4.4 Vapnik-Chervonenkis Dimension

This section studies some useful properties of the shatter coefficient of $\mathcal{A}$, a collection of subsets of $\mathcal{X}$. Recall that

$$s(\mathcal{A}, n) := \max_{(x_1,\ldots,x_n)\in\mathcal{X}^n} N_{\mathcal{A}}(x_1, \ldots, x_n).$$

From the definition, the results given below follow immediately.

Result 2.

(a) If $|\mathcal{A}| < \infty$, $s(\mathcal{A}, n) \leq |\mathcal{A}|$ *for all* $n \geq 1$.

(b) $s(\mathcal{A}, n) \leq 2^n$ *for all* n.

(c) If $s(\mathcal{A}, k) < 2^k$ *for some* k *then* $s(\mathcal{A}, n) < 2^n$ *for all* $n > k$.

Definition 2 (Shattering)**.** If $N_{\mathcal{A}}(x_1, \ldots, x_n) = 2^n$ for $(x_1, \ldots, x_n) \in \mathcal{X}^n$ then we say that the set $\{x_1, \ldots, x_n\}$ is *shattered* by $\mathcal{A}$.

There is a critical number of points which can be shattered by $\mathcal{A}$.

Definition 3 (V-C dimension)**.** For $\mathcal{A}$, we define $V_{\mathcal{A}}$ to be the largest $k \geq 1$ for which $s(\mathcal{A}, k) = 2^k$, and call $V_{\mathcal{A}}$ the Vapnik-Chervonenkis (V-C) dimension of $\mathcal{A}$. If $s(\mathcal{A}, n) = 2^n$ for all n, then we define $V_{\mathcal{A}} = \infty$.

Remark 6. $V_{\mathcal{A}}$ can be viewed as the size (or measure of capacity) of class $\mathcal{A}$. To prove that $V_{\mathcal{A}} = n$ we have to show $s(\mathcal{A}, k) = 2^k$ for all $k \leq n$. This means that for every $k \leq n$, there exists at least one set of k points, $(x_1, \ldots, x_k) \in \mathcal{X}^k$ which can be shattered by $\mathcal{A}$.

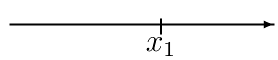
Lemma 3. *If* $|\mathcal{A}| < \infty$, *then* $V_{\mathcal{A}} \leq \log_2 |\mathcal{A}|$.

Proof. For $\mathcal{A}$ to have the V-C dimension of $V_{\mathcal{A}}$, we need $2^{V_{\mathcal{A}}}$ sets, yet $2^{V_{\mathcal{A}}} \leq |\mathcal{A}|$. It implies that $V_{\mathcal{A}} \leq \log_2 |\mathcal{A}|$. □

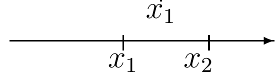
We look at a few examples of shatter coefficients and V-C dimension for some classes of sets.

Example 1.

(a) $\mathcal{A} = \{(-\infty, x] : x \in \mathbb{R}\}$ and $\mathcal{X} = \mathbb{R}$.

$s(\mathcal{A}, 1) = 2$ 

$\{\phi\}$, $\{x_1\}$

$s(\mathcal{A}, 2) = 3 < 2^2$ 

$\{\phi\}$, $\{x_1\}$, $\{x_1, x_2\}$

$\vdots$

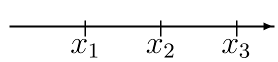
$s(\mathcal{A}, n) = n + 1 < 2^n$

$\Rightarrow V_{\mathcal{A}} = 1$.

(b) $\mathcal{A} = \{(a, b) : a < b \in \mathbb{R}\}$ and $\mathcal{X} = \mathbb{R}$.

$s(\mathcal{A}, 1) = 2$

$s(\mathcal{A}, 2) = 4 = 2^2$

$s(\mathcal{A}, 3) = 7 < 2^3$ 

$\{x_1, x_3\}$ cannot be picked out.

$\Rightarrow V_{\mathcal{A}} = 2$.

(c) $\mathcal{A} = \{\{(x_1, x_2) \in \mathbb{R}^2 : \ \beta_0 + \beta_1 x_1 + \beta_2 x_2 \geq 0\} : \ \beta_0, \beta_1, \beta_2 \in \mathbb{R}\}$ and $\mathcal{X} = \mathbb{R}^2$. See Figure 2)

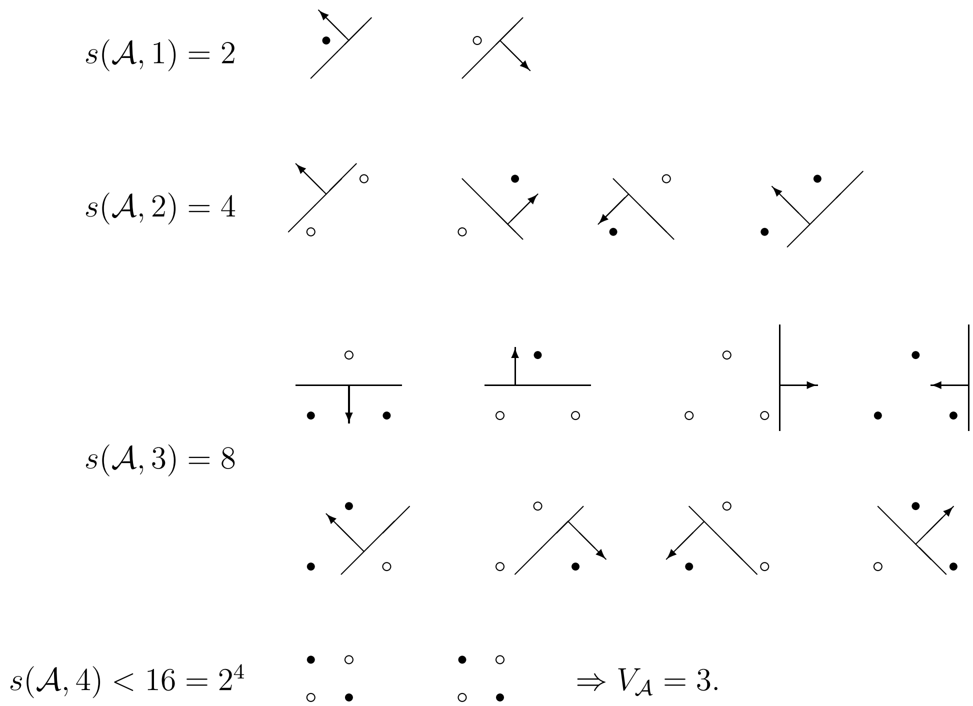


Figure 2: Shatter coefficients of half-spaces

Theorem 8. *For $\mathcal{X} = \mathbb{R}^d$, consider the class of sets induced by linear decision rules, that is,*

$$\mathcal{A} = \{H^+_{\beta_0,\beta} : \ \beta_0 \in \mathbb{R}, \ \beta \in \mathbb{R}^d\} \quad \text{where} \quad H^+_{\beta_0,\beta} = \{x \in \mathbb{R}^d : \ \beta_0 + \beta' x \geq 0\}.$$

The V-C dimension of $\mathcal{A}$ is $d+1$.

Proof. For $x_1, \ldots, x_n$ in $\mathbb{R}^d$, consider

$$B = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ x_1 & x_2 & \cdots & x_n \end{pmatrix}_{(d+1)\times n}.$$

Claim: $\{x_1, \ldots, x_n\}$ is shattered by $\mathcal{A} \iff$ the columns of B are linearly independent.

(i) If the columns of B are linearly independent, we show that $\{x_1, \ldots, x_n\}$ is shattered by $\mathcal{A}$.
If $\text{rank}(B) = n$, then $\text{rank}(B') = n$. Therefore, the row space of B has a dimension n and $B' : \mathbb{R}^{d+1} \to \mathbb{R}^n$ is an onto map. In other words, for every $v \in \mathbb{R}^n$ there exists an $u \in \mathbb{R}^{d+1}$ such that $B'u = v$. Consider $\sigma = (\sigma_1, \ldots, \sigma_n)' \in \{-1, 1\}^n$. So, for this σ, take $\beta^* = (\beta_0, \beta')' \in \mathbb{R}^{d+1}$ such that $B'\beta^* = \sigma$. That is, given n data points, $x_1, \ldots, x_n$,

to pick exactly, $x_{i_1}, \ldots, x_{i_k}$ out of these, set $\sigma_{i_j} = 1$ for all $j = 1, \ldots, k$ and the rest $\sigma_i = -1$. Then for the β^* (depending on σ),

$$(1, x_i')\beta^* = \beta_0 + \beta' x_i = \sigma_i = \begin{cases} 1 \geq 0 & \text{if } i \in \{i_1, \ldots, i_k\} \\ -1 < 0 & \text{otherwise.} \end{cases}.$$

The half space $H^+_{\beta^*}$ contains exactly the points $x_{i_1}, \ldots, x_{i_k}$. And we can do this for all $n \leq d+1$.

(ii) Conversely, for $\mathcal{A}$ to shatter $\{x_1, \ldots, x_n\}$, we should be able to find $\beta^* = (\beta_0, \beta')' \in \mathbb{R}^{d+1}$ such that $B'\beta^*$ lies in each of the 2^n orthants of $\mathbb{R}^n$. So, such 2^n points belong to range(B'). This in turn implies that span(2^n points) $= \mathbb{R}^n \subset$ range(B'). Thus, rank(B') = rank(B) = n. Hence the columns of B are linearly independent.

Since $V_{\mathcal{A}}$ is the maximal number of points which can be shattered by $\mathcal{A}$, and the columns of B can be linearly independent for at most $d+1$ points, $V_{\mathcal{A}} = d+1$. □

Remark 7. The above theorem can be extended to a class of generalized linear decision rules of the form

$$I\Big(\beta_0 + \sum_{j=1}^{d^*} \beta_j \psi_j(x) \geq 0\Big) \quad \text{for} \quad (\beta_0, \ldots, \beta_{d^*}) \in \mathbb{R}^{d^*+1}$$

where $\psi_1, \ldots, \psi_{d^*}$ are linearly independent. Then, the V-C dimension of

$$\mathcal{A} = \left\{ \Big\{ x \in \mathbb{R}^d : \ \beta_0 + \sum_{j=1}^{d^*} \beta_j \psi_j(x) \geq 0 \Big\} : \ \beta_0 \in \mathbb{R}, \beta \in \mathbb{R}^{d^*} \right\}$$

is $d^* + 1$.

The following theorem provides some elementary properties of shatter coefficient of classes obtained by set operations.

Theorem 9.

(a) If $\mathcal{A} = \mathcal{A}_1 \cup \mathcal{A}_2$ *then,* $s(\mathcal{A}, n) \leq s(\mathcal{A}_1, n) + s(\mathcal{A}_2, n)$.

(b) For $\mathcal{A}_c = \{A^c : \ A \in \mathcal{A}\}$, $s(\mathcal{A}_c, n) = s(\mathcal{A}, n)$.

(c) For $\mathcal{A} = \{A_1 \cap A_2 : \ A_1 \in \mathcal{A}_1, A_2 \in \mathcal{A}_2\}$, $s(\mathcal{A}, n) \leq s(\mathcal{A}_1, n)s(\mathcal{A}_2, n)$.

(d) For $\mathcal{A} = \{A_1 \cup A_2 : \ A_1 \in \mathcal{A}_1, A_2 \in \mathcal{A}_2\}$, $s(\mathcal{A}, n) \leq s(\mathcal{A}_1, n)s(\mathcal{A}_2, n)$.

(e) For $\mathcal{A} = \{A_1 \times A_2 : \ A_1 \in \mathcal{A}_1, A_2 \in \mathcal{A}_2\}$, $s(\mathcal{A}, n) \leq s(\mathcal{A}_1, n)s(\mathcal{A}_2, n)$.