Some properties of infinite VC-dimension systems

Professor Alexey Ya. Chervonenkis

Technical Report

CLRC-TR-01-04

March 1999

Royal Holloway University of London

Department of Computer Science
Egham, Surrey TW20 0EX, England

Preliminary

Let X be a set; R, a sigma-algebra of measurable subsets; P, probability distribution defined for all elements of R; S, some system of measureable sets $A \subseteq X$.

Let $X_l = \{x_1, x_2, \ldots, x_l\}$ be an independent sequence, taken from X with distribution P. Two sets A_1 and A_2 are undifferentiable on the sequence, iff for all x_i $(i = 1, l)$

$$((x_i \in A_1) \sim (x_i \in A_2)).$$

Now we define $\Delta^s(X_l)$ as the number of classes of sets $A \in S$, undifferentiable on X_l. If $\max(\log \Delta^s(X_l)) \equiv l$, we say that the system S has infinite VC dimension.

If $\Delta^S(X_l)$ is measurable, then we define entropy

$$H(S, l) = \overset{E}{_{Xl}} \log \Delta^s(X_l)$$

and entropy per element as $H(S, l)/l$

It is shown [1,2] that entropy per element has the limit while l goes to infinity

$$H(S, l)/l \to c(S, P), \quad (0 \leq c \leq 1)$$

$$l \to \infty,$$

and $\Delta^S(X_l)$ goes also to c by probability (due to results of this work one can see that it goes to c also with probability 1).

In this work we consider the case $c < 0$.

Main Assertion

Let us define for a measurable set B and a sequence X_l:

$X_{l1}(X_l, B)$ — the subsequence of X_l, of elements which belong to B, $l_1(X_l, B)$ — the length of this subsequence,

$$\Delta^S(X_l, B) \overset{def}{=} \Delta^S(X_{l1}$$

(in the case when $l_1 = 0$, we define $\Delta^S(X_l, B) = 1$).

If $\lim\limits_{l \to \infty} H(S, l) = c > 0$, then
there exists a subset $T \subseteq X$, such that for any $l > 0$ with probability 1
$$log\ \Delta^S(X_l, T) = l_1(X_l, T),$$
and $P(T) = c$.

(It is evident that the probability measure of such T can not be more than c).

Some additional definitions.

For a set system S_0, sequence $X_l(x_i \in X,\ i = 1, l)$ and binary sequence

$$Y_l = \{y_1, \ldots, y_l\},\ y_i = 0 \text{ or } 1$$

we denote

$$S(S_0, X_l, Y_l) = \{A:\ A \subseteq S, \text{ and for all } i\ (x_i \in A) \sim (y_i = 1)\}.$$

In other words $S(S_0, X_l, Y_l)$ consists of all sets A from S_0 such that for all i x_i belongs to the set A iff $y_i = 1$.

If $l = 1$

- $S(S_0, x, 1)$ consists of all sets from S_0, which include x,
- $S(S_0, x, 0)$ consists of all sets from S_0, which do not include x.

Consider

$$\begin{aligned} c(S_0) &= \lim\ H(S_0, l)/l \\ c_0(S_0, x) &= \lim\ H(S(S_0, x, 0), l)/l \quad (l \to \infty) \\ c_1(S_0, x) &= \lim\ H(S(S_0, x, 1), l)/l \end{aligned}$$

While $S_0 \supseteq S(S_0, x, 1)$, $S_0 \supseteq S(S_0, x, 0)$ and $S = S((S_0, x, 1) \cup S(S_0, x, 0)$, we have

$$\begin{aligned} &\Delta^{S_0}(X_l) \geq \Delta^S(S_0, x, 1), \\ &\Delta^{S_0}(S_0, X_l) \geq \Delta^S(S_0, x, 0), \\ &\Delta^{S_0}(S_0, X_l) \leq \Delta^S(S_0, x, 0) + \Delta^S(S_0, x, 1). \end{aligned}$$

It implies that

$$c(S_0) = \max\{c_0(S_0, x), c_1(S_0, x)\}.$$

In other words, either $c = c_0$ or $c = c_1$ and in both cases $c$is not less than each of c_0, c_1. But possibly for some x it appears that

$$c(S_0) = c_0(S_0, x) = c_1(S_0, x). \tag{1}$$

We shall call such x points of equal branching. The set of such points we shall call the equal branching set EB

$$EB(S_0) = \{x:\ x \in X \text{ and } c(S_0) = c_0(S_0, x) = c_1(S_0, x)\}.$$

It may seem that this set is possibly empty, but we shall show that this set is not only *not* empty, but that its measure

$$P(EB(S_0)) \geq c \tag{2}$$

Restriction process

The proof of this fact will be given below, but now taking (2) as true, we construct a process to find the set T, declared in the Main Assertion.

Given a sequence of independent samples $x_1, x_2, \ldots, x_i, \ldots$ (each distributed according to $P(x)$), we define the process of reducing the system S, according to the following recursive rule

$$S_0 = S; \quad B_0 = EB(S_0);$$

If x_i does not belong to B_{i-1}, then

$$S_i = S(S_{i-1}, x, 1), \text{ if } c_1(S_{i-1}, x_i) = c(S_{i-1});$$
$$S_i = S(S_{i-1}, x, 0), \text{ if } c_0(S_{i-1}, x_i) = c(S_{i-1});$$

If x_i does belong to B_{i-1}, then

$$S_i = S(S_{i-1}, x, 1) \text{ with probability } 0.5 \text{ or}$$
$$S_i = S(S_{i-1}, x, 0) \text{ with probability } 0.5.$$

In any case $B_i = EB(S_i)$.

According to the rule for all i

$$c(S_i) = c(S_0) = c(S);$$

and due to (2)

$$P(B_i) \geq c(S).$$

It is evident that $S_i \subseteq S_{i-1}$, and it is almost evident that

$$B_i \subseteq B_{i-1}.$$

We shall show with probability 1

$$P(B_i) \to c(S) \text{ with } i \text{ going to infinity.}$$

Now the set T, declared in the Main Assertion, may be found as intersection of all B_i, $(i = 1, 2, \ldots)$.

$$T = \bigcap_{i=1}^{\infty} B_i,$$

In general the set T may be not unique and the random process of its searching may lead to different resulting sets.

The proof

Lemma 1.

$$P(EB(S_0)) \geq c(S_0).$$

For a fixed sequence $X_l = \{x_1, \ldots, x_l\}$ we call a binary sequence $Y_l = \{y_1, \ldots, y_l\}$ ($y_i = 0$ or $y_i = 1$ permitted iff

$$S(S_0, X_l, Y_l) \text{ is not empty.}$$

The number of different permitted Y_l is just $\Delta^{S_0}(X_l)$.

Now we define a binary random process of length l in such a way that:

1. all permitted sequences Y_l are equally probable $(p(Y_l) = 1/\Delta^{s_0}(X_l))$,

2. all not permitted sequences Y_l are improbable $(p(Y_l) = 0)$.

The Shannon's entropy of the process will be

$$\bar{H}\ (l) = -\Sigma p(Y_l) \log\ p(Y_l) = log\ \Delta^{S_0}(X_l); \quad (p(Y_l) > 0)$$

Now conditional entropy will be

$$\begin{aligned} &\bar{H}\bar{H}\ (y_i/y_1, \ldots, y_{i-1}) = \\ &= -p(y_i = 0/y_1, \ldots, y_{i-1}) \log\ (y_i = 0/y_1, \ldots, y_{i-1}) - p(y_i = 0/y_1, \ldots, y_{i-1}) log\ (y_i = 0/y_1, \ldots, y_{i-1}). \end{aligned}$$

We shall denote, averaged over all $y_1, \ldots y_{i-1}$, conditional Shannon's entropy as

$$\bar{H}\ (y_i/y_1, \ldots, y_{i-1}) = \underset{Y_{i-1}}{E}\ \bar{H}\bar{H}\ (y_i/y_1, \ldots, y_{i-1})$$

Now total Shannon's entropy may be presented in the form

$$\bar{H}\ (l) = \bar{H}\ (y_1) + \bar{H}\ (y_2/y_1) + \ldots + \bar{H}\ (y_l/y_1, \ldots, y_{l-1}).$$

In general the elements of the sum are not monotonoious, but if we average this equation over all possible X_l (assuming that X_l is an independent sequence of samples, each distributed according to $P(x)$), the become monotoniously decreasing. Really, due to the fact that Shannon's information is not negative, we have

$$\bar{H}\ (y_i/y_1, \ldots, y_{i-1}) \leq \bar{H}\ (y_i/y_1, \ldots, y_{i-2}).$$

Averaging we have

$$\underset{X_l}{E}\ \bar{H}\ (y_i/y_1, \ldots, y_{i-1}) \leq \underset{X_l}{E}\ \bar{H}\ ((y_i/y_1, \ldots, y_{i-2})$$

But averaged entropy depends only on the number of predecessors, and not on their order.

So:

$$E\ \bar{H}\ (y_i/y_1,\ldots,y_{i-1}) \le E\ \bar{H}\ ((y_i/y_1,\ldots,y_{i-2}) \text{ and}$$

$$H(S_0,l) = E\ \bar{H}\ (l) = E\ \bar{H}\ (y_1) + E\ \bar{H}\ (y_2/y_1) + \ldots + E\ \ \bar{H}\ (y_l/y_1,\ldots,y_{l-1})$$

Then

$$\underset{X_l}{EH(y_1)} \ge H(S_0,l)/l.$$

And further

$$\underset{l\to\infty}{\underline{\lim}}\ EH(y_1) \ge c(S_0), \tag{3}$$

Nox if x_1 is not an equal branching point, then with probability 1 (on $x_2,\ldots,x_l$)

$$p(y_1=0)\to 1 \text{ or } p(y_1=0)\to 0 \text{ with } l\to\infty$$

and thus $\bar{H}\ (y_1)\to 0$ and

$$\underset{x_2\ldots x_l}{E}\ \ \bar{H}\ (y_1)\to 0.$$

Then, taking into account that $H(y_1)\le 1$,

$$E\ \bar{H}\ (y_1) = \underset{x_1\subseteq EB(S_0)}{E}\ \ \underset{x_2\ldots,x_l}{E}\ \bar{H}\ (y_1) + \underset{X_l \text{ not } \subseteq EB(S_0)}{E}\quad \underset{x_2\ldots,x_l}{E}\ \ \bar{H}\ (y_1)\ \ \ge\ \ P(EB(S_0)) + o(l).$$

Comparing this inequality with (3) we have

$$P(EB(S_0)) \ge c(S_0)$$

The Lemma 1 is proved.

Now we return to the Restriction Process, defined above.

Lemma 2.

$\underset{i\to\infty}{P(B_i)\to c(S)}$ with probability 1.

($P(B_i)$ in general is a random value, as far as sets B_i are random, being determined by preceding xs and ys.)

We know that $P(B_i)\ge c(S_0)$, $P(B_i)\le 1$ and it monotoniously decreases with i.

That is why it is enough to show that

$$EP(B_i) \to c(S) \quad \text{by } i \to \infty$$

Again, let us consider a fixed sequence $x-!, x_2, \dots, x_l$ and declare a binary sequence $Y_l = y_1, \dots, y_l$ permitted 1 iff

$$c(S(S_0, X_l, Y_l)) = c(S_0) \tag{4}$$

It can easily be seen that all sequences Y_l possibly generated by the Restriction Process are permitted 1, all permitted 1 Y_l are permitted, but in general not all permitted Y_l are permitted 1.

Now we define a binary random process in the following recursive way

If $y_1, \dots, y_l,\ 1$ is permitted 1 and $y_1, \dots, y_l,\ 0$ is not permitted 1 then
$p(y_i = 1/y_1, \dots, y_{i-1}) = 1$ and
$p(y_i = 0/y_1, \dots, y_{i-1}) = 0;$

If $y_1, \dots, y_{i-1},\ 0$ is permitted 1 and $y_1, \dots, y_{i-1},\ 1$ is not permitted 1 then
$p(y_i = 1/y_1, \dots, y_{i-1}) = 0$ and
$p(y_i = 0/y_1, \dots, y_{i-1}) = 1;$

If $y_1, \dots, y_{i-1},\ 1$ is permitted 1 and $y_1, \dots, y_{i-1},\ 0$ is also permitted 1 then
$p(y_i = 1/y_1, \dots, y_{i-1}) = 0.5$ and
$p(y_i = 0/y_1, \dots, y_{i-1}) = 0.5;$

Again we can calculate the Shannon's entropy of the binary process.

But now

$$\bar{H}\ (Y_l) \le log\Delta^s(X_l)$$

as far as the number of different permitted 1 sequences is not greater than $\Delta^s(X_l)$ and in general they are not equally probable.

Then

$$\bar{H}\bar{H}\ (y_i/y_1, \dots, y_{i-1}) = 0,$$

if $x - i$ does not belong to $EB(S(S_0, X_{i-1}, Y_{i-1})) = B_{i-1}$
and

$$\bar{H}\bar{H}\ (y_i/y_1, \dots, y_{i-1}) = 1, \text{ if}$$

x_i does belong to $EB(S(S_0, X_{i-1}, Y_{i-1})) = B_{i-1}$.

Averaging these relations over all X_l (as independent sample sequences) and using $P(EB_i) \ge c(S)$ we have:

$$H(S,l) \ge E\ \bar{H}\ (Y_l) = E\Sigma\ \bar{H}\bar{H}\ (y_i/y_1, \dots, y_{i-1}) = \sum_{i=1}^{l} EP(B_i) \ge lc(S)$$

Now, while $P(B_i) \ge P(B_{i+1})$, and $H(S,l)/l \to c(S)$ by $l \to \infty$,

$$\lim EP(B_i) = c(S),\ i \to \infty$$

Lemma 2 is proved.

Lemma 3.

Let

$$B_1 \supseteq B_2 \supseteq \ldots \supseteq B_i \ldots$$

be a sequence of measurable subsets of X,

$$Q = \text{ intersection of all } B_i : \; Q = \bigcap_{i=1}^{\infty} B_i, \quad P(Q)\, 0.$$

If for some $l\, 0$

$$E \log \Delta(X_l, B_i) \to lP(Q) \text{ by } i \to \infty,$$

then for this l

$$P(\log \Delta(X_l, Q) = l_1(X_1, Q)) = 1$$

We can change the assumption to the equivalent one:

$$P(\log \Delta(X_l, B_i) = l_1(X_l, B_i)) \to 1 \text{ with } i \to \infty$$

(It is true because $l_1 \geq \log \Delta(X_l, B_i), El_1(X_l, B_i) = lP(B_i) \to lP(Q)$, and $1 \leq \Delta(X_l, B_i) \leq 2^l$ has a fixed, finite set of possible values.)

Now we define the following events:

$$\begin{array}{ll} R1: & X_{l1}(X_l, B_i) \neq X_{l1}(X_l, Q) \\ R2: & \log \Delta(X_l, B_i) = l_1(X_l, B_i) \\ R3: & \log \Delta(X_l, Q) = l_1(X_l, Q) \\ R4: & (l_1(X_l, B_i) = X_1(X_l, Q) \text{and } \log \Delta(X_l, Q) = l_1(X_l, Q)) \end{array}$$

Then

$$\begin{array}{l} P(R_2) \leq P(R_1) + P(R_4); \; P(R_3) \geq P(R_4) \\ P(R_3) \to 0 \text{ by } i \to \infty; \\ P(R_2) \to 1 \text{ by } i \to \infty \text{ (modified assumption)}; \\ P(R_3) \leq 1 \text{ does not depend on } i. \end{array}$$

That is why $P(R_3) = 1$.

Lemma is proved.

Now to prove the main statement it is enough to show that the sequence B_i with probability 1 satisfies the assumption of Lemma 3.

Lemma 4.

For any positive l, with probability 1,

$$\begin{array}{ll} E \log \Delta(B_i, X_l) \to lc(S) \\ X_l \qquad\qquad i \to \infty \end{array}$$

(remembering that sets B_i, determined by the Restriction PRocess, are random).

Returning to the construction of Lemma 2, starting fro $S = S_i$. Then

$$\underset{X_l}{E} \bar{H}\ (Y_l) \geq lc(S_i) = lc(S);$$

Now, the number of permitted 1 sequences $y_1, \ldots, y_l$ (starting from S_i and B_i) is not greater than $\Delta(X_l, B_i$. (It can be easily shown by induction).

So

$$l_1(X_l, B_i) \geq \log \Delta(X_l, B_i) \geq \bar{H}\ (Y_l)$$

and then

$$lP(B_i) = \underset{X_l}{E} l_1(X_l, B_i) \geq E \log \Delta(X_l, B_i) \geq lc(S)$$

According to lemma 2, $P(B_i) \to c(S)$ with probability 1, and so we obtain the Lemma statement.

Combining the results of lemmas 3 and 4, we get the proof of the main statement.

Literature

- Vapnik, V.N. and Chervonenkis, A.Ya (1971), "On the uniform convergence of relative frequencies of events and their probabilities", Theory Probab. and Appl., Vol 16, 2, pp. 264-280.

- Vapnik, V.N. and Chervonenkis, A.Ya (1981), "Necessary and sufficient conditions for the uniform convergence of means to their expectations", Theory Probab. and Appl., Vol26, 3, pp. 532-553.