Solutions to
Understanding Machine Learning
by Shai Shalev-Shwartz and Shai Ben-David

Chapter 6 (VC-Dimension)

Zahra Taheri[1]

Apr 2020

Exercise 6.2

1. If $k = 0$ then obviously $VCdim(\mathcal{H}) = 0$. Let $k \geq 1$. We prove that $VCdim(\mathcal{H}) = min\{k, |\mathcal{X}| - k\}$. Firstly we show that if $C \subset \mathcal{X}$ such that $l := |C| = min\{k, |\mathcal{X}| - k\}$ then $\mathcal{H}$ shatters C. Let $C = \{c_1, \ldots, c_l\}$ and $(y_{c_1}, \ldots, y_{c_l}) \subset \{0,1\}^l$ be an arbitrary vector of labels of C and $s := \sum_{i=1}^{l} y_{c_i}$. Since $s \leq k$, there exists $C' \subset \mathcal{X} \setminus C$ such that $|C'| = k - s$. Define the function $h : \mathcal{X} \to \{0,1\}$ such that $h(c_i) = y_{c_i}$ for all $i \in \{1, \ldots, l\}$, $h(c') = 1$ for all $c' \in C'$, and $h(x) = 0$ for all $x \in \mathcal{X} \setminus (C \cup C')$. It is easy to see that $h \in \mathcal{H}$ and so, $\mathcal{H}$ shatters C. Therefore, $VCdim(\mathcal{H}) \geq min\{k, |\mathcal{X}| - k\}$. Now we prove that if $C \subset \mathcal{X}$ such that $|C| = min\{k, |\mathcal{X}| - k\} + 1$ then $\mathcal{H}$ does not shatter C. If $|C| = k + 1$ or $|C| = |\mathcal{X}| - k + 1$, then the all one vector or the all zero vector cannot be realized by $\mathcal{H}$, respectively. So, $\mathcal{H}$ does not shatter C and this completes the proof.

2. It is easy to see that

$$\begin{aligned}\mathcal{H}_{at-most-k} = &\{h \in \{0,1\}^{\mathcal{X}} : |\{x : h(x) = 1\}| = k\} \cup \{h \in \{0,1\}^{\mathcal{X}} : |\{x : h(x) = 0\}| = k\} \\ &\cup \{h \in \{0,1\}^{\mathcal{X}} : |\{x : h(x) = 1\}| < k\} \cup \{h \in \{0,1\}^{\mathcal{X}} : |\{x : h(x) = 0\}| < k\}.\end{aligned}$$

So by part (1), $VCdim(\mathcal{H}_{at-most-k}) \geq min\{k, |\mathcal{X}| - k\}$. We prove that $VCdim(\mathcal{H}_{at-most-k}) = k$. Firstly, we show that if $C \subset \mathcal{X}$ such that $|C| = k$ then $\mathcal{H}_{at-most-k}$ shatters C. Let $C = \{c_1, \ldots, c_k\}$ and $(y_{c_1}, \ldots, y_{c_k}) \subset \{0,1\}^k$ be an arbitrary vector of labels of C. By definition of $\mathcal{H}_{at-most-k}$, there exists $h \in \mathcal{H}_{at-most-k}$ such that $h(c_i) = y_{c_i}$ for all $i \in \{1, \ldots, k\}$, because $|C| = k$ and so the number of positive (negative) labels of C is at most k. So, $\mathcal{H}_{at-most-k}$ shatters C. Therefore, $VCdim(\mathcal{H}_{at-most-k}) \geq k$. Now let $C \subset \mathcal{X}$ such that $|C| = k + 1$. Then all one vector of labels cannot be realized by $\mathcal{H}_{at-most-k}$ and so, $\mathcal{H}_{at-most-k}$ does not shatter C.

[1] `https://github.com/zahta/Exercises-Understanding-Machine-Learning`

Exercise 6.4

In Sauer's lemma proof we proved that for every class $\mathcal{H}$ of finite VC-dimension d, and every subset A of the domain,

$$|\mathcal{H}_A| \overset{(1)}{\leq} |\{B \subseteq A : H \ \ shatters \ \ B\}| \overset{(2)}{\leq} \sum_{i=0}^{d} \binom{|A|}{i}$$

In the following, we give 4 examples in which the previous two inequalities are strict.

1. $((1),(2)) = (=,=)$ Let $\mathcal{H}$ be the class of threshold functions over $\mathbb{R}$, $h_a(x) = 1_{x<a}$. It is easy to see that $d := VCdim(\mathcal{H}) = 1$. So if $|A| = 1$, then we have

$$|\mathcal{H}_A| = |\{B \subseteq A : H \ \ shatters \ \ B\}| = \sum_{i=0}^{d} \binom{|A|}{i} = 2.$$

Also if $|A| = 2$, then

$$|\mathcal{H}_A| = |\{B \subseteq A : H \ \ shatters \ \ B\}| = \sum_{i=0}^{d} \binom{|A|}{i} = 3.$$

It is easy to see that for every subset A of the domain,

$$|\mathcal{H}_A| = |\{B \subseteq A : H \ \ shatters \ \ B\}| = \sum_{i=0}^{d} \binom{|A|}{i} = |A| + 1.$$

2. $((1),(2)) = (<,=)$ Let $m \geq 2$, $\mathcal{X} = \mathbb{R}^m$ and $\mathcal{H} = \{x \mapsto sign(\langle w, x\rangle + b) : w \in \mathbb{R}^m, b \in \mathbb{R}\}$ be the class of non-homogenous halfspaces in $\mathbb{R}^m$. We proved that $d := VCdim(\mathcal{H}) = m+1$. Let $m = 2$ and $A = \{a_1, a_2, a_3, a_4\}$ such that $a_1 = (2,1)$, $a_2 = (1,2)$, $a_3 = (2,3)$ and $a_4 = (3,2)$. It can be seen that all the labels except $(1,-1,1,-1)$ and $(-1,1,-1,1)$ can be realized by $\mathcal{H}$. So $|\mathcal{H}_A| = 14$. also we have

$$|\{B \subseteq A : H \ \ shatters \ \ B\}| = \sum_{i=0}^{3} \binom{|A|}{i} = 15.$$

3. $((1),(2)) = (<,<)$ Similar to part (2), let $m = 2$, $\mathcal{X} = \mathbb{R}^m$ and $\mathcal{H} = \{x \mapsto sign(\langle w, x\rangle + b) : w \in \mathbb{R}^m, b \in \mathbb{R}\}$ be the class of non-homogenous halfspaces in $\mathbb{R}^m$. Let $A = \{a_1, a_2, a_3\}$ such that $a_1 = (1,1)$, $a_2 = (2,2)$ and $a_3 = (3,3)$. It can be seen that all the labels except $(1,-1,1)$ and $(-1,1,-1)$ can be realized by $\mathcal{H}$. So $|\mathcal{H}_A| = 6$, $|\{B \subseteq A : H \ \ shatters \ \ B\}| = 7$ and $\sum_{i=0}^{3} \binom{|A|}{i} = 8$.

4. $((1),(2)) = (=,<)$ Let $\mathcal{H}$ be the class of axis aligned rectangles in $\mathbb{R}^2$, $\mathcal{H} = \{h_{(x_1,y_1,x_2,y_2)} : x_1 \leq y_1, x_2 \leq y_2\}$. It is easy to see that $d := VCdim(\mathcal{H}) = 4$ (see Section 6.3.3 of the book). If $A = \{a_1, a_2, a_3\}$

such that $a_1 = (1,1)$, $a_2 = (2,2)$ and $a_3 = (3,3)$, then exactly one label $(1,0,1)$ cannot be realized by $\mathcal{H}$. So $|\mathcal{H}_A| = |\{B \subseteq A : H \ \ shatters \ \ B\}| = 7$ and $\sum_{i=0}^{d} \binom{|A|}{i} = 8$.

Exercise 6.6

1. By part (1) of Exercise 3.4, $|\mathcal{H}^d_{con}| = 3^d + 1$.

2. For any subset C of the domain, if $|\mathcal{H}^d_{con}| < 2^{|C|}$, then C cannot be shattered by $\mathcal{H}^d_{con}$. So we need $VCdim(\mathcal{H}^d_{con}) \leq \lfloor log_2(|\mathcal{H}^d_{con}|)\rfloor$ to shatter C. So $VCdim(\mathcal{H}^d_{con}) \leq d \ log3$.

3. Let $y = (y_1, \dots, y_d)$ be an arbitrary vector of labels of $C = \{e_j\}_{j=1}^d$. Consider $I \subseteq \{1, \dots, d\}$ such that for all $i \in I$, $y_i = 0$, and $y_j = 1$, for all $j \in \{1, \dots, d\} \setminus I$. Define a boolean conjunction h as follows:

$$h := \begin{cases} h_{empty} & \text{if } I = \varnothing \\ x_1 \wedge \overline{x_1} & \text{if } I = \{1, \dots, d\} \\ \bigwedge_{i \in I} \overline{x_i} & \text{otherwise} \end{cases}$$

where h_{empty} is the empty conjunction which is interpreted as the all-positive hypothesis. It is easy to see that $(h(e_1), \dots, h(e_d)) = (y_1, \dots, y_d)$. Therefore $VCdim(\mathcal{H}^d_{con}) \geq d$.

4. Suppose, for a contradiction, that there exists a subset $C = \{c_1, \dots, c_{d+1}\}$ of the domain that is shattered by $\mathcal{H}^d_{con}$, i.e. $|\mathcal{H}^d_{con_C}| = 2^{d+1}$. Let $h_1, \dots, h_{d+1}$ be hypotheses in $\mathcal{H}^d_{con}$ such that for all $i, j \in [d+1]$,

$$h_i(c_j) := \begin{cases} 0 & \text{if } i = j \\ 1 & \text{otherwise} \end{cases}$$

For all $i \in [d+1]$, the conjunction that corresponds to h_i contains some literal l_i which is false on c_i and true on c_j for all $j \neq i$. By definition, a literal over the variables $x_1, \dots, x_d$ is a simple Boolean function that takes the form $f(x) = x_i$, for some $i \in [d]$, or $f(x) = 1 - x_i$ for some $i \in [d]$. So by Pigeonhole principle, there must be a pair $i < j \leq d+1$ such that l_i and l_j use the same x_k. If $l_i = l_j$, then l_i is true on c_i because l_j is true on c_i, that is a contradiction because l_i must be false on c_i. If $l_i \neq l_j$, then $\{l_i, l_j\} = \{x_k, 1 - x_k\}$. So for some $t \in [d+1] \setminus \{i, j\}$, $h_i(c_t)$ is negative because $h_j(c_t)$ is positive, that is a contradiction. So, $VCdim(\mathcal{H}^d_{con}) \leq d$.

5. Consider the class $\mathcal{H}^d_{mcon}$ of monotone Boolean conjunctions over $\{0,1\}^d$. Monotonicity here means that the conjunctions do not contain negations. We augment $\mathcal{H}^d_{mcon}$ with the all-negative hypothesis h^-. Firstly, similar to part (1) of Exercise 3.4, it can be shown that $|\mathcal{H}^d_{mcon}| = 2^d+1$. Since $VCdim(\mathcal{H}^d_{mcon}) \leq \lfloor log_2(|\mathcal{H}^d_{mcon}|)\rfloor$, we have $VCdim(\mathcal{H}^d_{mcon}) \leq d$. Similar to part (3), let $y = (y_1, \dots, y_d)$ be an arbitrary vector of labels of $C = \{1 - e_j\}_{j=1}^d$, where $1 = (1, \dots, 1)$. Consider $I \subseteq \{1, \dots, d\}$ such that for all $i \in I$,

$y_i = 0$, and $y_j = 1$, for all $j \in \{1, \ldots, d\} \setminus I$. Define a boolean conjunction h as follows:

$$h := \begin{cases} h_{empty} & \text{if } I = \varnothing \\ h^{-} & \text{if } I = \{1, \ldots, d\} \\ \bigwedge_{i \in I} x_i & \text{otherwise} \end{cases}$$

It is easy to see that $h \in \mathcal{H}^d_{mcon}$ and $(h(1-e_1), \ldots, h(1-e_d)) = (y_1, \ldots, y_d)$. Therefore, $VCdim(\mathcal{H}^d_{mcon}) \geq d$ and so we have $VCdim(\mathcal{H}^d_{mcon}) = d$.

Exercise 6.9

Let $\mathcal{H} = \{h_{a,b,s} : a \leq b, s \in \{-1, 1\}\}$ where

$$h_{a,b,s}(x) = \begin{cases} s & \text{if } x \in [a, b] \\ -s & \text{if } x \notin [a, b] \end{cases}$$

Let $C = \{a_1, a_2, a_3\}$ such that $a_1 < a_2 < a_3$. The set of all vector of labels of C is as follows:

$$\{(-1,-1,-1), (1,-1,-1), (-1,1,-1), (-1,-1,1), (1,1,-1), (1,-1,1), (-1,1,1), (1,1,1)\}.$$

(1) $h_{a,b,s}$ with $a = a_3 + 1$, $b = a_3 + 2$ and $s = 1$ realizes $(-1,-1,-1)$.

(2) $h_{a,b,s}$ with $a = a_1$, $b = \frac{a_1+a_2}{2}$ and $s = 1$ realizes $(1,-1,-1)$.

(3) $h_{a,b,s}$ with $a = a_2$, $b = \frac{a_2+a_3}{2}$ and $s = 1$ realizes $(-1,1,-1)$.

(4) $h_{a,b,s}$ with $a = a_3$, $b = a_3 + 1$ and $s = 1$ realizes $(-1,-1,1)$.

(5) $h_{a,b,s}$ with $a = a_3$, $b = a_3 + 1$ and $s = -1$ realizes $(1,1,-1)$.

(6) $h_{a,b,s}$ with $a = a_2$, $b = \frac{a_2+a_3}{2}$ and $s = -1$ realizes $(1,-1,1)$.

(7) $h_{a,b,s}$ with $a = a_1$, $b = \frac{a_1+a_2}{2}$ and $s = -1$ realizes $(-1,1,1)$.

(8) $h_{a,b,s}$ with $a = a_3 + 1$, $b = a_3 + 2$ and $s = -1$ realizes $(1,1,1)$.

So, $VCdim(\mathcal{H}) \geq 3$.

Let $C = \{a_1, a_2, a_3, a_4\}$ such that $a_1 < a_2 < a_3 < a_4$. Consider the vector of labels $y = (1,-1,1,-1)$. By definition of $\mathcal{H}$, y cannot be realized with $\mathcal{H}$. Therefore, $VCdim(\mathcal{H}) = 3$.

Exercise 6.10

Let A be a learning algorithm for the task of binary classification. Let $\mathcal{X}$ be the domain, $k \geq 2$ and m be any number smaller than or equal to $|\mathcal{X}|/k$, representing a training set size. Then by Exercise 5.3, there exists a distribution $\mathcal{D}$ over $\mathcal{X} \times \{0, 1\}$ such that:

1) There exists a function $f : \mathcal{X} \to \{0,1\}$ with $L_{\mathcal{D}}(f) = 0$.

2) $\mathbb{E}_{S\sim\mathcal{D}^m}[L_{\mathcal{D}}(A(S))] \geq \frac{1}{2} - \frac{1}{2k}$.

Now let $\mathcal{H}$ be a class of functions from $\mathcal{X}$ to $\{0,1\}$. 1. Let $VCdim(\mathcal{H}) \geq d$, for any d, and let C be a subset of the domain that is shattered by $\mathcal{H}$ and $|C| = d$. So $\mathcal{H}$ contains all functions from C to $\{0,1\}$. Without loss of generality we may assume that $\mathcal{X} = C$. Suppose that m is a training set size. We should assume that $m < d$, $k = d/m$ and $k \geq 2$. By Exercise 5.3, for every learning algorithm A, there exists a distribution $\mathcal{D}$ over $\mathcal{X} \times \{0,1\}$ such that $min_{h\in\mathcal{H}} L_{\mathcal{D}}(h) = 0$ and we have

$$\mathbb{E}_{S\sim\mathcal{D}^m}[L_{\mathcal{D}}(A(S))] \geq \frac{1}{2} - \frac{1}{2k} = min_{h\in\mathcal{H}} L_{\mathcal{D}}(h) + \frac{d-m}{2d}.$$

2. Let $VCdim(\mathcal{H}) = \infty$. Also let C be a subset of the domain $\mathcal{X}$ that is shattered by $\mathcal{H}$ and $|C| = \infty$. So $\mathcal{H}$ contains all functions from C to $\{0,1\}$. Without loss of generality we may assume that $\mathcal{X} = C$. Then by Corollary 5.2, $\mathcal{H}$ is not PAC learnable. Therefore, for every $\mathcal{H}$ that is PAC learnable, $VCdim(\mathcal{H}) < \infty$.

Exercise 6.11

Let $\mathcal{H}_1, \ldots, \mathcal{H}_r$ be hypothesis classes over some fixed domain set $\mathcal{X}$.

1. Let $d = max_i\, VCdim(\mathcal{H}_i)$ and assume for simplicity that $d \geq 3$. Let $\mathcal{H} = \bigcup_{i=1}^{r} \mathcal{H}_i$, $VCdim(\mathcal{H}) = k$ and C be a subset of the domain of size k that is shattered by $\mathcal{H}$. Therefore, $\mathcal{H}$ can produce all 2^k possible labelings of C. Also since for all $i \in [r]$, $VCdim(\mathcal{H}_i) \leq d \leq k$, by Sauer's lemma we have $\tau_{\mathcal{H}_i}(k) \leq (\frac{ek}{d})^d$. Therefore, $\tau_{\mathcal{H}_i}(k) < k^d$ because $d \geq 3$. On the other hand, by definition of the growth function we have $\tau_{\mathcal{H}}(k) \leq \sum_{i=1}^{r} \tau_{\mathcal{H}_i}(k)$. So, $\tau_{\mathcal{H}}(k) < rk^d$. Therefore, $2^k < rk^d$ and so $k < d\ log(k) + log(r)$. Hence by Lemma A.2., $k < 4d\ log(2d) + 2log(r)$ and so we have

$$VCdim(\mathcal{H}) < 4d\ log(2d) + 2log(r).$$

2. Let $d = max\{VCdim(\mathcal{H}_1), VCdim(\mathcal{H}_2)\}$, and $\mathcal{H} = \mathcal{H}_1 \cup \mathcal{H}_2$ and $VCdim(\mathcal{H}) = k$. Suppose, for a contradiction, that $k > 2d+1$ and so $k \geq 2d+2$. Since $\tau_{\mathcal{H}}(k) \leq \tau_{\mathcal{H}_1}(k) + \tau_{\mathcal{H}_2}(k)$, by Sauer's lemma we

have

$$\tau_{\mathcal{H}}(k) \leq \tau_{\mathcal{H}_1}(k) + \tau_{\mathcal{H}_2}(k) \leq \sum_{i=0}^{d} \binom{k}{i} + \sum_{i=0}^{d} \binom{k}{i}$$
$$= \sum_{i=0}^{d} \binom{k}{i} + \sum_{i=0}^{d} \binom{k}{k-i} = \sum_{i=0}^{d} \binom{k}{i} + \sum_{i=k-d}^{k} \binom{k}{i}$$
$$\leq \sum_{i=0}^{d} \binom{k}{i} + \sum_{i=d+2}^{k} \binom{k}{i} < \sum_{i=0}^{d} \binom{k}{i} + \sum_{i=d+1}^{k} \binom{k}{i}$$
$$= \sum_{i=0}^{k} \binom{k}{i} = 2^k$$

So $\tau_{\mathcal{H}}(k) < 2^k$, that is a contradiction because $\tau_{\mathcal{H}}(k) = 2^k$. Therefore, $VCdim(\mathcal{H}) = k \leq 2d + 1$.