-dimensional halfspaces in the independent MIL-PAC model: What is the minimal sample size of r-bags sufficient for efficient learning? What is the minimal sample size of r-bags without taking computational complexity into account? For the second question note that distributions over bags generated in the independent model are only a subclass of all possible distributions over bags;[2] thus the VC-dimension only provides an upper bound. Multi-instance learning under more general settings is discussed by Auer *et al.* and by Sabato and Tishby [1, 12].

Active learning is another variant of PAC learning. In this model the learner can decide whether to request the label of a random instance, and the complexity of an algorithm is measured by the number of label requests (see, e.g., Dasgupta [5]). It follows from results of Hanneke [9] and Friedman [7] that for learning hyperplanes over smooth distributions, the error of the hypotheses returned by the *mellow active learning algorithm* of Cohn *et al.* [3] decreases exponentially in the number of labels queried, with high probability.

Settles *et al.* proposed multi-instance active learning (MIAL) [14]. MIAL has been studied in several machine learning papers but, as far as we know, has not been considered so far in learning theory. There are several possibilities for formulating a model of active learning in the multi-instance model. Let us assume here that the learner gets unlabeled r-bags and then is charged for querying the label of a bag. Multi-instance learning of r-bags of d-dimensional halfspaces corresponds to learning concepts in (dr)-dimensional space of the form

$$\{(x_1, \ldots, x_r) : w \cdot x_i \geq t \text{ for some } i,\ 1 \leq i \leq r\}.$$

The results mentioned above imply positive results in this setting as well. The mellow algorithm for active learning has an efficient implementation whenever hypothesis finding can be done efficiently. This, again, does not work for bags of halfspaces. Thus it seems to be an open problem whether there is an efficient active learning algorithm with exponentially decreasing error rate.

Acknowledgement: We would like to thank Robert Langlois, Hans Ulrich Simon, and Balázs Szörényi for several interesting discussions.

[2] This explains why, unlike the standard setting, the efficient PAC learning algorithm of Blum and Kalai [2] does not lead to an efficient hypothesis finding algorithm for bags.

References

[1] P. Auer, P. M. Long, and A. Srinivasan. Approximating hyper-rectangles: Learning and pseudorandom sets. *J. Comput. Syst. Sci.*, 57:376–388, 1998.

[2] A. Blum and A. Kalai. A note on learning from multiple-instance examples. *Machine Learning*, 30(1):23–29, 1998.

[3] D. A. Cohn, L. A. Atlas, and R. A. Ladner. Improving generalization with active learning. *Machine Learning*, 15:201–221, 1994.

[4] R. Cordovil and P. Duchet. Cyclic polytopes and oriented matroids. *Eur. J. Comb.*, 21:49–64, 2000.

[5] S. Dasgupta. Two faces of active learning. *Theor. Comput. Sci.*, 412(19):1767–1781, 2011.

[6] T. G. Dietterich, R. H. Lathrop, and T. Lozano-Pérez. Solving the multiple instance problem with axis-parallel rectangles. *Artif. Intell.*, 89:31–71, 1997.

[7] E. Friedman. Active learning for smooth problems. In *COLT*, 2009.

[8] R. Graham, B. Rothschild, and J. H. Spencer. *Ramsey Theory*. Wiley, 2nd edition, 1990.

[9] S. Hanneke. *Theoretical Foundations of Active Learning*. PhD thesis, Carnegie Mellon University, 2009.

[10] O. E. Kundakcioglu, O. Seref, and P. M. Pardalos. Multiple instance learning via margin maximixation. *Applied Numerical Mathematics*, 60:358–369, 2010.

[11] J. Matoušek. *Lectures on Discrete Geometry*, volume 212 of *Graduate Texts in Mathematics*. Springer, 2002.

[12] S. Sabato and N. Tishby. Homogeneous multi-instance learning with arbitrary dependence. In *COLT*, 2009.

[13] S. Sabato and N. Tishby. Multi-instance learning with any hypothesis class. arXiv:1107.2021v1 [cs.LG], July 2011.

[14] B. Settles, M. Craven, and S. Ray. Multiple-instance active learning. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1289–1296, 2008.

[15] H. U. Simon. PAC-learning in the presence of one-sided classification noise. In *Int. Symp. Artificial Intelligence and Mathematics (ISAIM)*, 2012. (Electronic proceedings only).