Learning from Noisy Data and Markovian Processes

V L Raju Chinthalapati

University of Greenwich, London SE10 9LS, UK

Abstract

We discuss more realistic models of computational learning. We extend the existing literature on the Probably Approximately Correct (PAC) framework to finite Markov chains in two directions by considering: (1) the presence of classification noise (specifically assuming that the training data has currupted labelled examples), and (2) real valued function learning. In both cases we address the key issue of determining how many training examples must be presented to the learner in the learning phase for the learning to be successful under the PAC paradigm.

Keywords:
PAC Learning, Noisy Data, VC dimension, Classification Noise, Markovian Process, Real-valued and Boolean-valued Function Learning.

Email address: cv20@gre.ac.uk (V L Raju Chinthalapati)

April 18, 2015

1. Introduction

Supervised learning is a machine learning technique where the learner receives correct answers for all examples during the training period in order to infer the target function. In supervised learning, a learner receives example data and uses a learning method in order to get a hypothesis that fits well to the seen and, hopefully, to the unseen data. The example data consist of pairs (x, y), where x and y are input from an instance space X and output set, Y respectively. We call (x, y), a labelled input or a training example. In the case of seen data (or training data), the learner is provided with a teacher (an oracle) $EX(t, P)$ such that each call to $EX(t, P)$ for an instance returns a labelled instance (x, y) according to target hypothesis t; that is, $y = t(x)$. Here, x is drawn according to the probability distribution P. Here, P is a fixed, but unknown distribution on the instance space X. It is important that both training and unseen data instances are drawn from the same distribution P. It is our hope that after seeing enough examples generated by the oracle, a good learner would be in a position to infer an accurate hypothesis for yet unseen data. Informally, after seeing a large enough training sample drawn from the distribution P, a successful learner can deliver with high probability a hypothesis that almost correctly matches with the target hypothesis. For the above probabilistic setting, Valiant's [11] probably approximately correct (PAC) gives a general criterion of successful learning and formalises the terms high probability and almost correctly.

In this paper, we discuss more realistic models of computational learning. First of all, we consider Boolean-valued function learning in the presence of classification noise. In many real life applications, machine learning algorithms should perform well in the presence of noise. We discuss a more general PAC model that considers classification noise [1]. In the presence of classification noise, the learner has access to a corrupted oracle $EX_\eta(t, P)$. We assume that $EX_\eta(t, P)$ is able to generate instances $x \in X$ according to distribution P but it differs from $EX(t, P)$ by attributing an incorrect label to x subject to some unknown independent noise parameter $\eta < \frac{1}{2}$. Explicitly, the experiment performed by $EX_\eta(t, P)$ involves drawing an instance $x \in X$ according to distribution P, and then flipping a coin that comes up tails with probability η. If the coin comes up tails, $EX_\eta(t, P)$ reports x with the opposite of the label $t(x)$; otherwise, it reports x with correct label $t(x)$. Many of the PAC learning results can be preserved in the presence of classification

noise. Sloan [10] studies PAC learning in the presence of (1) malicious noise [6], (2) malicious misclassification noise, and (3) random attribute noise. We restrict ourselves to the PAC learning framework in the presence of classification noise and discuss how to upper bound the minimum number of noisy examples that are required for successful PAC learning. We obtain a new result involving the VC dimension. Secondly, in many real life examples, in addition to the classification noise, the training samples are not generated according to independent and identically distributed processes, but instead form a Markovian process. A key issue in that case is to determine how many training examples must be presented to the learner in the learning phase for the learning to be successful. For that, we generalise Gamarnik's work [3] to the learning in the presence of noisy examples that follow a Markov chain. Finally, we extend Gamarnik's work [3] on pattern classification problems to the case of learning real valued functions in a Markovian setting.

2. Boolean-valued function learning in the presence of classification noise

We start by describing the standard setting of the PAC model of learning. Let H be the set of all hypotheses that can be computed by the learner and C be the set of possible target hypotheses. Let $Z = X \times Y$. Then the learner receives a sequence of training examples $z = ((x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m)) = (z_1, z_2, \ldots, z_m) \in Z^m$, where $x_i \in X$ is drawn from distribution P. The sample $(x_1, x_2, \ldots, x_m)$ is drawn from X^m according to the probability distribution P^m on X^m. Here, $y_i = t(x_i)$ is provided by the oracle according to target hypothesis $t \in C$. The target concept can be either Boolean valued or real valued: in case of a Boolean valued target function, $Z = X \times \{0, 1\}$ (in case of real valued target function, $Z = X \times \Re$). In this section, we focus our attention on Boolean valued functions. After seeing a high enough number of training examples, the learner that uses learning algorithm L must output a hypothesis h estimating the target hypothesis t. The learning algorithm can be considered as a function that maps the set of all training samples Z^m, for all m, onto the hypothesis set H:

$$L : \cup_{m=1}^{\infty} Z^m \rightarrow H.$$

The performance of the hypothesis h output by L is thought of in terms of how accurate h will be on subsequent instances drawn from the distribution

P. In the case of Boolean valued functions, the true error of hypothesis h with respect to distribution P and target hypothesis t is the probability that h will misclassify an instance drawn with distribution P. Mathematically, the error is defined as follows:

$$er_P(h) = P\{x \in X : h(x) \neq t(x)\}.$$

The sample error of hypothesis h is a measure of its incompatibility with the target hypothesis on the training examples $z = (z_1, z_2, \ldots, z_m)$ and is defined as follows:

$$er_z(h) = \frac{1}{m}|\{i : 1 \leq i \leq m \ , \ h(x_i) \neq y_i\}|.$$

One can realise that, in general, it is difficult to find the true error. Since the sample error can be considered as an estimator for the true error, by using existing statistical results, we can bound $er_P(h)$ in terms of $er_z(h)$, with high probability.
We define the *disagreement number* to be the number of labelled instances (x, y) in the training sample which are such that $y \neq t(x)$. We have the following result of Angluin and Laird [1]. Here, η_b is a known upper bound on η. It is possible that the exact value of η might not be known.

Theorem 2.1 (Angluin and Laird). *If we draw a sample of size*

$$m = \frac{2}{\epsilon^2(1-2\eta_b)^2} \ln\left(\frac{2|H|}{\delta}\right)$$

from $EX_\eta(t, P)$ where $\eta \leq \eta_b < \frac{1}{2}$, where $t \in H$, and find any hypothesis $h \in H$ with minimal disagreement number, then

$$P^m(er_P(h) > \epsilon) \leq \delta.$$

Definition 2.1. *$S \subseteq X$ is shattered by H if and only if for every $R \subseteq S$, there exists $h \in H$ such that $h(x) = 1$ for all $x \in R$ and $h(x) = 0$ for all $x \in S \backslash R$. A set $S \in X$ is shattered by H if and only if for every dichotomy of S, there exists a hypothesis $h \in H$ that is consistent with the dichotomy, that means, h divides S into the two subsets.*

VC Dimension: The VC dimension VC(H) of the hypotheses set H defined over an instance set X is the cardinality of the largest subset of X that is shattered by H. If there are arbitrarily large sets that are shattered by H, then $VC(H) = \infty$.

We have the following new result which involves the VC dimension and the following result of Vapnik and Chervonenkis ([12]; see also *Theorem 4.3* from [2]).

Lemma 2.1 (Vapnik and Chervonenkis). *Suppose that H is a set of $\{0, 1\}$ - valued functions defined on a set X and that P is a probability distribution on $Z = X \times \{0, 1\}$. For $0 < \epsilon < 1$ and m a positive integer, we have*

$$P^m\{|er_P(h) - er_z(h)| \geq \epsilon \ \ for\ some\ h \in H\} \leq 4\left(\tfrac{2em}{d}\right)^d e^{\frac{-\epsilon^2 m}{8}}.$$

Theorem 2.2. *Let d be the VC dimension of the hypothesis set H. If we draw a sample of size*

$$m \geq \frac{64}{(1-2\eta)^2\epsilon^2}\left[d\ln\left(\frac{128}{(1-2\eta)^2\epsilon^2}\right) + \ln\left(\frac{8}{\delta}\right)\right],$$

from $EX_\eta(t, P)$ for any $\eta \leq \eta_b < \frac{1}{2}$, and find any hypothesis $h \in H$ with minimal disagreement number, then

$$P^m(er_P(h) > \epsilon) \leq \delta.$$

Proof: Let Q be the distribution on $X \times [0, 1]$ giving the same distribution on $X \times [0, 1]$ as obtained by choosing inputs from X according to P and then noisily labelling it. From the construction of the probability measure Q, it is clear that

$$\forall\ h \in H, er_Q(h) = \eta + (1-2\eta)er_P(h)$$

and therefore

$$P^m(\exists\ h \in H\ with\ er_P(h) \geq \epsilon, er_z(h) < \eta + \frac{s}{2}) \leq$$

$$Q^m(\exists\, h \in H \text{ with } er_Q(h) \geq \eta + s, er_z(h) < \eta + \frac{s}{2}),$$

where $s = (1-2\eta)\epsilon$. So,

$P^m(\exists\, h$ with $er_P(h) > \epsilon$ which minimises disagreements on $z)$

$\leq P^m(er_z(t) \geq \eta + \frac{s}{2}) + P^m(\exists\ \ h \text{ with } er_P(h) \geq \epsilon, er_z(h) \leq \eta + \frac{s}{2})$

$\leq P^m(er_z(t) \geq \eta + \frac{s}{2}) + Q^m(\exists\, h \in H \text{ with } er_Q(h) \geq \eta + s, er_z(h) < \eta + \frac{s}{2})$

$\leq P^m(er_z(t) \geq \eta + \frac{s}{2}) + Q^m(\exists\, h \in H \text{ with } er_Q(h) \geq er_z(h) + \frac{s}{2})$

$\leq P^m(er_z(t) \geq \eta + \frac{s}{2}) + Q^m(\exists\, h \in H \text{ with } |er_Q(h) - er_z(h)| \geq \frac{s}{2})$

Using Lemma 2.1,

$$Q^m(\exists\, h \in H \text{ with } |er_Q(h) - er_z(h)| \geq \frac{s}{2}) \leq 4\left(\frac{2em}{d}\right)^d e^{-\frac{s^2 m}{32}}.$$

Now,

$$4\left(\frac{2em}{d}\right)^d e^{-\frac{s^2 m}{32}} \leq \frac{\delta}{2} \tag{1}$$

if

$$\frac{s^2}{4} \geq \frac{8}{m} \ln\left(8\frac{\left(\frac{2em}{d}\right)^d}{\delta}\right),$$

which means

$$m \geq \frac{32}{s^2}\left(d\ln m + d\ln\left(\frac{2e}{d}\right) + \ln\left(\frac{8}{\delta}\right)\right). \tag{2}$$

Since $\ln x \leq \alpha x - \ln\alpha - 1$ for all $\alpha, x > 0$, then for $\alpha = \frac{s^2}{64d}$,

$$\ln m \leq \frac{s^2}{64d} m - \ln\left(\frac{s^2}{64d}\right) - 1.$$

So,

$$\frac{32}{s^2}\left(d\ln m + d\ln\left(\frac{2e}{d}\right) + \ln\left(\frac{8}{\delta}\right)\right) \ \leq\ \frac{m}{2} + \frac{32d}{s^2}\ln\left(\frac{64d}{s^2}\right) - \frac{32d}{s^2}$$

$$
\begin{aligned}
& + \frac{32d}{s^2} \ln\left(\frac{2e}{d}\right) + \frac{32}{s^2} \ln\left(\frac{8}{\delta}\right) \\
= \quad & \frac{m}{2} + \frac{32d}{s^2} \ln\left(\frac{128}{s^2}\right) \\
& + \frac{32}{s^2} \ln\left(\frac{8}{\delta}\right). \qquad (3)
\end{aligned}
$$

From (2), (3) and (5) it suffices for (1) to hold to have

$$m \geq \frac{m}{2} + \frac{32d}{s^2} \ln\left(\frac{128}{s^2}\right) + \frac{32}{s^2} \ln\left(\frac{8}{\delta}\right),$$

that is,

$$m \geq \frac{64}{s^2}\left(d \ln\left(\frac{128}{s^2}\right) + \ln\left(\frac{8}{\delta}\right)\right),$$

so,

$$m \geq m_0 = \frac{64}{(1-2\eta)^2\epsilon^2}\left(d \ln\left(\frac{128}{(1-2\eta)^2\epsilon^2}\right) + \ln\left(\frac{8}{\delta}\right)\right).$$

Furthermore, by using Hoeffding's result [5],

$$
\begin{aligned}
P^m(er_z(t) \geq \eta + \frac{s}{2}) \quad & \leq \quad e^{-2\left(\frac{s}{2}\right)^2 m} \\
& \leq \quad \frac{\delta}{2} \qquad (4)
\end{aligned}
$$

From (4), if

$$m \geq \frac{2}{s^2} \ln\left(\frac{2}{\delta}\right), \qquad (5)$$

which is certainly the case if $m \geq m_0$

■

[9] mentions a result without proof that is similar to Theorem 2.2 and attributes to [1] . Theorem 2.2 provides a result involving VC dimension rather than $|H|$ and is therefore more widely applicable (since finite $|H|$ implies finite dimension but not conversely). It also generalises the classic result (Theorem 2.1) of Angluin and Laird.

3. Generalization of PAC learning in the presence of noisy examples that follow a Markov chain

We assume that the learner receives a sequence of training examples

$$\underline{z} = ((x_0, y_0), (x_1, y_1), \ldots, (x_T, y_T)) = (z_0, z_1, \ldots, z_T) \in Z^T,$$

where $x_i \in X$ forms a Markovian process. Note that in order to differentiate the training sample $((x_0, y_0), (x_1, y_1), \ldots, (x_T, y_T))$ from the earlier $((x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m))$, we consider the notation $\underline{z}$. We assume that the learner has access to a corrupted oracle $EX_\eta(t)$. We assume that $EX_\eta(t)$ produces instances $x \in X$ according to some underlying Markov process, but it attaches to an example x the opposite label to $t(x)$ subject to some unknown independent probability $\eta < \frac{1}{2}$. As earlier, we assume we know η_b such that $\eta \leq \eta_b < \frac{1}{2}$.
Let π_0 be the probability distribution of the initial state x_0. As it is appropriate to consider a finite set of instances X in many applications, we assume that this is an aperiodic and irreducible finite state Markovian process that has a unique stationary distribution π. In this case, for some $\beta, \psi > 0$ the following inequality holds [8].

$$|Pr\{x_i = x^{'} | x_0 = x\} - \pi(x^{'})| \leq \beta e^{-\psi i}, \forall x, x^{'} \in X \tag{6}$$

We write $T = nl$, where n and l are integers. Let $\underline{z} = (z_0, z_1, \ldots, z_T) \in Z^T$, where $z_i = (x_i, y_i)$, be a training sample and let

$$\underline{z}^{'} = ((x_0, y_0), (x_n, y_n), (x_{2n}, y_{2n}), \ldots, (x_{nl}, y_{nl}))$$

be the length $l + 1$ sub-sample of $\underline{z}$ that is created by considering (as in [3]) every n^{th} observation of $\underline{z}$. One key idea is to consider large enough n such that $x_{nj}, j = 0, 1, \ldots, l$ are almost i.i.d. For a given function $f \in H$, the error of f with respect to stationary distribution π is $er_\pi(f)$. Here, $er_\pi(f)$ is the probability that if x is generated according π, then $f(x) \neq t(x)$. For a sample $\underline{z}$, the sample error is $er_{\underline{z}}(f)$ and for the sub-sample $\underline{z}^{'}$, it is $er_{\underline{z}^{'}}(f)$. For $\epsilon > 0$, let us denote by $A_{H,\epsilon}$ the set

$$\{\underline{z} \in Z^T : \exists\ f \in H \text{ minimising } er_{\underline{z}^{'}}(f) \text{ and having error } er_\pi(f) > \epsilon\}.$$

We shall let $1\{A_{H,\epsilon}(\underline{z})\}$ be 1 if $z \in A_{H,\epsilon}$ and 0 otherwise (and shall sometimes simply write $1\{A_{H,\epsilon}\}$). Then,

$$Pr_{\pi_0}(A_{H,\epsilon}) = \sum_{(x_0,x_n,\ldots,x_{nl})\in X^{l+1}} 1\{A_{H,\epsilon}\}\pi_0(x_0)\prod_{i=1}^{l} p^n(x_{ni}|x_{n(i-1)})$$

where $p^n(x|x^{'})$ is the n^{th} step transition probability from state $x^{'}$ to state x. Here, $Pr_{\pi_0}(A)$ denotes the probability that event A will occur when the initial state is drawn according to the distribution π_0. We have the following result, which is a slight variant of Gamarnik's result from [3] (the original result of Gamarnik [3] considers a different set than $A_{H,\epsilon}$).

Lemma 3.1.

$$Pr_{\pi_0}(A_{H,\epsilon}) \leq \sum_{(x_n,x_{2n},\ldots,x_{nl})\in X^{l}} 1\{A_{H,\epsilon}\}\prod_{i=1}^{l} \pi(x_{ni}) + l|X|\beta e^{-\psi n}.$$

Proof: *Step 1:* From 6, $p^{(n)}(x_{nl}|x_{n(l-1)}) \leq \pi(x_{nl}) + \beta e^{-\psi n}$. So, it follows that

$$\begin{aligned} Pr_{\pi_0}(A_{H,\epsilon}) \quad \leq & \sum_{(x_0,x_n,\ldots,x_{nl})\in X^{l+1}} 1\{A_{H,\epsilon}\}\pi_0(x_0)\prod_{i=1}^{l-1} p^{(n)}(x_{ni}|x_{n(i-1)})\pi(x_{nl}) \\ & + \sum_{(x_0,x_n,\ldots,x_{nl})\in X^{l+1}} 1\{A_{H,\epsilon}\}\pi_0(x_0)\prod_{i=1}^{l-1} p^{(n)}(x_{ni}|x_{n(i-1)})\beta e^{-\psi n}. \end{aligned}$$

Step 2: The second term on the right-hand side can be bounded above as follows:

$$\begin{aligned} & \sum_{(x_0,x_n,\ldots,x_{nl})\in X^{l+1}} 1\{A_{H,\epsilon}\}\pi_0(x_0)\prod_{i=1}^{l-1} p^{(n)}(x_{ni}|x_{n(i-1)})\beta e^{-\psi n} \\ & \leq \sum_{(x_0,x_n,\ldots,x_{n(l-1)})\in X^{l}} \pi_0(x_0)\prod_{i=1}^{l-1} p^{(n)}(x_{ni}|x_{n(i-1)})|X|\beta e^{-\psi n} \\ & \leq |X|\beta e^{-\psi n}, \end{aligned}$$

as $1\{A_{H,\epsilon}\} \leq 1$, and

$$\sum_{(x_0,x_n,\ldots,x_{n(l-1)})\in X^{l}} \pi_0(x_0)\prod_{i=1}^{l-1} p^{(n)}(x_{ni}|x_{n(i-1)}) = 1.$$

Therefore,

$$Pr_{\pi_0}(A_{H,\epsilon}) \leq \sum_{(x_0,x_n,\ldots,x_{nl})\in X^{l+1}} 1\{A_{H,\epsilon}\}\pi_0(x_0)\prod_{i=1}^{l-1} p^{(n)}(x_{ni}|x_{n(i-1)})\pi(x_{nl}) + |X|\beta e^{-\psi n}.$$

By repeating Step 1, we obtain

$$Pr_{\pi_0}(A_{H,\epsilon}) \leq \sum_{(x_0,x_n,\ldots,x_{nl})\in X^{l+1}} 1\{A_{H,\epsilon}\}\pi_0(x_0)\prod_{i=1}^{l-2} p^{(n)}(x_{ni}|x_{n(i-1)})\pi(x_{(l-1)n})\pi(x_{nl})$$

$$+ \sum_{(x_0,x_n,\ldots,x_{nl})\in X^{l+1}} 1\{A_{H,\epsilon}\}\pi_0(x_0)\prod_{i=1}^{l-2} p^{(n)}(x_{ni}|x_{n(i-1)})\pi(x_{(l-1)n})\beta e^{-\psi n} + |X|\beta e^{-\psi n}.$$

Repeating Step 2, the second part of the above expression can be bounded by $|X|\beta e^{-\psi n}$. So,

$$Pr_{\pi_0}(A_{H,\epsilon}) \leq \sum_{(x_0,x_n,\ldots,x_{nl})\in X^{l+1}} 1\{A_{H,\epsilon}\}\pi_0(x_0)\prod_{i=1}^{l-2} p^{(n)}(x_{ni}|x_{n(i-1)})\pi(x_{n(l-1)})\pi(x_{nl}) + 2|X|\beta e^{-\psi n}.$$

Now, by performing Step 1 and Step 2 iteratively, one can obtain

$$Pr_{\pi_0}(A_{H,\epsilon}) \leq \sum_{(x_0,x_n,\ldots,x_{nl})\in X^{l+1}} 1\{A_{H,\epsilon}\}\pi_0(x_0)\prod_{i=1}^{l}\pi(x_{ni}) + l|X|\beta e^{-\psi n}.$$

But,

$$\sum_{(x_0,x_n,\ldots,x_{nl})\in X^{l+1}} 1\{A_{H,\epsilon}\}\pi_0(x_0)\prod_{i=1}^{l}\pi(x_{ni}) \leq \sum_{(x_n,x_{2n},\ldots,x_{nl})\in X^{l}} 1\{A_{H,\epsilon}\}\prod_{i=1}^{l}\pi(x_{ni}),$$

and therefore

$$Pr_{\pi_0}(A_{H,\epsilon}) \leq \sum_{(x_n,x_{2n},\ldots,x_{nl})\in X^{l}} 1\{A_{H,\epsilon}\}\prod_{i=1}^{l}\pi(x_{ni}) + l|X|\beta e^{-\psi n}.$$

■

The upper bound on the probability of the event $A_{H,\epsilon}$ from Lemma 3.1 is useful to find the number of training examples that are required for successful learning in the presence of classification noise.

Theorem 3.1. *With $A_{H,\epsilon}$ as defined earlier, we have $Pr_{\pi_0}(A_{H,\epsilon}) < \delta$ provided*

$$l \geq \frac{2}{\epsilon^2(1-2\eta_b)^2} \ln\left(\frac{4|H|}{\delta}\right)$$

and

$$n \geq \frac{1}{\psi}\left(\ln l + \ln\left(\frac{2|X|\beta}{\delta}\right)\right).$$

Note that Theorem 3.1 implies a sample size T that satisfies

$$T \geq \frac{1}{\psi}\frac{2}{\epsilon^2(1-2\eta_b)^2} \ln\left(\frac{4|H|}{\delta}\right)\left(\ln\left(\frac{2}{\epsilon^2(1-2\eta_b)^2}\right) + \ln\ln\left(\frac{4|H|}{\delta}\right) + \ln\left(\frac{2|X|\beta}{\delta}\right)\right).$$

Proof: We can use Lemma 3.1. Consider the quantity

$$\sum_{(x_n, x_{2n}, \ldots, x_{nl}) \in X^l} 1\{A_{H,\epsilon}\} \prod_{i=1}^{l} \pi(x_{ni}).$$

This is equal to the probability that when $x_{ni} \quad (i = 1, \ldots, l)$ are drawn independently at random, each according to distribution π, and each is then labelled noisily, as described, event $A_{H,\epsilon}$ holds. We can therefore use Angluin and Laird's result [1] (which is discussed in Theorem 2.1) about PAC learning in the presence of noisy examples and assure that

$$\sum_{(x_n, x_{2n}, \ldots, x_{nl}) \in X^l} 1\{A_{H,\epsilon}\} \prod_{i=1}^{l} \pi(x_{ni}) < \frac{\delta}{2}$$

provided

$$l > \frac{2}{\epsilon^2(1-2\eta_b)^2} \ln\left(\frac{4|H|}{\delta}\right).$$

Furthermore, we can assure

$$l|X|\beta e^{-\psi n} \leq \frac{\delta}{2}$$

if

$$n > \frac{1}{\psi}\left((\ln l + \ln(|X|\beta)) + \ln\left(\frac{2}{\delta}\right)\right).$$

It follows that if

$$l \geq l_0 = \frac{2}{\epsilon^2(1-2\eta_b)^2} \ln\left(\frac{4|H|}{\delta}\right)$$

and

$$n \geq \frac{1}{\psi}(\ln l + \ln(|X|\beta) + \ln\left(\frac{2}{\delta}\right)),$$

then

$$Pr_{\pi_0}(A_{H,\epsilon}) < \delta.$$

■

Theorem 3.1 suggests the upper bound (nl) for the number of training samples that are required for successful learning based on the extended PAC framework to finite Markov chains in the presence of classification noise. Note that we can use Theorem 2.2 and extend Theorem 3.1 to involve the VC dimension.

4. A PAC framework for Real-valued Functions

We now discuss extensions of the PAC learning framework for Boolean functions to a framework for learning real valued functions. Seminal work on this was conducted by Haussler [4] and has been much extended since [7, 2]. We now assume the learner computes real-valued functions. Let the hypothesis set H be the collection of all real valued functions on the instance space X that can be computed by the learner L. In this more general model, we assume that the learner receives training samples (x_i, y_i), each of which is drawn according to some fixed, but unknown, distribution P on $X \times [0,1]$ (note that the label $y_i \in [0,1]$ is a real number). In the case of real valued functions, it may not be appropriate to define

$$er_P(f) = P\{(x,y) \in Z : f(x) \neq y\},$$

as with high probability we end up with $er_P(f) = 1$. It is more appropriate to quantify $er_P(f)$ as on average how far $f(x)$ is away from $t(x)$. A loss function can be considered as penalizing the mistakes that the hypothesis f makes. The most common and convenient loss function is quadratic loss [2]

$$er_P(f) = E[f(x) - y]^2,$$

and its estimate using the sample sequence $z = (z_1, z_2, \ldots, z_m)$ is the "empirical loss"

$$er_z(f) = \frac{1}{m}\sum_{i=1}^{m}(f(x_i) - y_i)^2.$$

If we did not restrict the magnitude of y_i then the quadratic loss could be arbitrarily large. In order to eliminate this problem, we have assumed that y falls in a bounded interval $[0, 1]$. We restrict ourselves to agnostic learning (where nothing is known about hypothesis set H) [7]. Here, after seeing enough labeled examples, it is hoped that with high probability, the learner L ends up with a hypothesis h such that $er_P(h)$ is close to the quantity

$$opt_P(H) = \inf_{g \in H} \; er_P(g),$$

which is a measure of the best possible error achievable using H.

Definition 4.1. *L is a PAC learning algorithm using the class H if there exists for each $\epsilon, \delta \in (0, 1)$, an integer $m_0(\epsilon, \delta)$ such that, for all samples of size at least $m_0(\epsilon, \delta)$, the following holds: with probability at least $1 - \delta$, for all distributions P on $X \times [0, 1]$, we have*

$$er_P(h) \leq opt_P(H) + \epsilon,$$

where h is the hypothesis produced by L.

In analysing real-valued learning, covering numbers have proven useful [4]. We say $W \subset \Re^m$ is an ϵ-cover of H (with respect to the norm $\|.\|_1$) on a sequence $x = (x_1, x_2, \ldots, x_m) \in X^m$ if for all $f \in H$ there exist a $u \in W$ such that

$$\frac{1}{m}\sum_{i=1}^{m} |u_i - f(x_i)| \leq \epsilon.$$

We define the covering number $N(\epsilon, H, m)$ as

$$N(\epsilon, H, m) = \max_{X \in X^m} \min \{|U| : U \text{ is an } \epsilon\text{−cover of } H \text{ on } (x_1, x_2, \ldots, x_m)\}.$$

The following result is a key ingredient in extending the PAC framework: see [2], for example.

Theorem 4.1. *Let H be a set of hypothesis defined on instance space X and mapping into the interval [0,1]. Let P be any probability distribution over $Z = X \times [0,1]$, ϵ any real number between 0 and 1, and m any positive integer. Then*

$$P^m\{some\ \ f\ \ in\ \ H\ \ has\ \ |er_P(f) - er_z(f)| \geq \epsilon\} \leq N\left(\frac{\epsilon}{16}, H, 2m\right) \exp\left(\frac{-\epsilon^2 m}{32}\right).$$

For the PAC learning model of real-valued function learning, it is considered that the learner receives a sequence of ordered pairs of the form $(x, y) \in Z$ as training data. Each $(x \in X)$, labelled with the value $y \in \Re$ according to some unknown target function that outputs y (we can assume $y \in [0,1]$ by scaling, if necessary). We assume that the input process, the sequence of training data provided to the learner, is a finite Markovian process x_i, $i = 0, 1, \ldots, T$. Let π_0 be the probability distribution of the initial state x_0. As it is reasonable to consider that patterns and cycles occur in the real world data, we assume that every $x_i \in X$ is repeated at irregular times and is reachable from the other elements of X. That means we assume that x_i is an aperiodic and irreducible, finite state Markovian process. It has a unique stationary distribution π, and for some $\beta, \psi > 0$ the following inequality holds:

$$|Pr\{x_i = x^{'} | x_0 = x\} - \pi(x^{'})| \leq \beta e^{-\psi i}, \forall x, x^{'} \in X.$$

Let n, l be positive integers and define $T = nl$. Let $\underline{z} = (z_0, z_1, \ldots, z_T) \in Z^T$ be a training sample and $\underline{z}^{'} = ((x_0, y_0), (x_n, y_n), (x_{2n}, y_{2n}), \ldots, (x_{nl}, y_{nl}))$ be the length $l + 1$ subsample of $\underline{z}$ that is created by considering every n^{th} observation of $\underline{z}$. As before, we consider n significantly large, so that $x_{nj}, j = 0, 1, \ldots, l$ can be considered as almost i.i.d. Let H be the finite set of all real-valued functions the learner can compute. For a given function $f \in H$, the error of f with respect to the stationary distribution π is defined as the expected value of $(f(x) - y)^2$, where the expectation is calculated with respect to $z = (x, y)$ drawn according to π. So, the error can be expressed as $er_\pi(f) = E[f(x) - y]^2$. For sample $\underline{z}$ and subsample $\underline{z}'$, we define sample and subsample errors of a function $f \in H$ as follows:

$$er_{\underline{z}}(f) = \frac{1}{T} \sum_{i=1}^{T} (y_i - f(x_i))^2$$

and

$$er_{\underline{z}'}(f) = \frac{1}{l}\sum_{j=1}^{l}(y_{nj} - f(x_{nj}))^2.$$

Theorem 4.2. *With the notation as defined above, if*

$$l \geq \frac{32}{\epsilon^2}\ln\left(\frac{|H|}{\delta}\right)$$

and

$$n \geq \frac{1}{\psi}\left(\ln l + \ln\left(\frac{2|X|\beta}{\delta}\right)\right),$$

then with probability at least $1-\delta$*, for all* $f \in H$*,* $er_{\pi}(f) \leq er_{\underline{z}'}(f) + \epsilon$*.*

Proof:
We now denote the event $er_{\pi}(f) > \ er_{\underline{z}'}(f) + \epsilon$ by $A_{H,\epsilon}$. We have

$$Pr_{\pi_0}(A_{H,\epsilon}) = \sum_{(x_0,x_n,\ldots,x_{nl})\in X^{l+1}} 1\{A_{H,\epsilon}\}\pi_0(x_0)\prod_{i=1}^{l} p^n(x_{ni}|x_{n(i-1)}),$$

where $p^n(x_i|x_{i-1})$ is the n^{th} step transition probability from state x_{i-1} to state x_i.
From the proof of lemma 3.1, it is clear that

$$Pr_{\pi_0}(A_{H,\epsilon}) \leq \sum_{(x_n,x_{2n},\ldots,x_{nl})\in X^{l}} \pi_0(x_0)\prod_{i=1}^{l}\pi(x_{in}) + l|X|\beta e^{-\psi n}.$$

Now, arguing as earlier,

$$\sum_{(x_n,x_{2n},\ldots,x_{nl})\in X^{l}} 1\{A_{H,\epsilon}\}\prod_{i=1}^{l}\pi(x_{ni})$$

can be bounded using the fact that it is the probability of the event $A_{H,\epsilon}$ when random variables x_{ni} are drawn *independent and identically distributed* according to π. We can therefore use the PAC result for real valued functions, Theorem 4.2, to obtain

$$\sum_{(x_n,x_{2n},\ldots,x_{nl})\in X^l} 1\{A_{H,\epsilon}\}\prod_{i=1}^{l}\pi(x_{ni}) < 4Ne^{\frac{-\epsilon^2 l}{32}}$$

where $N \leq |H|$ is a covering number $N\left(\frac{\epsilon}{16}, H, 2m\right)$.
We can therefore assure

$$\sum_{(x_n,x_{2n},\ldots,x_{nl})\in X^l} 1\{A_{H,\epsilon}\}\prod_{i=1}^{l}\pi(x_{ni}) < \frac{\delta}{2}$$

provided

$$l > \frac{32}{\epsilon^2}\ln\left(\frac{8|H|}{\delta}\right).$$

Furthermore, we can assure

$$l|X|\beta e^{-\psi n} \leq \frac{\delta}{2}$$

if

$$n > \frac{1}{\psi}(\ln l + \ln(|X|\beta)) + \ln\left(\frac{2}{\delta}\right).$$

The theorem follows. ∎
Note that this implies

$$\begin{aligned} T &\geq nl \\ &= \frac{32}{\epsilon^2}\ln\left(\frac{8|H|}{\delta}\right)\frac{1}{\psi}\left(\ln\left(\frac{32}{\psi^2}\right) + \ln\ln\left(\frac{8|H|}{\delta}\right) + \ln\left(\frac{2|X|\beta}{\delta}\right)\right). \end{aligned}$$

Theorem 4.2 suggests that a learning algorithm minimizing the sub-sample error $er_{\underline{z}'}(f)$ will be successful. The following corollary bounds the error in terns of the empirical error on the whole sample.

Corollary 4.1. *With the notation as above, suppose*

$$T \geq \frac{1}{\psi}\frac{32}{\epsilon^2}\ln\left(\frac{8|H|}{\delta}\right)\left(\ln\left(\frac{32}{\epsilon^2}\right) + \ln\ln\left(\frac{8|H|}{\delta}\right) + \ln\left(\frac{2|X|\beta}{\delta}\right)\right).$$

Then, with probability at least $1-\delta$*, for any distribution* P *on* $X \times [0,1]$*, for all* $f \in H$*,*

$$er_\pi(f) \leq C(\epsilon,\delta) er_{\underline{z}}(f) + \epsilon$$

where

$$C(\epsilon,\delta) = \frac{1}{\psi}\left(\ln\left(\frac{32}{\epsilon^2}\right) + \ln\ln\left(\frac{8|H|}{\delta}\right) + \ln\left(\frac{2|X|\beta}{\delta}\right)\right).$$

Proof: Let

$$l = \frac{32}{\epsilon^2}\ln\left(\frac{8|H|}{\delta}\right)$$

and let

$$n \geq \frac{1}{\psi}\left(\ln l + \ln\left(\frac{2|X|\beta}{\delta}\right)\right).$$

We have

$$\begin{aligned} er_{\underline{z}}(f) &= \frac{1}{T}\sum_{i=1}^{T}(y_i - f(x_i))^2 \\ &= \frac{1}{nl}\sum_{i=1}^{nl}(y_i - f(x_i))^2. \end{aligned}$$

Also,

$$er_{\underline{z}'}(f) = \frac{1}{l}\sum_{j=1}^{l}(y_{nj} - f(x_{nj}))^2.$$

Then, $nl\ er_{\underline{z}}(f) \geq l\ er_{\underline{z}'}(f)$ and so $\mathrm{er}_{\underline{z}'}(f) \leq n\ er_{\underline{z}}(f)$. Therefore,

$$Pr_{\pi_0}(er(f) > n\ er_{\underline{z}}(f) + \epsilon) \leq Pr_{\pi_0}(er(f) > er_{\underline{z}'}(f) + \epsilon),$$

and the Corollary follows from Theorem 4.2. ∎

Corollary 4.1 bounds the error in terms of the error on the whole sample.

5. Conclusions

In this paper, we have extended in two distinct ways the basic PAC model and previous work on extending it. First, we developed a model of learning in which the training data is noisy and is generated by a Markov process. Then, we considered the case in which the data is Markovian, but in which the data labels are real-valued. The resulting sample size bounds may be quite large (as is often the case in PAC-type bounds) but they provide some justification, theoretically, for learning in these contexts by choosing a hypothesis that fits the data well.

Acknowledgement

The author would like to thank Martin Anthony for his constructive suggestions on learning real valued functions.

References

[1] D. Angluin and P. Laird. Learning from noisy examples. *Machine Learning*, 2:343–370, 1988.

[2] M. Anthony and P. L. Bartlett. *Neural Network Learning: Theoretical Foundations*. Cambridge University Press, 1999.

[3] D. Gamarnik. Extension of the PAC framework to finite and countable markov chains. *IEEE Trans. Inform. Theory*, 49(1):338–345, 2003.

[4] D. Haussler. Decision theoretic generalizations of the PAC model for neural net and other learning applications. *Information and Communication*, 100(1):78–150, 1992.

[5] W. Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of American Statistical Association*, 58(301):13–30, 1963.

[6] M. Kearns. Learning in the presence of malicious errors. *SIAM Journal on Computing*, 22:807–837, 1993.

[7] M. J. Kearns and R. E. Schapire. Toward efficient agnostic learning. *Machine Learning*, 17:115–141, 1994.

[8] J. R. Norris. *Markov Chains*. Cambridge University Press, Cambridge, UK, 1997.

[9] H. U. Simon. PAC-learning in the presence of one-sided classification noise. International Symposium on Artificial Intelligence and Mathematics, 2012.

[10] R. H. Sloan. Four types of noise in data for PAC learning. *Information Processing Letters*, 54:157–162, 1995.

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

[12] V. N. Vapnik and A. Y. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and its Applications*, 16(2):264–280, 1971.