Examples of the VC Dimension

prof. dr Arno Siebes

Algorithmic Data Analysis Group
Department of Information and Computing Sciences
Universiteit Utrecht

Recall: VC dimension

The previous time we introduced the VC dimension of a hypothesis class $\mathcal{H}$ as:

The VC dimension of a set of hypotheses $\mathcal{H}$ is the size of the largest set $C \subseteq X$ such that C is shattered by $\mathcal{H}$. If $\mathcal{H}$ can shatter arbitrary sized sets, its VC dimension is infinite.

Where a finite set is shattered by $\mathcal{H}$ if

$$|\mathcal{H}_C| = 2^{|C|}$$

We now study the VD dimension of some finite classes, more in particular: classes of boolean functions.

Finite Hypothesis Classes

If a finite hypothesis class $\mathcal{H}$ shatters a finite class C then

$$|\mathcal{H}| \geq |\mathcal{H}_C| = 2^{|C|}$$

This immediately implies that

$$VC(\mathcal{H}) \leq \log(|\mathcal{H}|)$$

Clearly, the VC dimension can be smaller

- consider threshold functions that can take thresholds in $\{1, \ldots k\}$
- $|\mathcal{H}| = k$, while $VC(\mathcal{H}) = 1$

In other words,

- the difference between $VC(\mathcal{H})$ and $\log(|\mathcal{H}|)$ can be arbitrary big
- but $\log(|\mathcal{H}|)$ is never the smallest

Monotone Monomials

Recall the class C_n of boolean expressions over n literals. A smaller class C_n^+ (sometimes denoted by M_n^+) consists of the monotone (positive) monomials

- *no negations*, just conjunctions of the variables

Clearly, a variable is either in such an expression or not. Hence,

$$|C_n^+| = 2^n$$

Hence, by the previous page:

$$VC\left(C_n^+\right) \leq \log\left(2^n\right) = n$$

But, as we noted on the previous page, it could be smaller, a lot smaller.

- however, it isn't.

To prove that we are going to create a set of n elements that is shattered by C_n^+.

$VC\left(C_n^+\right) = n$

Let S consist of all 0/1-vectors of length n that have exactly

- $n-1$ 1's
- and 1 0.

Denote by x_i that element of S that has 0 for the i-th coordinate.

- if $j = i : \pi_j(x_i) = 0$
- if $j \neq i : \pi_j(x_i) = 1$

Let $R \subseteq S$ be any subset of S. Define $h_R \in C_n^+$ as

- the conjunction of all variables u_j such that $x_j \notin R$

Then we have:

$$h_R(x) = \begin{cases} 1 & \text{if } x \in R \\ 0 & \text{if } x \in S \setminus R \end{cases}$$

That is, we have a classifier for any $R \subseteq S$: S is shattered. Hence,

$$VC\left(C_n^+\right) = n$$

How About C_n?

It is easy to see that

- $VC(C_1) = 2$

the monomials

- x and $\neg x$ will do that for you.

Moreover, since $C_n^+ \subset C_n : VC(C_n^+) \leq VC(C_n)$

- any set that can be shattered by C_n^+ can be shattered by C_n

So, it may appear that by allowing negations we increase the VC dimension, because we now have that

$$n \leq VC(C_n) \leq \log(|C_n|) = \log(3^n) = n\log(3)$$

But, we don't

- except for the case $n = 1$

No set of size $n+1$ can be shattered by C_n if $n \geq 2$

$VC(C_n) = n$

Let $S = \{s^1, \ldots, s^{n+1}\}$ be a set of $n+1$, 0/1 vectors of length n, that is shattered by C_n

- define $S_i = S \setminus \{s^i\}$

Because S is shattered by C_n there exists a $m_i \in C_n$ such that

- $S_i = S \cap m_i$, thus, $\forall i, j : m_i(s^j) = 0 \leftrightarrow i = j$ ($0 =$ false)

But this means that:

- each s^i contains a component $s^i_{h(i)}$
- each m_i contains a literal $l_{k(i)}$
- such that $l_{k(i)}$ is false on $s^i_{h(i)}$, i.e., $l_{k(i)}(s^i_{h(i)}) = 0$

Given that there are only n variables

- at least 2 of these literals $l_{k(1)}, \ldots, l_{k(i+1)}$
- must refer to the same variable, say $l_{k(1)}$ and $l_{k(2)}$
- either $l_{k(1)} = l_{k(2)}$, then $l_{k(1)}(s^1_{h(1)}) = l_{k(1)}(s^2_{h(2)}) = 0$, i.e, $m_1(s^1) = m_1(s^2) = 0$. Contradiction
- or $l_{k(1)} = \neg l_{k(2)}$, then either $l_{k(1)}$ or $l_{k(2)}$ is false on s^3. Either $m_1(s^3) = 0$ or $m_2(s^3) = 0$. Again a contradiction

$D_n^{(+)}$ by Duality

Denote by

- D_n^+ the set of all disjunctions over at most n variables, again no negations
- D_n the set of disjunctions over at most n literals

Note that for $\phi \in C_n$ and $x \in \{0,1\}^n$ we have

$$\phi(x) \leftrightarrow \neg\phi(\neg x)$$

That is we have a duality between C_n and D_n and similarly between C_n^+ and D_n^+

By this duality we immediately have:

- $VC(D_n) = n$ and
- $VC(D_n^+) = n$

In the end, it is just consistently switching

- 1's to 0's and vice versa

Monotone Formulas

We have seen that both

- C_n^+, conjunctions of variables, has VD dimension n
- and D_n^+, disjunctions of variables, has VD dimension n

The natural follow up question is

- what happens if we allow both conjunctions and disjunctions
- but no negations

This is the class of *monotone boolean formulas*,

- sometimes denoted by M_n
- note, without a $+$; perhaps because allowing negations as well yields the class of all boolean functions
  - which we will discuss later

The problem is thus: determine $VC(M_n)$

Sperner's Theorem

To compute the VC dimension of M_n we need a result from combinatorics known as Sperner's Theorem.

Let X be a set of n elements

- a chain of subsets of X is a family of subsets A_i such that $\emptyset \subseteq A_1 \subset A_2 \subset \cdots \subset A_k \subseteq X$
- an antichain is a family of subsets F such that for any two elements $A, B \in F$:

$$A \not\subset B \wedge B \not\subset A$$

Sperner: if F is an antichain of X, then

$$|F| \leq \binom{n}{\lfloor n/2 \rfloor}$$

Note, an antichain is also known as a Sperner family of subsets.

Maximal Chains

Without loss of generality we assume that $X = \{1, \dots, n\}$. A maximal chain in X obviously has length $n + 1$

$$\emptyset = A_0 \subset A_1 \subset \cdots \subset A_n = X$$

Such a maximal chain puts a total order on the elements of X

- the smallest element is the single element of A_1
- the one-but-smallest is the new element in A_2
- and so on and so on

Similarly, each total order on X defines a chain

- A_1 consists of the smallest element
- A_2 consists of the two smallest elements
- and so on and so on

That is, the total number of maximal chains equals the number of permutations: $n!$

Maximal Chains and Antichains

Let $A \subseteq X$, with $|A| = k$. A maximal chain that contains A

- i.e., $A = A_k$ in that chain

consists of

- A maximal chain for the set A
- followed by a chain for $X \setminus A$
  - each set in the latter chained is extended by the union with A, of course

This means that there are $k!(n-k)!$ chains containing A.

Note that if F is an antichain, than any chain can contain at most one element of F

- If A and B are in a chain, then either $A \subset B$ or $B \subset A$
- If A and B are in F, then both $A \not\subset B$ and $B \not\subset A$

Proving Sperner

Recall that F is an antichain. The number of maximal chains that contain an element of F (and thus exactly 1) is

- $\sum_{A \in F} |A|!(n - |A|)! = \sum_{A \in F} n! \frac{|A|!(n-|A|)!}{n!} = n! \sum_{A \in F} \frac{1}{\binom{n}{|A|}}$

Because there are in total $n!$ maximal chains, we have

- $\sum_{A \in F} \frac{1}{\binom{n}{|A|}} \leq 1$

For binomial coefficients, the middle ones are the largest, hence

- $\sum_{A \in F} \frac{1}{\binom{n}{\lfloor n/2 \rfloor}} \leq \sum_{A \in F} \frac{1}{\binom{n}{|A|}} \leq 1$

Since

- $\sum_{A \in F} \frac{1}{\binom{n}{\lfloor n/2 \rfloor}} = \frac{|F|}{\binom{n}{\lfloor n/2 \rfloor}}$

We have that

$$|F| \leq \binom{n}{\lfloor n/2 \rfloor}$$

Back to Monotone Formula's

Let S be the set of all assignments to $\{x_1, \ldots, x_n\}$ such that exactly

- $\lfloor n/2 \rfloor$ variables are mapped to 1 (true)

Clearly, $|S| = \binom{n}{\lfloor n/2 \rfloor}$

- this is the definition of $\binom{a}{b}$

Now choose some $0/1$ labelling on S

- i.e., choose an arbitrary function $g : S \to \{0, 1\}$
- we need to show that M_n contains that function

Define T (from true) by

$$T = \{A \in S \mid g(A) = 1\}$$

We need to construct a monotone formula f such that

$$f(A) = 1 \leftrightarrow A \in T \leftrightarrow g(A) = 1$$

Two Special Cases and f

g maps al variables to 0 (false)

- iff $S = \emptyset$

Clearly, the function false $\in M_n$. Hence we can assume $S \neq \emptyset$

If $n = 1$, we have only 1 variable which is either mapped to 1 or to 0

- a function that is obviously in M_1

Hence we may assume that $n > 1$

Let f be the monotone function

$$f(z_1, \ldots x_n) = \bigvee_{A \in T} \bigwedge_{i : A(x_i) = 1} x_i$$

Given the assumptions made above, the disjunction isn't empty and neither is the conjunction

$VC(M_n) \geq \binom{n}{\lfloor n/2 \rfloor}$

Let $B \in T$, then the monomial

$$\bigwedge_{i:B(x_i)=1} x_i$$

is mapped to 1 by B and, thus, by f

For $B \in S \setminus T$, note that each monomial

$$\bigwedge_{i:A(x_i)=1} x_i$$

in f assigns 1 to exactly $\lfloor n/2 \rfloor$ variables and 0 to the rest. Since $B \in S \setminus T$

- it assigns 0 to at least one of these $\lfloor n/2 \rfloor$ variables

Which means that f assigns 0 to B,

In other words, M_n shatters S which has $\binom{n}{\lfloor n/2 \rfloor}$ elements. Hence $VC(M_n) \geq \binom{n}{\lfloor n/2 \rfloor}$.

$VC(M_n) \leq \binom{n}{\lfloor n/2 \rfloor}$

Let S be a set of assignments such that $|S| > \binom{n}{\lfloor n/2 \rfloor}$. For each $A \in S$ define:

$$V_A = \{i \mid A(x_i) = 1\}$$

Because of the size of S, Sperner's theorem tells us the V_A'a cannot be an antichain. Hence, there are $A_1, A_2 \in S$ such that

$$A_1(x_i) = 1 \rightarrow A_2(x_i) = 1$$

Since the functions in M_n are monotone, this means:

$$\forall f \in M_n : f(A_1) = 1 \rightarrow f(A_2) = 1$$

In other words a labelling that maps A_1 to 1 and A_2 to 0 cannot be constructed in M_n. In other words: $VC(M_n) \leq \binom{n}{\lfloor n/2 \rfloor}$ Hence

$$VC(M_n) = \binom{n}{\lfloor n/2 \rfloor}$$

Adding Negations

In the case of C_n and D_n we saw that

- adding negation did not increase the VC dimension

So, it is reasonable to expect that

- the VC dimension of all boolean functions is the same as that of M_n

This is, however,

not true!

The VC dimension of that set of hypotheses is strictly bigger.

Computing the exact dimension is pretty hard

- in fact, I am not aware of an exact expression

Bounding the dimension is easier

- for k-DNF we can compute a Θ bound

For the general case, we need some extra machinery. But first we look at k-DNF

k-DNF

Recall that k-DNF consists of disjunctions

- each component (disjunct, consisting of conjunctions) is the conjunction of at most k literals.

Computing the VC dimension exactly isn't easy, giving a bound is:

For $n, k \in \mathbb{N}$, let $D_{n,k}$ be the set of k-DNF functions (expressions) over $\{0,1\}^n$ (i.e., in n variables). Then $VC(D_{n,k}) = \Theta(n^k)$

Recall:

- $g(n) = O(f(n))$ if there exist c, n_0 such that $\forall n \geq n_0 : g(n) \leq f(n)$ (i.e., upper bound)
- $g(n) = \Omega(f(n))$ if there exist c, n_0 such that $\forall n \geq n_0 : g(n) \geq f(n)$ (i.e., lower bound)
- $g(n) = \Theta(f(n))$ if $g(n) = O(f(n))$ and $g(n) = \Omega(f(n))$

$VC(D_{n,k}) = O(n^k)$

The number of monomials of degree at most k (not identical false or empty) is:

$$\sum_{i=1}^{k} \binom{n}{i} 2^i = O(n^k) \text{ for fixed } k$$

(2^i, since the literals you choose are either a variable or its negation).

Each k-DNF formula is the disjunction of a set of such terms

$$|D_{n,k}| = 2^{O(n^k)}$$

Which means:

$$VC(D_{n,k}) = O(n^k)$$

$VC(D_{n,k}) = \Omega(n^k)$

Let $S \subseteq \{0,1\}^n$ consist of those vectors

- that have exactly k entries equal to 1

Let $R \subseteq S$

- for each $y = (y_1, \ldots, y_n) \in R$
- form the term t_y as the conjunction of the literals u_i such that $y_i = 1$
- t_y has exactly k literals and
- $\forall z \in S : t_y(z) = 1 \leftrightarrow z = y$

Hence,

$$\bigvee_{y \in R} t_y \text{ is a classifier for } R$$

That is, S is shattered by $D_{n,k}$. Since $|S| = \binom{n}{k} = \Omega(n^k)$ (for fixed k). We have:

$$VC(D_{n,k}) = \Omega(n^k)$$

An Observation

From he results we have reached – perhaps even more from the proofs of these results – one sees that

- the richer the model class, the higher the VC dimension.

This is, of course, completely logical as we have by definition that

$$\mathcal{H}_1 \subseteq \mathcal{H}_2 \rightarrow VC(\mathcal{H}_1) \leq VC(\mathcal{H}_2)$$

This observation, however, hints at a way to find good models:

- start with a very simple model class and pick the best hypothesis
- if that is good, you are done. If not take a slightly richer class

This line of thought gives rise to structural risk minimization

- rather than empirical risk minimization

which we'll discuss next week

The Growth Function

Exact bounds for larger classes of boolean functions are not known. We do, however, have a more general result which is based on the *growth function*.

The VC dimension only looks at the largest set that $\mathcal{H}$ can shatter. The growth function $\tau_{\mathcal{H}} : \mathbb{N} \to \mathbb{N}$ looks much broader to the classifications $\mathcal{H}$ contains:

$$\tau_{\mathcal{H}}(m) = \max_{C \subset X : |C| = m} |\mathcal{H}_C|$$

That is,

$$\tau_{\mathcal{H}}(m) = \max_{C \subset X : |C| = m} |\{f(c_1), \ldots, f(c_m)\}_{f \in \mathcal{H}}|$$

each $f \in \mathcal{H}$ produces a $0/1$ vector of length m and $\tau_{\mathcal{H}}$ tells you

- how many different vectors $\mathcal{H}$ can produce maximally

Growth Above VC

Clearly, if $m \leq d = VC(\mathcal{H})$ then $\tau_{\mathcal{H}} = 2^m$

- if there is a d sized set that $\mathcal{H}$ can shatter, the for each smaller integer there is also a set that $\mathcal{H}$ can shatter
- restrict (actually project) the shattering to the lower dimensional space.

It is more instructive what happens if $m > d$. The fact that $\mathcal{H}$ cannot shatter a set of size m

- doesn't mean that it is completely useless for sets of that size

It might, e.g., classify almost always almost correctly

- or it might do a horrible job for any m sized set.

Sauer's Lemma tells us what to expect above d.

Pajor's Lemma

Let $\mathcal{H}$ be any hypothesis class with $VC(\mathcal{H}) = d < \infty$. For any $C = \{c_1, \ldots, c_m\}$

$$|\mathcal{H}_C| \leq |\{B \subseteq C \mid \mathcal{H} \text{ shatters } B\}|$$

To prove this by induction, first note that for $m = 1$, either both sides are 1 or both are 2

- the empty set is shattered by all hypothesis classes.

Now assume that the inequality holds for all $k < m$

- Let $C = \{c_1, c_2, \ldots, c_m\}$ and
- let $C' = \{c_2, \ldots, c_m\}$

Define the two sets

$$Y_0 = \{(y_2, \ldots, y_m) \mid (0, y_2, \ldots, y_m) \in \mathcal{H}_C \vee (1, y_2, \ldots, y_m) \in \mathcal{H}_C\}$$
$$Y_1 = \{(y_2, \ldots, y_m) \mid (0, y_2, \ldots, y_m) \in \mathcal{H}_C \wedge (1, y_2, \ldots, y_m) \in \mathcal{H}_C\}$$

Note that $|\mathcal{H}_C| = |Y_0| + |Y_1|$, because Y_1 contains those vectors of $\mathcal{H}_C$ that generate a vector in Y_0 twice rather than once.

Proof Part 1

Since $Y_0 = \mathcal{H}_{C'}$ we have by the induction assumption that

$$
\begin{aligned}
|Y_0| = |\mathcal{H}_{C'}| &\leq |\{B \subseteq C' \mid \mathcal{H} \text{ shatters } B\}| \\
&= |\{B \subseteq C \mid c_1 \notin B \wedge \mathcal{H} \text{ shatters } B\}|
\end{aligned}
$$

Next, define $\mathcal{H}'$ to contain pairs of hypotheses that agree on C' but disagree on c_1:

$$
\begin{aligned}
\mathcal{H}' = \{h \in \mathcal{H} \mid \exists h' \in \mathcal{H} : &(1 - h'(c_1), h_2(c_2), \ldots, h_m(c_m)) \\
&= (h(c_1), h(c_2), \ldots, h_m(c_m))\}
\end{aligned}
$$

Note that

- if $\mathcal{H}'$ shatters $B \subseteq C'$ it also shatters $B \cup \{c_1\}$ and vice versa
- $Y_1 = \mathcal{H}'_{C'}$

So, by induction we can compute $|Y_1|$

Proof Part 2

Because $|C'| < m$ our induction assumption yields

$$\begin{aligned}|Y_1| = |\mathcal{H}'_{C'}| &\leq |\{B \subseteq C' \mid \mathcal{H}' \text{ shatters } B\}| \\ &= |\{B \subseteq C' \mid \mathcal{H}' \text{ shatters } B \cup \{c_1\}\}| \\ &= |\{B \subseteq C \mid c_1 \in B \land \mathcal{H}' \text{ shatters } B\}| \\ &\leq |\{B \subseteq C \mid c_1 \in B \land \mathcal{H} \text{ shatters } B\}|\end{aligned}$$

Bringing all intermediate results together gives us:

$$\begin{aligned}|\mathcal{H}_C| &= |Y_0| + |Y_1| \\ &\leq |\{B \subseteq C \mid c_1 \notin B \land \mathcal{H} \text{ shatters } B\}| \\ &\quad + |\{B \subseteq C \mid c_1 \in B \land \mathcal{H} \text{ shatters } B\}| \\ &= |\{B \subseteq C \mid \mathcal{H} \text{ shatters } B\}|\end{aligned}$$

Which was to be proven.

Sauer's Lemma

Let $\mathcal{H}$ be any hypothesis class with $VC(\mathcal{H}) = d < \infty$.

- $\forall m : \tau_{\mathcal{H}}(m) \leq \sum_{i=0}^{d} \binom{m}{i}$
- if $m \geq d$: $\tau_{\mathcal{H}}(m) < (em/d)^d$

Proof: Since $VC(\mathcal{H}) = d$, $\mathcal{H}$ shatters *no* set with more than d elements. Thus

$$|\{B \subseteq C \mid \mathcal{H} \text{ shatters } B\}| \leq \sum_{i=0}^{d} \binom{m}{i}$$

$$\begin{aligned}\sum_{i=0}^{d} \binom{m}{i} &= \sum_{i=0}^{d} \binom{m}{i} \left(\frac{m}{d}\right)^i \left(\frac{d}{m}\right)^i \leq \left(\frac{m}{d}\right)^d \sum_{i=0}^{d} \binom{m}{i} \left(\frac{d}{m}\right)^i \\ &\leq \left(\frac{m}{d}\right)^d \sum_{i=0}^{m} \binom{m}{i} \left(\frac{d}{m}\right)^i = \left(\frac{m}{d}\right)^d \left(1 + \frac{d}{m}\right)^m < \left(\frac{em}{d}\right)^d\end{aligned}$$

For last inequality use $x > 0 \rightarrow (1 + x/m)^m < e^x$

A Simple Consequence

Let $\mathcal{H}$ be a finite hypothesis set with at least two hypothesis, defined on a finite domain X

- unfortunately, 1 hypothesis isn't going to work because:

Two hypothesis $h_1, h_2 \in \mathcal{H}$ are different if

- $\exists x \in X : h_1(x) \neq h_2(x)$

That is, h_1 and h_2 are different if they are different classifications on the complete domain X

- there are, by definition, $\tau_{\mathcal{H}}(|X|)$ such classifications.

That is:

$$\tau_{\mathcal{H}}(|X|) = |H|$$

By Sauer's lemma we have $|H| < \left(\frac{e|X|}{d}\right)^d$. Which means that

$$VC(\mathcal{H}) \geq \frac{\log |\mathcal{H}|}{n + \log e}$$

Back to Boolean Functions

If $VC(\mathcal{H}) \geq 3$ the inequality of the previous slide can be improved to $VC(\mathcal{H}) \geq \frac{\log|\mathcal{H}|}{n}$.

Hence, for any large enough class B_n of boolean functions on $\{0,1\}^n$ we have that

$$\frac{\log|B|}{n} \leq VC(B_n) \leq \log|B|$$

Clearly, these bounds are much weaker than the ones we had for M_n

- but, then again, we talk about (almost) arbitrary sets here

Until now we studied classes of functions from $\{0,1\}^n$ to $\{0,1\}$. An obvious generalization is to study sets of functions

- from $\mathbb{R}^n$ to $\{0,1\}$.

We look at one such class, polynomials on $\mathbb{R}$

Polynomials as Classifiers

Recall how we saw lines and hyperplanes as classifiers

- simply by distinguishing the half spaces above and below the line/hyper plane

For polynomials we can do the same. First we define the set of polynomials of degree at most n by

$$P_n = \sum_{i=0}^{n} a_i x^i$$

for $a_i \in \mathbb{R}$. Next, for any $p \in P_n$ define the function $p_+ : \mathbb{R} \to \{0, 1\}$ by

$$p_+(x) = \begin{cases} 1 & \text{if } f(x) > 0 \\ 0 & \text{if } f(x) \leq 0 \end{cases}$$

The set of all these classifiers is known as $\mathrm{pos}(P_n)$ which we denote by $P_n^{>0}$. The question now is: determine $VC(P_n^{>0})$?

The Intuition

The fundamental theorem of algebra tells us that over the complex numbers, a polynomial of degree n can be written as

$$\beta \prod_{i=1}^{n} (x - \alpha_i) \qquad \alpha_i, \beta \in \mathbb{C}$$

In other words, the graph of a real valued degree n polynomial

- crosses the x-axis at most n times

Each such crossing

- switches the classification from 1 to 0 or vice versa

Hence we can shatter at most $n + 1$ points in $\mathbb{R}$

Each labelling of $n + 1$ points on the x-axis shows a number of adjacent change pairs $(1, 0)$ or $(0, 1)$

- construct your polynomial such that the roots are between the two points of a change pair

This will give you a separating polynomial

From Intuition to Proof

Making this intuition precise using the language of the graphs of polynomials involves lots of infuriating bookkeeping details

- wiggly lines are hard to keep under control

To make life easier

- for those who know some linear algebra

we map (embed) our data into a higher dimensional space

- and stretch the wiggly line in a linear structure: a hyperplane
- for the cognoscenti, we are using the "kernel trick", well known from SVMs, with a polynomial kernel

The mapping we use is:

$$\phi : z \to (1, z, z^2, \ldots, z^n)$$

mapping$c \in \mathbb{R}$ to the vector $(1, c, c^2, \ldots, c^n)^T \in \mathbb{R}^{n+1}$

Using ϕ

A polynomial p is given by

$$p = \sum_{i=0}^{n} a_i x^i$$

We can rewrite this as a dot product by

$$p = \sum_{i=0}^{n} a_i x^i = (a_0, a_1, \ldots a_n) \cdot (1, x, x2, \ldots, x^n)$$

The second expression should remind you of a hyperplane, perhaps all the more when evaluated on a particular instance

$$\begin{aligned} p(c) &= (a_0, a_1, \ldots a_n) \cdot (1, c, c^2, \ldots, c^n) \\ &= (a_0, a_1, \ldots a_n) \cdot \phi(c) = \phi(p)(\phi(c)) \end{aligned}$$

where $\phi(p)$ denotes the function on $\mathbb{R}^{n+1}$

Polynomials, Hyperplanes, and Thresholds

More in particular, if we turn from P_n to $P_n^{>0}$

- i.e., we turn from functions to classifiers

We see that

- $p(c) > 0$ on $\mathbb{R}$ translates to $\phi(p)(\phi(c)) > 0$ on $\mathbb{R}^{n+1}$

Now, the expression: $\phi(p)$ denotes both

- a threshold function on $\mathbb{R}^{n+1}$
- and a hyperplane on $\mathbb{R}^n$

That is, we have a 1-1 correspondence between

- polynomial classifiers and
- threshold/hyperplane classifiers

This correspondence helps us to prove our results "linearly".

$VC(P_n^{>0}) \leq n+1$

Let $S \subseteq \mathbb{R}^n$ be a set, that is shattered by $P_n^{>0}$. That is, for every $S^+ \subseteq S$ there exists a $p_+ \in P_n^{>0}$ such that

- $p_+(s) = 1$ if $s \in S^+$
- $p_+(s) = 0$ if $s \in S \setminus S^+$

In other words, there is a $p \in P_n$ such that

- $\sum_{i=0}^n a_i s^i > 0$ if $s \in S^+$
- $\sum_{i=0}^n a_i s^i \leq 0$ if $s \in S \setminus S^+$

Written in the language of dot products this says that there is a vector $a = (a_1, \ldots, a_n)$ and a constant a_0 such that

- $(a_1, \ldots, a_n) \cdot (s, s^2, \ldots, s^n) + a_0 > 0$ if $s \in S^+$
- $(a_1, \ldots, a_n) \cdot (s, s^2, \ldots, s^n) + a_0 \leq 0$ if $s \in S \setminus S^+$

Since $z \rightarrow (z, z^2, \ldots, z^n)$ simply maps $\mathbb{R} \rightarrow \mathbb{R}^n$, we have a separating hyperplane on $\mathbb{R}^n$. Hence, $|S| \leq n+1$

Independent Vectors are Shattered

To prove that $VC(P_n^{>0}) \geq n+1$, we first prove that a set $\{x_1, \ldots, x_n\} \subset \mathbb{R}^n$ of independent vectors is shattered by threshold functions on $\mathbb{R}^n$.

Let A be the $n \times n$ matrix with the x_i vectors as columns. This is an invertible matrix

- otherwise the x_i would not be independent

Let v be any of the 2^n -1/+1 vectors that denote the labellings of the x_i

- then, the matrix equation $Aw = v$ has a unique solution
- $w = A^{-1}v$

The vector w gives you the threshold function that shatters the x_i for labelling v.

Hence, if we can prove that there exists a set $\{x_0, \ldots, x_n\} \subset \mathbb{R}$ that ϕ maps to a set of independent vectors in $\mathbb{R}^{n+1}$ we are done.

P_n is a Vector Space

For that we need:

Let $f, g \in P_n$ and $\lambda \in \mathbb{R}$. Then clearly

- $f + g = \sum_{i=0}^{n}(a_i + b_i)x^i \in P_n$ and
- $\lambda f = \sum_{i=0}^{n}(\lambda a_i)x^i \in P_n$

In other words, P_n is a vector space over $\mathbb{R}$

Moreover, P_n is a $n+1$-dimensional vector space with base

$$\{1, x, \ldots, x_n\}$$

For, clearly, these functions are linearly independent

$$\left[\forall x \in \mathbb{R} : \lambda_0 + \lambda_1 x + \lambda_2 x^2 + \cdots + \lambda_n x^n = 0\right] \Leftrightarrow [\forall i : \lambda_i = 0]$$

and every element of P_n can (by definition) be written as a linear combination of these functions

$n + 1$ Independent Vectors

To see that ϕ creates $n+1$ independent vectors we argue from contradiction.

Assume that for every $X = \{x_0, \ldots, x_n\} \subset \mathbb{R}$ we have that the set of vectors $\phi(X) = \{\phi(x_0), \ldots \phi(x_n)\}$ is dependent

- then the vector subspace spanned by $\{\phi(x) \mid x \in \mathbb{R}\}$ of $\mathbb{R}^{n+1}$ has at most dimension n
- that is, it is contained in some hyperplane
- this means that there are λ_i, not all equal to 0, such that

$$\forall x \in \mathbb{R} : \sum_{i=0}^{n} \lambda_i(\phi(x)) = \sum_{i=0}^{n} \lambda_i x^i = 0$$

But that contradicts that $\{1, x, \ldots, x_n\}$ is a basis.

$VC(P_n^{>0}) \geq n+1$

We have:

- there exists a $X = \{x_0, \dots, x_n\} \subset \mathbb{R}$
- such that $\phi(X) = \{\phi(x_0), \dots \phi(x_n)\}$ is independent
- hence, $\phi(X)$ is shattered by threshold functions
- hence, X is shattered by the corresponding polynomials

In other words, $VC(P_n^{>0}) \geq n+1$. We already had that $VC(P_n^{>0}) \leq n+1$, hence we have

$$VC(P_n^{>0}) = n+1$$

For the more general case, having more variables $x_1, \dots, x_m$ see exercise 6.12 in the book

A Simple Consequence

The fact that $VC(P_n^{>0}) = n + 1$ implies that the set of all polynomials

$$P = \bigcup_{n=1}^{\infty} P_n$$

has $VC(P) = \infty$

- if $VC(P)$ would be finite, say k we have a contradiction with $VC(P_k) = k + 1$

Hence, we cannot simply learn the best fitting polynomial using the ERM rule

- recall that sets with infinite VC dimension are not PAC learnable

For that one needs a more subtle approach

- Structural Risk Minimization

Which we mentioned before and is discussed later in this course.