LFD Problem Set 4

John Cohen

September 30, 2024

Exercise 2.4

Consider the input space $X = \{1\}xR^d$ (including the constant coordinate $x_0 = 1$). Show that the VC dimension of the perceptron (with $d+1$ parameters, counting w_0) is exactly $d+1$ by showing that it is at least $d+1$ and at most $d+1$, as follows.

(a) To show that $d_{VC} \geq d+1$, find $d+1$ points in X that the perceptron can shatter. [Hint: Construct a nonsingular $(d+1)$ x $(d+1)$ matrix whose rows represent the d+1 points, then use the nonsingularity to argue that the perceptron can shatter these points.]

Construct a nonsingular $(d+1)$ x $(d+1)$ matrix X whose rows represent the d+1 points matrix, where (i,1) and (i,i) = 1 for all i=1...d+1 and the rest are 0's. X is nonsingular (e.g., its rows are linearly independent). All nonsingular matrices are invertible. Find a vector of weights w s.t. Xw = y. $w = X^{-1}y$. The perceptron can shatter any set of d+1 points and thus $d_{VC} \geq d+1$.

(b) To show that $d_{VC} \leq d+1$, show that no set of $d+2$ points in X can be shattered by the perceptron. [Hint: Represent each point in X as a vector of length $d+1$, then use the fact that any $d+2$ vectors of length $d+1$ have to be linearly dependent. This means that some vector is a linear combination of all the other vectors. Now, if you choose the class of these other vectors carefully, then the classification of the dependent vector will be dictated. Conclude that there is some dichotomy that cannot be implemented, and therefore that for $N \geq d+2$, $m_H(N) < 2^N$.]

Consider the same input set X from above where each point is a vector 1x(d+1). With d+2 such vectors, by the pigeon hole principle it is impossible for all of them to be linear independent. In a space of d+1 there are only d+1 independent vectors. As such, at least one vector must be a linear combination of another. We cannot apply -1 one vector and +1 to the another same instance of that vector becasue the perceptron cannot realize this labeling. Thus the perceptron cannot shatter the set of d+2 points and thus d+2 is a break point. With this, if N truly is greater than or equal to d+2 then we know $m_H(N) < 2^N$. It follows that $d_{VC} \leq d+1$.

Using both parts proving $d+1 \leq d_{VC} \leq d+1$, we can only conclude by the squeeze theorem that $d_{VC} = d+1$.

Problem 2.3

Compute the maximum number of dichotomies, $m_H(N)$, for these learning models, and consequently compute d_{VC}, the VC dimension.

(a) Positive or negative ray: H contains the functions which are +1 on $[a, \infty)$ (for some a) together with those that are +1 on $(-\infty, a]$ (for some a).

The growth function from previous homeworks is N+1. With negative rays, its just the same however we must subtract the 2 dichotomies that are all -1 and all +1 giving $m_H(N) = 2N$. At $m_H(3) = 6 < 8 = 2^3$. These means $d_{VC} = 2$ with a break point of 3.

(b) Positive or negative interval: H contains the functions which are +1 on an interval [a,b] and -1 elsewhere or -1 on an interval [a,b] and +1 elsewhere.

This growth function is $m_H(N) = 2\binom{N}{2} + 2$ or $m_H(N) = N^2 - N + 2$. Breaking at $m_H(4) = 14 < 16 = 2^4$, k = 4 and thus $d_{VC} = 3$.

(c) Tow concentric spheres in R^d: H contains the functions which are +1 for $a \leq \sqrt{x_1^2 + ... + x_d^2} \leq b$.

This growth function is $m_H(N) = \binom{N+1}{2} + 1$ or $m_H(N) = \frac{N^2+N+2}{2}$. Breaking at $m_H(3) = 7 < 8 = 2^3$, k = 3 and thus $d_{VC} = 2$.

Problem 2.8

Which of the following are possible growth functions $m_H(N)$ for some hypothesis set:

$$1+N; 1+N+\frac{N(N-1)}{2}; 2^N; 2^{[\sqrt{N}]}; 2^{[N/2]}; 1+N+\frac{N(N-1)(N-2)}{6}$$

Remember: Theorem 2.4: If $m_H(N) < 2^k$ for some value k (k is a break-point by these statement), then

$$m_H(N) \leq \sum_{i=0}^{k-1} \binom{N}{i}$$

for all N. The RHS is a polynomial in N of degree k-1.
From this we can conclude 2 cases:

(1): If such a break point exists, the growth function $m_H(N)$ is bounded by a polynomial of degree k1.

(2): If no finite break point exists (i.e., the VC dimension is infinite), then $m_H(N) = 2^N$

(i) 1+N;
Break point k=2 as $m_H(N) = 1 + 2 = 3 < 4 = 2^2$.
Growth function is bounded by polynomial of degree 1.
Thus this growth is possible corresponding to a VC dimension of $d_{VC} = 1$

(ii) $1 + N + \frac{N(N-1)}{2}$;
$m_H(N) = \binom{N}{0} + \binom{N}{1} + \binom{N}{2}$
Break point k = 3 as $m_H(N) = 1 + 2 + 3 = 7 < 8 = 2^3$.
Growth function is bounded by polynomial of degree 2 (k-1).
Thus this growth is possible corresponding to a VC dimension of $d_{VC} = 2$

(iii) 2^N;
Refer to case 2 above. To spell it out, there is no break-point and the growth function is the maximum number of dichotomies. The VC dimension is ∞ and this is very much possible.

(iv) $2^{[\sqrt{N}]}$;
Growth function is exponential in $\sqrt{N}$ with is below N. That is, this function is neither 2^k for infinity nor polynomial growth for finite d_{VC}.
Thus it is not possible.

(v) $2^{[N/2]}$;
Same as (iv) above.

(vi) $1 + N + \frac{N(N-1)(N-2)}{6}$
Simplified: $m_H(N) = \binom{N}{0} + \binom{N}{1} + \binom{N}{3}$
$m_H(2) = 1 + 2 + 0 = 3 < 4 = 2^2$
here we see k=2 as $m_H(k) < 2^k$.
By Theorem 2.4 and with k=2, $m_H(N) \leq \binom{N}{0} + \binom{N}{1}$
Substitute: $1 + N + \frac{N(N-1)(N-2)}{6} \leq 1 + N$ for all N.
This statement is obviously false. For example N=3 breaks this.
Thus, this growth is not possible.

Problem 2.10

Show that $m_H(2N) \leq m_H(N)^2$, and hence obtain a generalization bound which only involves $m_H(N)$.

Consider a set S s.t. $S = S_1 \cup S_2, |S_1| = |S_2| = N, S_1 \cap S_2 = 0$.
The number of possible labelings for S_1 and S_2 each is at the very most $m_H(N)$. Therefore, the number of possible labelings for S is at most $m_H(2N) \leq m_H(N) * m_H(N) = m_H(N)^2$.
This implies $\sqrt{\frac{8}{N} ln(\frac{4m_H(2N)}{\delta})} \leq \sqrt{\frac{8}{N} ln(\frac{4m_H(N)^2}{\delta})}$ giving a new generalization bound of $E_{out}(g) \leq E_{in}(g) + \sqrt{\frac{8}{N} ln(\frac{4m_H(N)^2}{\delta})}$

Problem 2.12

For an H with $d_{VC} = 10$, what sample size do you need (as prescribed by the generalization bound) to have 95% confidence that your generalization error is at most 0.05?

Given: $d_{VC} = 10, \epsilon = (1 - 0.95) = 0.05, \delta = 0.05$

$$N \geq \frac{8}{0.05^2} ln(\frac{4(2N)^{10} + 4}{0.05})$$

Based on what we did in class, this is an iterative process involving an initial guess. My starting point is N=1000. Let's start:

$$N \geq \frac{8}{0.05^2} ln(\frac{4(2(1000))^{10} + 4}{0.05}) \approx 257251$$
$$N \geq \frac{8}{0.05^2} ln(\frac{4(2(257251))^{10} + 4}{0.05}) \approx 434853$$
$$N \geq \frac{8}{0.05^2} ln(\frac{4(2(434853))^{10} + 4}{0.05}) \approx 451652$$
$$N \geq \frac{8}{0.05^2} ln(\frac{4(2(452865))^{10} + 4}{0.05}) \approx 452950$$
$$N \geq \frac{8}{0.05^2} ln(\frac{4(2(452950))^{10} + 4}{0.05}) \approx 452956$$
$$N \geq \frac{8}{0.05^2} ln(\frac{4(2(452956))^{10} + 4}{0.05}) \approx 452957$$
$$N \geq \frac{8}{0.05^2} ln(\frac{4(2(452957))^{10} + 4}{0.05}) \approx 452957$$

N converges at $N \approx 452957$.