University of Illinois at Urbana-Champaign

Statistical Learning Theory

Final Project

Fat Shattering and Learnability of Real-Valued Functions

Student:
Puoya Tabaghi

Instructor:
Prof. Bruce Hajek

May 12, 2017

Contents

1 Introduction

In most definitions of learning, a learner sees a sequence of values of an unknown function at random points, and the goal is to choose an accurate approximation to that function, with high probability. We already showed that for binary valued functions, Vapnik Chervonenkis (VC)-dimension of a function class characterizes its learnability in the sense that a function class is learnable if and only if its VC dimension is finite. Kearns and Schapire [2] introduced a generalization of the VC dimension, which we call the fat-shattering function, and showed that a class of probabilistic concepts is learnable only if the class has a finite fat-shattering function. In this project, we consider the learnability of $[0, 1]$-valued function classes. We show that a class of $[0, 1]$-valued functions is learnable from a finite training sample with observation noise satisfying some mild conditions if and only if the class has a finite fat-shattering function. The main goal is to show that the finiteness of the fat-shattering function is necessary for learning. We also consider small-sample learnability, for which the sample size is allowed to grow only polynomially with the required performance parameters.

2 Definitions and Main Result

2.1 Classes of Noise Distributions

A function f is said to have bounded variation if there is a constant $C > 0$ such that for every ordered sequence $x_0 < \cdots < x_n$ in $\mathbb{R}$ we have

$$\sum_{k=1}^{n} |f(x_k) - f(x_{k-1})| \leq C$$

In that case, the *total variation* of f on $\mathbb{R}$ is

$$V(f) = \sup\{\sum_{k=1}^{n} |f(x_k) - f(x_{k-1})| : x_0 < \cdots < x_n\}$$

Definition: An *admissible* noise distribution class $\mathcal{D}$ is a class of distributions on $\mathbb{R}$ that satisfies

- Each distribution in $\mathcal{D}$ has zero mean and finite variance

- Each distribution in $\mathcal{D}$ is absolutely continuous and its pdf has bounded variation: there is a function $v : \mathbb{R}^+ \to \mathbb{R}^+$, called the *total variation function*, such that if f is the pdf of a distribution in $\mathcal{D}$ with variance σ^2, then $V(f) < v(\sigma)$.

If $\mathcal{D}$ also satisfies the following condition, we say it is a *bounded addmissible* distribution class.

- There is non-decreasing function $s : \mathbb{R}^+ \to \mathbb{R}^+$, called the support function, such that if $D \in \mathcal{D}$ with variance σ^2, then the support of D is contained in a closed interval of length $s(\sigma)$.

If $\mathcal{D}$ is an admissible noise distribution, we say it is an *almost-bounded admissible* distribution class.

- Each distribution $D \in \mathcal{D}$ has an even pdf and light tails: there are constants s_0 and c_0 in R^+ such that, for all distribution $D \in \mathcal{D}$ with variance σ^2, and all $s > s_0\sigma$

$$D\{\eta : |\eta| > s/2\} \leq c_0 e^{-s/\sigma}$$

2.2 The Learning Problem

Choose a set F of functions from X to $[0, 1]$. For $m \in \mathbb{N}$, $f \in F$, $x \in X^m$ and $\eta \in \mathbb{R}^m$, let

$$\text{sam}(x, \eta, f) = ((x_1, f(x_1) + \eta_1), \cdots, (x_m, f(x_m) + \eta_m)).$$

A deterministic learning algorithm is defined to be a mapping from $\cup_m (X \times \mathbb{R})^m$ to $[0, 1]^X$. A randomized learning algorithm L is a pair (A, P_Z) where P_Z is a distribution on set Z and A is a mapping from $\cup_m (X \times \mathbb{R})^m \times Z^m$ to $[0, 1]^X$. Given a sample of length m the randomized algorithms chooses a sequence $z \in Z^m$ at random from P_Z^m and passes it to the deteministic mapping A as a parameter . For probability distribution P on X, $f \in F$ and $h : X \to [0, 1]$ define

$$\text{er}_{P,f}(h) = \int_X |h(x) - f(x)| dP(x).$$

Definition: Let $\mathcal{D}$ be a class of distributions on $\mathbb{R}$. Choose $0 < \epsilon, \delta < 1$ and $\sigma > 0$ and $m \in \mathbb{N}$. We say a learning algorithm $L = (A, P_Z)$ $(\epsilon, \delta, \sigma)$-learns F from m

examples with noise D if for all distributions P on X, all functions $f \in F$ and all distributions $D \in \mathcal{D}$ with variance σ^2

$$P^m \times D^m \times {P_Z}^m\{(x, \eta, z) : \mathrm{er}_{P,f}(A(\mathrm{sam}(x, \eta, f), z)) \geq \epsilon\} < \delta$$

Similarly, L (ϵ, δ)-learns F from m examples withous noise if for all distributions P on X and all functions $f \in F$

$$P^m \times {P_Z}^m\{(x, z) : \mathrm{er}_{P,f}(A(\mathrm{sam}(x, 0, f), z)) \geq \epsilon\} < \delta.$$

Function class F is *learnable* with noise $\mathcal{D}$ if there is a learning algorithm L and a function $m_0 : (0,1) \times (0,1) \times \mathbb{R}^+ \to N$ such that for all $0 < \epsilon, \delta < 1$ for all $\sigma > 0$ algorithm L $(\epsilon, \delta, \sigma)$-learns F from $m_0(\epsilon, \delta, \sigma)$ examples with noise $\mathcal{D}$. Function class F is *small sample learnable* with noise $\mathcal{D}$ if in addition the function m_0 is bounded above by a polynomial in $1/\epsilon$, $1/\delta$, and σ.

Choose $x_1, \cdots, x_d \in X$. We say $x_1, \cdots, x_d \in X$ are γ-shattered by F if there exists $r \in [0,1]^d$ such that for each $b \in \{0,1\}^d$ there is an $f \in F$ (a witness) such that for each i

$$f(x_i) \begin{cases} \geq r_i + \gamma & \text{if} \quad b_i = 1 \\ \leq r_i - \gamma & \text{if} \quad b_i = 0. \end{cases}$$

where γ is the width of shattering, [2]. A geometric interpretation of this definition is to regard $(r_1, \cdots, r_d)$ as the origin of a coordinate system in d-dimensional Euclidean space; Then $\mathcal{F}$ shatters x's if the set $\{(f(x_1), \cdots, f(x_d)) : f \in \mathcal{F}\}$ intersects all 2^d orthants of the coordinate system at least γ far away from origin. For each γ, let $\mathrm{fat}_F(\gamma) = \max\{d \in \mathbb{N} : F \text{ } \gamma\text{-shatters some } x_1, \cdots x_d\}$ if such a maximum exists, and ∞ otherwise. If $\mathrm{fat}_F(\gamma)$ is finite for all γ we say F has a finite fat-shattering function.

2.3 The Main Result

Theorem 1, [1]: Suppose F is a permissible class of $[0,1]$ valued functions defined on X. If $\mathcal{D}$ is a bounded admissble distribution class then F is learnable with observation noise $\mathcal{D}$ if and only if F has finite fat shattering function. If $\mathcal{D}$ is an almost bounded admissible distrbution class then F is small sample learnable with observation noise $\mathcal{D}$ if and only if there is a polynomial p that satisfies $\mathrm{fat}_F(\gamma) < p(1/\gamma)$ for all $\gamma > 0$.

3 Lower Bound

In this section, we give a lower bound on the number of examples necessary to learn a real-valued function class in the presence of observation noise.

3.1 Learnability With Noise Implies Quantization Learnability

In this subsection, we try to relate the problem of learning a real valued function class with observation noise to the problem of learning a quantized version of that class without noise.

Definition: For $\alpha \in \mathbb{R}^+$ define the quantization function

$$Q_\alpha(y) = \alpha \lceil \frac{y - \alpha/2}{\alpha} \rceil$$

For a set $S \subset \mathbb{R}$, let $Q_\alpha(S) = \{Q_\alpha(y) : y \in S\}$ for a function class $F \in [0,1]^X$, let $Q_\alpha(F)$ be the set $\{Q \circ f : f \in F\}$ of $Q_\alpha([0,1])$ valued functions defined on X.

We want to show that an algorithm that can learn a real-valued function class with observation noise can be used to construct an algorithm that can learn a quantized version of the function class to slightly worse accuracy and confidence with the same number of examples, provided the quantization width is sufficiently small. To do so, we need to prove the following lemma which will be useful later.

Lemma 1: Let $\mathcal{D}$ be an admissible noise distribution class with total variation function v. Let $\sigma > 0$ and $0 < \alpha < 1$. Let D be a distribution in $\mathcal{D}$ with variance σ^2. Let η, ζ and ν be random variables and suppose that η and ν are distributed according to D and ζ is distributed uniformly on $[-\alpha/2, \alpha/2]$

1. For any $y \in [0,1]$ if P_1 is the distribution of $y + \eta$ and P_2 is the distribution of $Q_\alpha(y + \eta) + \zeta$, we have

$$d_{TV}(P_1, P_2) \leq \alpha v(\sigma)$$

2. For any $y \in [0,1]$ if P_3 is the distribution of $Q_\alpha(y + \eta)$ and P_4 is the distribution of $Q_\alpha(y) + Q_\alpha(\nu)$ we have

$$d_{TV}(P_3, P_4) \leq \alpha v(\sigma)$$

Proof of Lemma 1: 1. Let p be the pdf of D. Thus, the pdf of $y+\eta$ would be $p_1(t) = p(t-y)$. You can easily show that the pdf of $Q_\alpha(y+\eta)+\zeta$ would be

$$p_2(t) = \frac{1}{\alpha}\int_{Q_\alpha(t)-\alpha/2}^{Q_\alpha(t)+\alpha/2} p(x-y)dx.$$

So,

$$\begin{aligned}
d_{TV}(P_1,P_2) &= \int_{-\infty}^{+\infty} |p_1(x)-p_2(x)|dx \\
&= \int_{-\infty}^{+\infty} |p(x-y) - \frac{1}{\alpha}\int_{Q_\alpha(x)-\alpha/2}^{Q_\alpha(x)+\alpha/2} p(\theta-y)d\theta|dx \\
&= \int_{-\alpha/2}^{\alpha/2} \sum_{n=-\infty}^{\infty} |p(x-y+n\alpha) - \frac{1}{\alpha}\int_{-\alpha/2}^{\alpha/2} p(\theta-y+n\alpha)d\theta|dx \\
&\leq \int_{-\alpha/2}^{\alpha/2} \sum_{n=-\infty}^{\infty} \sup_{z\in(-\alpha/2,\alpha/2)} |p(x-y+n\alpha) - p(z-y+n\alpha)|dx \\
&\leq \alpha v(\sigma)
\end{aligned}$$

Note that the mean value theorem is employed for the first inequality: $\exists z_1, z_2 \in [-\alpha/2, \alpha/2]$ s.t.

$$p(z_1-y+n\alpha) \leq \frac{1}{\alpha}\int_{-\alpha/2}^{\alpha/2} p(\theta-y+n\alpha)d\theta \leq p(z_2-y+n\alpha)$$

2. The distribution of $Q_\alpha(y+\eta)$ is discrete, and satisfies

$$P_3(t) = \int_{n\alpha-\alpha/2}^{n\alpha+\alpha/2} p(x-y)dx$$

if $t = n\alpha$ for some $n \in \mathbb{Z}$, and $P_3(t) = 0$ otherwise. Since v has distribution D, the distribution P_4 of the random variable $Q_\alpha(y)+Q_\alpha(v)$ is also discrete, and satisfies,

$$P_4(t) = \int_{n\alpha-\alpha/2}^{n\alpha+\alpha/2} p(x)dx$$

if $t = n\alpha + Q_\alpha(y)$ for some $n \in \mathbb{Z}$, and $P_4(t) = 0$ otherwise. So,

$$\begin{aligned} d_{TV}(P_3, P_4) &= \sum_{n=-\infty}^{\infty} |P_3(n\alpha) - P_4(n\alpha)| \\ &= \sum_{n=-\infty}^{\infty} |\int_{n\alpha-\alpha/2}^{n\alpha+\alpha/2} p(x-y)dx - \int_{n\alpha-\alpha/2}^{n\alpha+\alpha/2} p(x-Q_\alpha(y))dx| \\ &\leq \int_{-\alpha/2}^{\alpha/2} \sum_{n=-\infty}^{\infty} |p(x-y+n\alpha) - p(x-Q_\alpha(y)+n\alpha)|dx \\ &\leq \alpha v(\sigma) \end{aligned}$$

Lemma 2: Suppose F is a set of functions from X to $[0,1]$, $\mathcal{D}$ is an admissible noise distribution class with total variation function v, A is a learning algorithm, $0 < \epsilon, \delta < 1$, $\sigma \in \mathbb{R}^+$, $m \in \mathbb{N}$. If the quantization width $\alpha \in \mathbb{R}^+$ satisfies

$$\alpha \leq \min(\frac{\delta}{v(\sigma)m}, 2\epsilon)$$

and A $(\epsilon, \delta, \sigma)$-learns F from m examples with noise $\mathcal{D}$ then there is a randomized learning algorithm (C, P_Z) that $(2\epsilon, 2\delta)$-learns $Q_\alpha(F)$ from m examples.

Proof of Lemma 2: We will describe the a randomized algorithm (C) that is constructed from algorithm A and show that it $(2\epsilon, 2\delta)$-learns the quantized function class $Q_\alpha(F)$ (see Figure 1). Fix a noise distribution D in $\mathcal{D}$ with variance σ^2, a function $f \in F$ and a distribution P on X. Since A $(\epsilon, \delta, \sigma)$-learns F, we have

$$P^m \times D^m\{(x, \eta) : \mathrm{er}_{P,f}(A(\mathrm{sam}(x, \eta, f))) \geq \epsilon\} < \delta$$

That is the probability that Algorithm A chooses a bad function is small. We will show that this implies that the probability that algorithm C chooses a bad function is also small, where the probability is over all $x \in X^m$ and all values of the random variables that algorithm C uses. Now, fix a sequence $x = (x_1, \cdots, x_m) \in X^m$, and define the events

$$E = \{\eta \in \mathbb{R}^m : \mathrm{er}_{P,f}(A(\mathrm{sam}(x, \eta, f))) \geq \epsilon\}$$

$$E_1 = \{y \in \mathbb{R}^m : \mathrm{er}_{P,f}(A(x_1, y_1, \cdots, x_m, .y_m)) \geq \epsilon\}$$

that is E is the set of noise sequences that makes A chooses a bad function and E_1 is the corresponding set of y sequences.

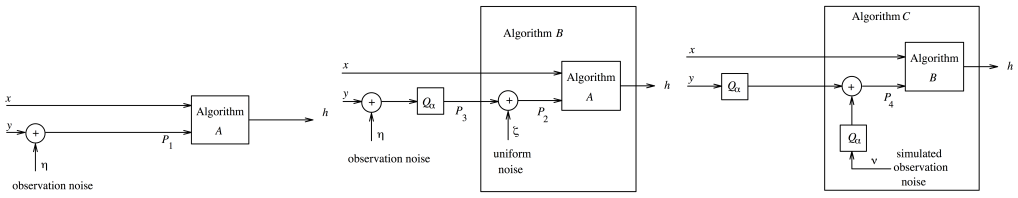

Figure 1: Learning algorithm for real-valued functions can be used to construct a learning algorithm for quantized functions.

Clearly

$$D^m(E) = (\prod_{i=1}^{m} P_{1|x_i})(E1)$$

where $P_{1|x_i}$ is the distribution of $f(x_i) + \eta$. Let ζ be a random variable with distribution U_α, where U_α is the uniform distribution on $(-\alpha/2, \alpha/2)$. Let algorithm B be the randomized algorithm that adds noise ζ to each y value it recieves and passes the sequence to algorihm A . That is,

$$B(x_1, y_1, \cdots, x_m, y_m) = A(x_1, y_1 + \zeta_1, \cdots, x_m, y_m + \zeta_m)$$

Let $P_{2|x_i}$ be the distribution of $Q_\alpha(f(x_i)+\eta)+\zeta$. From lemma 1, $d_{TV}(P_{1|x_i}, P_{2|x_i}) \leq \alpha v(\sigma)$.

Lemma 3: If P_i and Q_i are distributions on a set $Y (i = 1, \cdots, m)$, and E is a measurable subset of Y^m, then

$$|(\prod_{i=1}^{m} P_i)(E) - (\prod_{i=1}^{m} Q_i)(E)| \leq \frac{1}{2} \sum_{i=1}^{m} d_{TV}(P_i, Q_i)$$

Lemma 3 implies

$$(\prod_{i=1}^{m} P_{2|x_i})(E1) \leq D^m(E) + \frac{m\alpha v(\sigma)}{2} \leq D^m(E) + \delta/2$$

where the second inequality follows from the hypothesis that $\alpha \leq \delta/(mv(\sigma))$. Let $P_{3|x_i}$ be the distribution of $Q_\alpha(f(x_i) + \eta)$ and let

$$E_3 = \{y \in \mathbb{R}^m : E(\mathrm{er}_{P,f}(B(x_1, y_1, \cdots, x_m, y_m))) \geq \epsilon\}$$

where the expectation is over all values of ζ the uniform noise that B introduces. In this case, E_3 is the set of y sequence that make B chooses bad function. Clearly

$$(\prod_{i=1}^{m} P_{3|x_i})(E_3) = (\prod_{i=1}^{m} P_{2|x_i})(E1)$$

Let v be a random variable with distribution D. Let algorithm C be the randomized algorithm that adds noise $Q_\alpha(v)$ to each y value it receives and passes the sequence to algorithm B. that is

$$C(x_1, y_1, \cdots, x_m, y_m) = B(x_1, y_1 + Q_\alpha(v_1), \cdots, x_m, y_m + Q_\alpha(v_m))$$

let $p_{4|x_i}$ be the distribution of $Q_\alpha(f(x_i))+Q_\alpha(v)$. From lemma 1, $d_{TV}(P_{4|x_i}, P_{3|x_i}) \leq \alpha v(\sigma)$ and lemma 3 implies

$$(\prod_{i=1}^{m}(P_{4|x_i}))(E_3) \leq (\prod_{i=1}^{m}(P_{3|x_i}))(E_3) + \frac{m\alpha v(\sigma)}{2} \leq D^m(E) + \delta$$

It follows that the probability under $P^m \times U_\alpha^m \times D^m$ that x, ζ, and v satisfy

$$\begin{aligned} &\mathrm{er}_{P,f}(A(x_1, Q_\alpha(f(x_1)) + Q_\alpha(v_1) + \zeta_1, \cdots, x_m, Q_\alpha(f(x_m)) + Q_\alpha(v_m) + \zeta_m)) = \\ &\mathrm{er}_{P,f}(C(x_1, Q_\alpha(f(x_1)), \cdots, x_m, Q_\alpha(f(x_m)))) \geq \epsilon \end{aligned} \tag{1}$$

is less than 2δ. Since $\alpha \leq 2\epsilon$, for all $x \in X$, $|f(x) - Q_\alpha(f(x))| \leq \epsilon$ and therefore (1) implies

$$\mathrm{er}_{P,Q_\alpha(f)}(C(x_1, Q_\alpha(f(x_1)), \cdots, x_m, Q_\alpha(f(x_m)))) < 2\epsilon$$

with probability at least $1 - 2\delta$. This is true for any $Q_\alpha(f)$ in $Q_\alpha(F)$, so this algorithm $(2\epsilon, 2\delta)$-learns $Q_\alpha(F)$ from m examples.

3.2 Lower Bounds for Quantized learning

In the previous subsection, we showed that if a class F can be $(\epsilon, \delta, \sigma)$-learned with a certain number of examples, then an associated class $Q_\alpha(F)$ of discrete-valued functions can be $(2\epsilon, 2\delta)$-learned with the same number of examples. In this subsection, we show that an algorithm for learning a class of discrete-valued functions can effectively be used as a subroutine in an algorithm for learning binary-valued functions. We then apply a lower bound result for binary-valued functions.

Definition: For each $d \in \mathbb{N}$, let $POWER_d$ be the set of all functions from $\{1, \cdots, d\}$ to $\{0, 1\}$.

Theorem 2 [3]: Let C be a nontrivial, well-behaved concept class. If the VC dimension of C is $d < \infty$, then for any $0 < \epsilon < 1/2$ and sample size less than

$$\max(\frac{1-\epsilon}{\epsilon} \ln \frac{1}{\delta}, d(1-2(\epsilon(1-\delta)+\delta)))$$

no function $A : S_C \to H$, for any hypothesis space H, is a learning function for C.

Corollary 1: Let A be a randomized learning algorithm which always outputs $\{0,1\}$ valued hypothesis. If A is given fewer than $d/2$ examples, A fails to $(1/8, 1/8)$-learn $POWER_d$. (For proof, replace $(\epsilon, \delta) = (1/8, 1/8)$ in Theorem 2.)

Lemma 4: Choose a set F of functions from X to $Q_\alpha([0,1])$, $d \in \mathbb{N}$ and $\gamma > 0$ such that $\text{fat}_F(\gamma) \geq d$. If a randomized learning algorithm A is given fewer than

$$\frac{d-400}{4+192\ln 1/\alpha}$$

examples, A fails to $(\gamma/32, 1/16)$-learn F without noise.

Proof Sketch of Lemma 4: We start with the contrapositive statement and assume that A $(\gamma/32, 1/16)$-learn F without noise. Then, we construct another algorithm $\tilde{A}$ that $(1/8, 1/8)$-learns $POWER_d$ from $m + \lceil 96(\ln 8 + m \ln \lceil 1/\alpha \rceil) \rceil$ examples without noise. Then, we apply Corollary 1 and it completes the proof.

3.3 The Lower Bound

Theorem 3: Suppose F is a set of $[0,1]$-valued functions defined on X, $\mathcal{D}$ is an admissible noise distribution class with total variation function v, $0 < \gamma < 1$, $0 < \epsilon \leq \gamma/65$, $0 < \delta \leq 1/32$, $\sigma \in \mathbb{R}^+$, and $d \in \mathbb{N}$. If $\text{fat}_F(\gamma) \geq d > 800$, then any algorithm that $(\epsilon, \delta, \sigma)$-learns F with noise $\mathcal{D}$ requires at least m_0 examples, where

$$m_0 > \min\{\frac{d}{800\ln(2+dv(\sigma)/10)}, \frac{d}{800\ln(2+d/120)}, \frac{d}{400\ln(40/\gamma)}\}$$

Proof Sketch of Theorem 3: Combine Lemmas 4 and 2.

4 The Upper Bound

In this section, we present a theorem that gives an upper bound on the number of examples required to learn a function class with noise. This will finishes the proof

of Theorem 1. We omitted the proof since it will lengthen the report. One can refer to [1] for the proof.

Theorem 4: For any permissible class F of functions from X to $[0,1]$, there is a learning algorithm A such that, for all bounded admissible distribution classes $\mathcal{D}$ with support function s, for all probability distributions P on X, and for all $0 < \epsilon < 1/2$, $0 < \delta < 1$ and $\sigma > 0$, if $d = \text{fat}_F(\frac{\epsilon^2}{576s(\sigma+1)})$, then $A(\epsilon, \delta, \sigma)$-learns F from

$$\frac{1152(1+s(\sigma))^4}{\epsilon^4}(12d(25+\ln\frac{d(1+s(\sigma))^6}{\epsilon^8})^2+\ln\frac{4}{\delta})$$

examples with noise $\mathcal{D}$.

Corrolary 2: Let F be a class of functions from X to $[0,1]$. Let p be a polynomial and suppose $\text{fat}_F(\gamma) < p(1/\gamma)$ for all $0 < \gamma < 1$. Then for any almost-bounded admissible distribution class $\mathcal{D}$, F is small-sample learnable with noise $\mathcal{D}$.

Proof Sketch for Corrolary 2: We construct a bounded distribution by cutting the tail of the almost-bounded noise distribution. Then, we show that total variation distance between these two can be made small enough so that it doesn't voilate the statements presented in the upper bound proof.

5 Conclusions

In this project, we consider the problem of learning real-valued functions from random examples when the function values are corrupted with noise. With mild conditions on independent observation noise, we provide characterizations of the learnability of a real-valued function class in terms of a generalization of the Vapnik-Chervonenkis dimension, the fat-shattering function, introduced by Kearns and Schapire. We show that, given some restrictions on the noise, a function class is learnable in our model if and only if its fat-shattering function is finite. With different restrictions, satisfied for example by gaussian noise, we show that a function class is learnable from polynomially many examples if and only if its fat-shattering function grows polynomially.

References

[1] Bartlett, Peter L., Philip M. Long, and Robert C. Williamson. "Fat-shattering and the learnability of real-valued functions." Journal of Computer and System Sciences 52.3 (1996): 434-452.

[2] Kearns, Michael J., and Robert E. Schapire. "Efficient distribution-free learning of probabilistic concepts." Journal of Computer and System Sciences 48.3 (1994): 464-497.

[3] Blumer, Anselm, et al. "Learnability and the Vapnik-Chervonenkis dimension." Journal of the ACM (JACM) 36.4 (1989): 929-965.