Machine Learning from Data Assignment 4

Greg Stewart

October 1, 2018

Exercise 2.4

Consider the input space $X = \{1\} \times \mathbb{R}^d$. Show that the VC dimension of the perceptron, with $d+1$ parameters, 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.*

Let there be a set of $d+1$ points in $\mathbb{R}^d$ that are shattered by the perceptron, e.g.

$$X = \begin{pmatrix} x_1^T \\ x_2^T \\ \vdots \\ x_{d+1}^T \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ 1 & 1 & 0 & \cdots & 0 \\ 1 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & 0 & 0 & \cdots & 1 \end{pmatrix}$$

We need to find a

$$y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_{d+1} \end{pmatrix}$$

where $y_i \in \{-1, +1\}$ that satisfies $\text{sign}(Xw) = y$, where w is some set of weights. This can easily be accomplished with

$$w = X^{-1}y$$

because X is invertible. Therefore, $d_{VC} \geq d+1$.

(b) *To show $d_{VC} \leq d+1$, show that no set of $d+2$ points in X can be shattered by the perceptron.*

Given a set of $d+2$ points, $\{x_1, x_2, \ldots, x_{d+2}\}$, we have more vectors than dimensions, so not all of them can be linearly independent. Therefore, for some point x_i in the set, we can write it as a linear combination:

$$x_i = \sum_{j=1, j\neq i}^{d+2} \alpha_j x_j.$$

It is straightforward to create a dichotomy that can't actually be generated:

$$y_j = \begin{cases} \text{sign}(\alpha_j) & j \neq i \\ -1 & j = i \end{cases}$$

We can assume the labels are correxts as $\text{sign}(\alpha_j) = \text{sign}(w^T x_j)$. Then $\alpha_j w^T x_j > 0$. So for the i^{th} point, we have

$$w^T x_i = \sum_{j \neq i} \alpha_j w^T x_j > 0$$

so $y_i = +1$ which contradicts the previously constructed dichotomy. Therefore

$$d_{VC} \leq d + 1$$

.

Since we have both $d_{VC} \geq d + 1$ and $d_{VC} \leq d + 1$, it must be the case 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}.

(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).*

$$m_H(N) = 2N$$

The pattern beging to become obvious at $N = 3$. The VC dimension is

$$d_{VC} = 2.$$

(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.*

(c) *Two concentric spheres in $\mathbb{R}^4$: H contains the functions which are +1 for $a \leq \sqrt{x_1^2 + \cdots + x_d^2} \leq b$.*

As these are spheres, the middle term of the equality can be thought of as a radius, reducing the problem to 1-dimensional positive intervals, i.e.

$$a \leq r \leq b.$$

This means the growth function is

$$m_H(N) = \binom{N+1}{2} + 1 = \frac{N^2}{2} + \frac{N}{2} + 1.$$

Thus the VC dimension is

$$d_{VC} = 2$$

Problem 2.8

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

$1 + N$: Yes, this is possible.

$1 + N + \frac{N(N-1)}{2}$: Yes, this is possible.

2^N: Yes, this is possible, and is in fact the upper bound.

$2^{\lfloor \sqrt{N} \rfloor}$: No, this is not possible because it is neither polynomial nor 2^N.

$2^{\lfloor N/2 \rfloor}$: No, not possible again because it is neither polynomial nor 2^N.

$1 + N + \frac{N(N-1)(N-2)}{6}$: Yes, this is possible.

Problem 2.10

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

In the worst case, where $m_H(N) = 2^N$, we have

$$m_H(2N) = 2^{2N} \leq (2^N)^2 = 2^{2N}$$

Now, if the growth function is not exponential, it must be a finite polynomial, so we can write it as

$$m_H(N) = a_1 N^{d_{VC}} + a_2 N^{d_{VC}-1} + \cdots + a_{d_{VC}} N + b$$

Then for $m_H(2N)$ we have

$$m_H(N) = a_1 2^{d_{VC}} N^{d_{VC}} + a_2 2^{d_{VC}-1} N^{d_{VC}-1} + \cdots + 2a_{d_{VC}} N + b$$

And for $m_H(N)^2$ we have

$$\begin{aligned} m_H(N)^2 &= (a_1 N^{d_{VC}} + a_2 N^{d_{VC}-1} + \cdots + a_{d_{VC}} N + b)(a_1 N^{d_{VC}} + a_2 N^{d_{VC}-1} + \cdots + a_{d_{VC}} N + b) \\ &= a_1^2 N^{2d_{VC}} + a_1 a_2 N^{d_{VC}(d_{VC}-1)} + \cdots + a_{d_{VC}}^2 N^2 + b^2 \end{aligned}$$

The largest term in $m_H(N)^2$ (and many more terms after it) is larger than than the largest term in $m_H(2N)$. i.e.

$$a_1 2^{d_{VC}} N^{d_{VC}} + a_2 2^{d_{VC}-1} N^{d_{VC}-1} + \cdots + 2a_{d_{VC}} N + b \leq a_1^2 N^{2d_{VC}} + a_1 a_2 N^{d_{VC}(d_{VC}-1)} + \cdots + a_{d_{VC}}^2 N^2 + b^2$$

Therefore we know that

$$m_H(2N) \leq m_H(N)^2.$$

With this conclusion we can restate the generalization bound and replace $m_H(2N)$ with $m_H(N)^2$.

$$E_{out}(g) \leq E_{in}(g) + \sqrt{\frac{8}{N} \ln \frac{4m_H(N)^2}{\delta}}$$

Problem 2.12

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

We use formula derived from the generalization bound to obtain this iteratively:

$$N \geq \frac{8}{\epsilon^2} \ln \frac{4((2N)^{d_{VC}} + 1)}{\delta}$$

Using $\epsilon = 0.05, \delta = 0.05$, and the given VC dimension, let's start with $N = 1000$.

$$\begin{aligned} N = 1000 &\implies N = 62668 \\ N = 62668 &\implies N = 89150 \\ N = 89150 &\implies N = 91406 \\ N = 91406 &\implies N = 91566 \\ N = 91566 &\implies N = 91577 \\ N = 91577 &\implies N = 91578 \end{aligned}$$

So let's settle on using a sample size of $N \approx 91600$ to get the desired outcome.