Available at
www.ComputerScienceWeb.com
POWERED BY SCIENCE DIRECT®

Information Processing Letters 87 (2003) 229–234

Information Processing Letters

www.elsevier.com/locate/ipl

An upper bound on the sample complexity of PAC-learning halfspaces with respect to the uniform distribution

Philip M. Long

Genome Institute of Singapore, 1 Science Park Road, The Capricorn, #05-01, Singapore 117528, Republic of Singapore

Received 20 September 2002; received in revised form 8 April 2003

Communicated by P.M.B. Vitányi

Abstract

We show that halfspaces in n dimensions can be PAC-learned with respect to the uniform distribution with accuracy ε and confidence δ using $O(\frac{1}{\varepsilon}(n+\log\frac{1}{\delta}))$ examples.

Keywords: Machine learning; Sample complexity; PAC-learning; Halfspaces; Computational complexity

1. Introduction

In the PAC model [14], a learning algorithm is given examples $(x_1, f(x_1)), \ldots, (x_m, f(x_m))$ of the behavior of an unknown $\{0,1\}$-valued function f applied to independently randomly drawn elements of its domain. The learner then outputs a hypothesis h, and its goal is for h to accurately approximate f. The "target function" f is an arbitrary member of a class F of functions that the learning algorithm knows ahead of time. It is assumed that the domain elements $x_1, \ldots, x_m$ in the learner's input are chosen independently at random according to a probability distribution D. The same distribution D is used to measure the accuracy of h: this accuracy is the probability that h would incorrectly classify another element of the domain of f chosen according to D. If, given m examples, with probability at least $1-\delta$ (with respect to the draws of the random examples) the learner outputs a hypothesis whose accuracy is at least as good as ε it is said to (ε,δ) learn from m examples.

In this paper, we examine the number of examples required for (ε,δ) PAC-learning in the case in which D is the uniform distribution over the unit ball in $\mathbb{R}^n$ and F is the set of indicator functions for halfspaces whose separating hyperplanes go through the origin. For this problem, we show that

$$O\left(\frac{n+\log\frac{1}{\delta}}{\varepsilon}\right)$$

examples suffice. This improves on the best previously known bounds of

$$O\left(\frac{n}{\varepsilon}\log\frac{1}{\varepsilon}+\frac{1}{\varepsilon}\log\frac{1}{\delta}\right)$$

[6,15,3] and

$$O\left(\frac{n}{\varepsilon}\log\frac{1}{\delta}\right)$$

E-mail address: gislongp@nus.edu.sg (P.M. Long).

0020-0190/$ – see front matter

[7] that follow from more general results, and matches a known

$$\Omega\left(\frac{n+\log\frac{1}{\delta}}{\varepsilon}\right)$$

lower bound [9] for this particular problem to within a constant factor. As did the general analysis in [2] (see also [10]), our proof proceeds by showing that the set of halfspaces has a small *cover*—a collection of possible hypotheses with the property that every halfspace has a good approximation in the cover. The improved bound is obtained by showing that there is a small cover which also has the property that no target has very many elements in the cover that approximate it moderately well. This is helpful because hypotheses with errors a little worse than ε are especially dangerous, because it is particularly difficult to recognize that they are not accurate enough. The remainder of our paper is organized as follows. Section 2 takes care of some preliminaries. The proof of the bound is in Section 3, and some other related work is described in Section 4.

2. Preliminaries

2.1. Learning

An example is an element of $X \times \{0,1\}$ and a sample is a finite sequence of examples. A hypothesis is a function from X to $\{0,1\}$. For a hypothesis h, a function f from X to $\{0,1\}$, and a probability distribution D over X, define the error of h with respect to f and D to be

$$\mathrm{er}_{f,D}(h) = \Pr_{x\sim D}\big(h(x) \neq f(x)\big).$$

For a sample $S = ((x_1, y_1), \ldots, (x_m, y_m))$, let $\mathrm{er}_S(h)$ be the fraction of incorrect classifications that h makes on S, i.e.,

$$\mathrm{er}_S(h) = \frac{1}{m}\big|\big\{i\colon h(x_i) \neq y_i\big\}\big|.$$

A learning strategy takes as input a sample, and outputs a hypothesis. For a probability distribution D over X, a set F of functions from X to $\{0,1\}$ is said to be (ε,δ)-learnable with respect to D from m examples if there is a learning strategy such that, for any $f \in F$, if m examples $x_1, \ldots, x_m$ are drawn independently at random according to D, and $(x_1, f(x_1)), \ldots, (x_m, f(x_m))$ is passed to A, then, with probability at least $1-\delta$, the hypothesis h output by A satisfies $\mathrm{er}_{f,D}(h) \leqslant \varepsilon$.

2.2. Halfspaces, the uniform distribution, and distance

For each positive integer n, let U_n be the uniform distribution on the surface of the unit ball in $\mathbb{R}^n$. For each $\vec{w} \in \mathbb{R}^n$, let $h_{\vec{w}}$ be the indicator function for the halfspace with normal vector $\vec{w}$ whose separating hyperplane goes through the origin. Thus $h_{\vec{w}}(\vec{x}) = 1 \Leftrightarrow \vec{w}\cdot\vec{x} \geqslant 0$. Let H_n be $\{h_{\vec{w}}\colon \vec{w} \in \mathbb{R}^n\}$. For $f, g \in H_n$, let

$$\rho_{U_n}(f,g) = \Pr_{x\sim U_n}\big(f(x) \neq g(x)\big).$$

2.3. Balls in $\mathbb{R}^n$ and their volumes

Let $V_n(r)$ denote the volume of a ball of radius r in $\mathbb{R}^n$, with respect to the usual Euclidean distance. The following are well known.

Lemma 1. *For all $n \geqslant 2$, $V_n(r) = r^n V_n(1)$ and $2 \leqslant \frac{\sqrt{n}V_n(1)}{V_{n-1}(1)} \leqslant 3$.*

Proof. In Appendix A. □

3. PAC-learning halfspaces with respect to the uniform distribution

The following is our main result.

Theorem 2. *H_n can be (ε,δ)-PAC-learned with respect to U_n from $\mathrm{O}((n+\log\frac{1}{\delta})/\varepsilon)$ examples.*

We will make use of the standard Chernoff bound.

Lemma 3. *If $Y_1, \ldots, Y_m$ are i.i.d. $\{0,1\}$-valued random variables and $\Pr(Y_i = 1) = p$, then for all $0 < \gamma \leqslant 1$,*

$$\Pr\left(\sum_{i=1}^m Y_i > (1+\gamma)\mathrm{E}\left(\sum_{i=1}^m Y_i\right)\right) \leqslant \mathrm{e}^{-\gamma^2 pm/3},$$

$$\Pr\left(\sum_{i=1}^m Y_i < (1-\gamma)\mathrm{E}\left(\sum_{i=1}^m Y_i\right)\right) \leqslant \mathrm{e}^{-\gamma^2 pm/2}.$$

For each positive integer n, let U_{H_n} be the distribution over H_n obtained by sampling $\vec{w}$ uniformly from the unit ball, and taking $h_{\vec{w}}$. Our first lemma addresses the following question: given some halfspace h, if another halfspace g is chosen uniformly at random (i.e., according to U_{H_n}), how likely is it that it will be "close" to h. The upper bound is from [9], but its proof is included for completeness.

Lemma 4. *There are constants $c_1, c_2 > 0$ such that for any $n \in \mathbb{N}$, $n \geqslant 2$, for any $h \in H_n$, and for any $0 < \alpha \leqslant 1/2$,*

$$\Pr_{g\sim U_{H_n}}\left(\rho_{U_n}(g,h) \leqslant \alpha\right) \geqslant (c_1\alpha)^{n-1},$$

and for any $0 < \alpha \leqslant 1$,

$$\Pr_{g\sim U_{H_n}}\left(\rho_{U_n}(g,h) \leqslant \alpha\right) \leqslant (c_2\alpha)^{n-1}.$$

Proof. See Appendix B. □

The following lemma adds a small wrinkle to the traditional volume argument. The usual volume argument could be used to show that there is a small set G_n of halfspaces that contains an accurate approximation to each element of H_n. The modification shows that a similarly small set G_n also has the property that for any element h of H_n, not *too* many elements of G_n are close to h.

Lemma 5. *There is a constant $c_3 > 0$ such that, for all $n \geqslant 2$ and all $\varepsilon \in (0,1]$, there is a $G_n \subseteq H_n$ for which for all $h \in H_n$,*

- *there is an $g \in G_n$ such that $\rho_{U_n}(g,h) \leqslant \varepsilon/4$, and*
- *for all $\alpha \geqslant \varepsilon$, $|\{g \in G_n\colon\ \rho_{U_n}(g,h) \leqslant \alpha\}| \leqslant (c_3\alpha/\varepsilon)^{n-1}$.*

Proof. Fix n. Consider G_n constructed by repeatedly choosing an arbitrary element of H_n at a distance (with respect to ρ_{U_n}) greater than $\varepsilon/4$ from all of the previously chosen elements of G_n, for as long as this is possible. Choose $h \in H_n$ and $\alpha \geqslant \varepsilon$. At any point in time during this process, by the triangle inequality, the balls of radius $\varepsilon/8$ centered at the elements of G_n are pairwise disjoint. In other words, for any distinct $g_1, g_2 \in G_n$, we have

$$\left\{f\colon\ \rho_{U_n}(g_1,f) \leqslant \varepsilon/8\right\} \cap \left\{f\colon\ \rho_{U_n}(g_2,f) \leqslant \varepsilon/8\right\} = \emptyset.$$

Thus, if we denote OR by $\vee$,

$$\Pr_{f\sim U_{H_n}}\left(\bigvee_{g\in G_n\colon\ \rho_{U_n}(g,h)\leqslant\alpha}\left(\rho_{U_n}(g,f) \leqslant \varepsilon/8\right)\right) = \sum_{g\in G_n\colon\ \rho_{U_n}(g,h)\leqslant\alpha} \Pr_{f\sim U_{H_n}}\left(\rho_{U_n}(g,f) \leqslant \varepsilon/8\right).$$

If c_1 is as in the statement of Lemma 4, then that lemma implies that

$$\Pr_{f\sim U_{H_n}}\left(\bigvee_{g\in G_n\colon\ \rho_{U_n}(g,h)\leqslant\alpha}\left(\rho_{U_n}(g,f) \leqslant \varepsilon/8\right)\right) \geqslant \left|\left\{g \in G_n\colon\ \rho_{U_n}(g,h) \leqslant \alpha\right\}\right|\left(\frac{c_1\varepsilon}{8}\right)^{n-1}. \tag{1}$$

By the triangle inequality (see Fig. 1), Lemma 4 implies

$$\begin{aligned}\Pr_{f\sim U_{H_n}}&\left(\bigvee_{g\in G_n\colon\ \rho_{U_n}(g,h)\leqslant\alpha}\left(\rho_{U_n}(g,f) \leqslant \varepsilon/8\right)\right)\\ &\leqslant \Pr_{f\sim U_{H_n}}\left(\rho_{U_n}(h,f) \leqslant \alpha + \varepsilon/8\right)\\ &\leqslant \left(c_2(\alpha+\varepsilon/8)\right)^{n-1}\\ &\leqslant \left((9c_2/8)\alpha\right)^{n-1},\end{aligned} \tag{2}$$

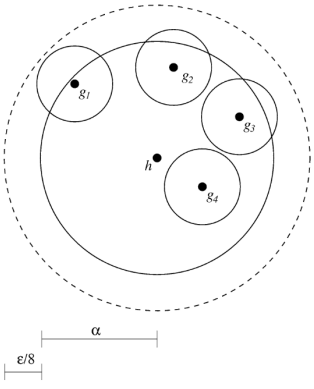

Fig. 1. By the triangle inequality, the union of the balls of radius $\varepsilon/8$ centered at elements of G_n at a distance at most α from h is contained in the ball of radius $\alpha + \varepsilon/8$ around h.

since $\alpha \geqslant \varepsilon$. Putting together (1) and (2) and solving, we get

$$\left|\left\{g \in G_n\colon \rho_{U_n}(g,h) \leqslant \alpha\right\}\right| \leqslant \left(\frac{9c_2\alpha}{c_1\varepsilon}\right)^{n-1}. \tag{3}$$

Note that (3), in the case $\alpha = 1$, implies that the process used to generate G_n terminates. When it does, all elements are within distance $\varepsilon/4$ of some element of G_n, since otherwise, another round would be possible. This completes the proof. □

Proof of Theorem 2. Consider the algorithm that, given ε,

- constructs G_n as in Lemma 5, and
- outputs the element of G_n with the minimum number of disagreements with the sample.

Choose $f \in H_n$, and let f^* be an element of G_n minimizing $\mathrm{er}_{f,U_n}(f^*)$.

Let h be the hypothesis output by the above algorithm, which is a function of the random sample. Then

$$\begin{aligned}&\Pr\left(\mathrm{er}_{f,U_n}(h) > \varepsilon\right)\\ &\quad\leqslant \Pr\left(\mathrm{er}_S(f^*) > \varepsilon/2 \text{ or } \exists g \in G_n,\right.\\ &\qquad\left.\mathrm{er}_{f,U_n}(g) > \varepsilon \text{ and } \mathrm{er}_S(g) \leqslant \varepsilon/2\right).\end{aligned} \tag{4}$$

Lemma 5 implies that $\mathrm{er}_{f,U_n}(f^*) \leqslant \varepsilon/4$. Applying the Chernoff bound (Lemma 3), we have

$$\Pr\left(\mathrm{er}_S(f^*) > \varepsilon/2\right) \leqslant \mathrm{e}^{-\varepsilon m/12}. \tag{5}$$

Let

$$p = \Pr\left(\exists g \in G_n,\ \mathrm{er}_{f,U_n}(g) > \varepsilon \text{ and } \mathrm{er}_S(g) \leqslant \varepsilon/2\right).$$

Here is where the proof departs from the usual. We will decompose the elements of G_n into layers, where the elements in each layer have approximately the same error. Specifically, layer i contains elements of G_n with error between $i\varepsilon$ and $(i+1)\varepsilon$. Let us say that an element of G_n that has true error greater than ε and empirical error at most $\varepsilon/2$ is *seductive*. We can apply the usual union bound to bound the probability that any member of a given layer is seductive by the product of the number of elements in that layer and the largest probability that any individual hypothesis in the layer is seductive. For the layers in which i is small, Lemma 5 ensures that there are not many hypotheses in the layer. For the layers in which i is large, the probability that each individual hypothesis is seductive is small enough to compensate for the potentially greater number of them.

We have

$$\begin{aligned}p \leqslant \sum_{i=1}^{\lceil 1/\varepsilon\rceil} &\Pr\left(\exists g \in G_n,\ i\varepsilon < \mathrm{er}_{f,U_n}(g) \leqslant (i+1)\varepsilon\right.\\ &\left.\text{and } \mathrm{er}_S(g) \leqslant \varepsilon/2\right).\end{aligned}$$

If c_3 is as in the statement of Lemma 5, applying that lemma and Lemma 3, we get

$$p \leqslant \sum_{i=1}^{\lceil 1/\varepsilon\rceil} \left(c_3(i+1)\right)^{n-1} \exp\left(-\left(1-\frac{1}{2i}\right)^2 i\varepsilon m/2\right).$$

Overestimating the first $i+1$ with e^i, and underestimating $1-\frac{1}{2i}$ by $1/2$, and rearranging a bit, we get that if $m > 8n/\varepsilon$, then

$$\begin{aligned}p &\leqslant c_3^n \sum_{i=1}^{\lceil 1/\varepsilon\rceil} \exp\left(i(n-\varepsilon m/8)\right)\\ &\leqslant \frac{c_3^n \exp(n-\varepsilon m/8)}{1-\exp(n-\varepsilon m/8)}\\ &= \frac{\exp((1+\ln c_3)n-\varepsilon m/8)}{1-\exp(n-\varepsilon m/8)}.\end{aligned}$$

Combining this with (4) and (5), we have

$$\begin{aligned}&\Pr\left(\mathrm{er}_{f,U_n}(h) > \varepsilon\right)\\ &\quad\leqslant \mathrm{e}^{\varepsilon m/12} + \frac{\exp((1+\ln c_3)n-\varepsilon m/8)}{1-\exp(n-\varepsilon m/8)}.\end{aligned}$$

From here simple calculations complete the proof. □

4. Related work

For classes of VC-dimension d, general bounds of $\mathrm{O}(\frac{d}{\varepsilon}\log\frac{1}{\varepsilon} + \frac{1}{\varepsilon}\log\frac{1}{\delta})$ [15,3] and $\mathrm{O}(\frac{d}{\varepsilon}\log\frac{1}{\delta})$ [7] are known. The only improvement on these that does not follow from Littlestone's analysis of PAC algorithms obtained from algorithms with mistake bounds [8], is the analysis of [1] of axis-aligned rectangles in $\mathbb{R}^n$. In that case, an optimal $\mathrm{O}(\frac{n+\log\frac{1}{\delta}}{\varepsilon})$ bound was also obtained, using a completely different technique. Servedio showed that a very simple algorithm efficiently

learns halfspaces with respect to the uniform distribution even in the presence of independent misclassification noise [12]; the best bound known on the number of examples used by this algorithm in the noise-free case is $\widetilde{O}(n/\varepsilon^2)$, however.

Acknowledgements

I'd like to thank Vinsensius Vega for his comments on a draft of this paper. I'd also like to acknowledge the support of National University of Singapore Academic Research Fund Grant RP3992710.

Appendix A. Proof of Lemma 1

The formula

$$V_n(1) = \frac{\pi^{n/2}}{\Gamma(n/2+1)}$$

is derived in many calculus books (see [13,4]). Applying Stirling's formula (see [11,5]) and simplifying yields

$$\begin{aligned}
&\sqrt{2\pi}\,\exp\left(\frac{1}{2}+\frac{5}{36n^2-30n}\right)\left(1-\frac{1}{n}\right)^{n/2}\\
&\quad\leqslant \frac{\sqrt{n}\,V_n(1)}{V_{n-1}(1)}\\
&\quad\leqslant \sqrt{2\pi}\,\exp\left(\frac{1}{2}+\frac{7}{36n^2-30n-6}\right)\left(1-\frac{1}{n}\right)^{n/2}.
\end{aligned}$$

Applying calculus to bound each factor independently yields

$$\frac{\sqrt{2\pi\mathrm{e}}}{2} \leqslant \frac{\sqrt{n}\,V_n(1)}{V_{n-1}(1)} \leqslant \sqrt{2\pi}\,\mathrm{e}^{23/39}\mathrm{e}^{-1/2},$$

which directly implies the lemma. □

Appendix B. Proof of Lemma 4

Note that $\rho_{U_n}(h_{\vec{v}}, h_{\vec{w}})$, the probability that $h_{\vec{v}}$ and $h_{\vec{w}}$ classify a uniformly randomly drawn point differently, is equal to the angle between $\vec{v}$ and $\vec{w}$ (in radians) divided by π. Also, choosing an element of H_n randomly by sampling its normal vector from the unit ball is equivalent to choosing an element of H_n by sampling its normal vector uniformly from the interior of the unit ball.

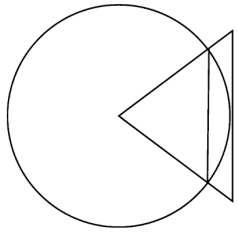

Fig. 2. An example of the cones used in the proof of Lemma 4. The region whose volume we want has part of the ball as part of its boundary; we lower and upper bound it by the volumes of an inscribed cone and a circumscribed cone respectively (here pictured as triangles).

We will use the latter view in our argument. Thus, we wish to approximate the volume in weight space of the collection of hypotheses h for which $\rho_{U_n}(g,h) \leqslant \alpha$. We can lower bound this volume by calculating the volume of the cone whose tip is the origin and whose base contains the elements of the unit ball whose angle with $\vec{v}$ is exactly $\alpha\pi$. Our upper bound will be obtained by calculating the volume of the smallest cone containing the region whose base is tangent to the unit ball. (See Fig. 2.) If $V_n(r)$ is the volume of a ball of radius r in $\mathbb{R}^n$,

$$\begin{aligned}
&\frac{1}{V_n(1)}\int_0^{\cos(\alpha\pi)} V_{n-1}\bigl(x\sin(\alpha\pi)\bigr)\,\mathrm{d}x\\
&\quad\leqslant \Pr_{\vec{w}\sim U_n}\bigl(\rho_n(h_{\vec{v}}, h_{\vec{w}}) \leqslant \alpha\bigr)\\
&\quad\leqslant \frac{1}{V_n(1)}\int_0^{1} V_{n-1}\bigl(x\sin(\alpha\pi)\bigr)\,\mathrm{d}x.
\end{aligned}$$

Using the fact that $V_{n-1}(r) = r^{n-1}V_{n-1}(1)$, we get

$$\begin{aligned}
&\frac{V_{n-1}(1)}{V_n(1)}\bigl(\sin(\alpha\pi)\bigr)^{n-1}\int_0^{\cos(\alpha\pi)} x^{n-1}\,\mathrm{d}x\\
&\quad\leqslant \Pr_{\vec{w}\sim U_n}\bigl(\rho_n(h_{\vec{v}}, h_{\vec{w}}) \leqslant \alpha\bigr)\\
&\quad\leqslant \frac{V_{n-1}(1)}{V_n(1)}\bigl(\sin(\alpha\pi)\bigr)^{n-1}\int_0^{1} x^{n-1}\,\mathrm{d}x
\end{aligned}$$

which implies

$$\begin{aligned}&\frac{V_{n-1}(1)}{nV_n(1)}\big(\sin(\alpha\pi)\big)^{n-1}\big(\cos(\alpha\pi)\big)^n\\&\quad\leqslant \Pr_{\vec{w}\sim U_n}\big(\rho_n(h_{\vec{v}},h_{\vec{w}})\leqslant\alpha\big)\\&\quad\leqslant \frac{V_{n-1}(1)}{nV_n(1)}\big(\sin(\alpha\pi)\big)^{n-1}.\end{aligned}$$

Lemma 1 then implies

$$\begin{aligned}&\frac{2}{\sqrt{n}}\big(\sin(\alpha\pi)\big)^{n-1}\big(\cos(\alpha\pi)\big)^n\\&\quad\leqslant \Pr_{\vec{w}\sim U_n}\big(\rho_n(h_{\vec{v}},h_{\vec{w}})\leqslant\alpha\big)\\&\quad\leqslant \frac{3}{\sqrt{n}}\big(\sin(\alpha\pi)\big)^{n-1}\end{aligned}$$

and the identity $\sin(u)\cos(u)=\sin(2u)/2$ yields

$$\begin{aligned}&\frac{2\cos(\alpha\pi)}{\sqrt{n}}\big(\sin(\alpha\pi)/2\big)^{n-1}\\&\quad\leqslant \Pr_{\vec{w}\sim U_n}\big(\rho_n(h_{\vec{v}},h_{\vec{w}})\leqslant\alpha\big)\\&\quad\leqslant \frac{3}{\sqrt{n}}\big(\sin(\alpha\pi)\big)^{n-1}.\end{aligned}$$

Approximating sin using its Taylor expansion,

$$\begin{aligned}&\frac{2\cos(\alpha\pi)}{\sqrt{n}}(\alpha\pi-4\alpha^3\pi^3/3)^{n-1}\\&\quad\leqslant \Pr_{\vec{w}\sim U_n}\big(\rho_{U_n}(h_{\vec{v}},h_{\vec{w}})\leqslant\alpha\big)\\&\quad\leqslant \frac{3}{\sqrt{n}}(\alpha\pi)^{n-1}\end{aligned}$$

and the lemma now follows from some straightforward calculations (using the assumption that $\alpha\leqslant 1/2$ for the lower bound along with the fact that $1\leqslant n^{1/2(n-1)}\leqslant 2$ for all $n\geqslant 2$). □

References

[1] P. Auer, P.M. Long, A. Srinivasan, Approximating hyper-rectangles: learning and pseudo-random sets, J. Computer System Sci. 57 (3) (1998) 376–388.

[2] G. Benedek, A. Itai, Learnability with respect to fixed distributions, Theoret. Comput. Sci. 86 (2) (1991) 377–389.

[3] A. Blumer, A. Ehrenfeucht, D. Haussler, M.K. Warmuth, Learnability and the Vapnik–Chervonenkis dimension, J. ACM 36 (4) (1989) 929–965.

[4] H.P. Boas, Peeling an onion, Tel Aviv University Math 696 course notes, 1995.

[5] B. Bollobas, Random Graphs, Academic Press, New York, 1985.

[6] T.M. Cover, Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition, IEEE Trans. Electron. Comput. EC-14 (1965) 326–334.

[7] D. Haussler, N. Littlestone, M.K. Warmuth, Predicting $\{0,1\}$-functions on randomly drawn points, Inform. and Comput. 115 (2) (1994) 129–161.

[8] N. Littlestone, From on-line to batch learning, in: Proceedings of the 1989 Workshop on Computational Learning Theory, 1989, pp. 269–284.

[9] P.M. Long, On the sample complexity of PAC learning half-spaces against the uniform distribution, IEEE Trans. Neural Networks 6 (6) (1995) 1556–1559.

[10] D. Pollard, Convergence of Stochastic Processes, Springer, Berlin, 1984.

[11] H. Robbins, A remark on Stirling's formula, Amer. Math. Monthly 62 (1955) 26–29.

[12] R. Servedio, On PAC learning using Winnow, Perceptron, and a Perceptron-like algorithm, in: Proceedings of the 1999 Conference on Computational Learning Theory, 1999, pp. 296–307.

[13] J. Stewart, Calculus, Brooks/Cole, 1995.

[14] L.G. Valiant, A theory of the learnable, Comm. ACM 27 (11) (1984) 1134–1142.

[15] V.N. Vapnik, Estimation of Dependencies based on Empirical Data, Springer, Berlin, 1982.