São Paulo School of Data Science

Vapnik-Chervonenkis Theory

Ulisses Braga-Neto

ECE Department

Texas A&M University

Vapnik-Chervonenkis Dimension

- The Vapnik–Chervonenkis (VC) dimension is a measure of the size, i.e., the complexity, of a class of classifiers $\mathcal{C}$.
- It agrees very naturally with our intuition of complexity as the ability of a classifier to cut up the space finely.
- Furthermore, we will see it can be used to bound, in a distribution-free manner, the difference between the apparent error and the true error of a classifier in $\mathcal{C}$ (we will see that the simpler $\mathcal{C}$ is, the tighter we can bind this, which again agrees with intuition).

Shatter Coefficients

- Intuitively, the complexity of a classification rule must have to do with its ability to "pick out" subsets of a given set of points.
- For a given n, consider a set of points $x_1, \ldots, x_n$ in R^d. Given a set $A \subseteq R^d$, then

$$A \cap \{x_1, \ldots, x_n\} \subseteq \{x_1, \ldots, x_n\}$$

  is the subset of $\{x_1, \ldots, x_n\}$ "picked out" by A.
- Now consider a family $\mathcal{A}$ of (measurable) sets in R^d, and let

$$N_{\mathcal{A}}(x_1, \ldots, x_n) = |\{A \cap \{x_1, \ldots, x_n\} | \, A \in \mathcal{A}\}|$$

  i.e., the total number of subsets of $\{x_1, \ldots, x_n\}$ that can be picked out by sets in $\mathcal{A}$.

Shatter Coefficients - II

- The n-th *shatter coefficient* of the family $\mathcal{A}$ is defined as

$$s(\mathcal{A}, n) = \max_{\{x_1,\ldots,x_n\}} N_{\mathcal{A}}(x_1, \ldots, x_n)$$

- The shatter coefficients $s(\mathcal{A}, n)$ measure the richness (the size, the complexity) of $\mathcal{A}$.
- Note that $s(\mathcal{A}, n) = k$ if $N_{\mathcal{A}}(x_1, \ldots, x_n) \leq k$ for all sets of n points, *and* at least one instance $\{z_1, \ldots, z_n\}$ can be found such that $N_{\mathcal{A}}(z_1, \ldots, z_n) = k$.
- Note also that $s(\mathcal{A}, n) \leq 2^n$ for all n. (why?)

The VC Dimension

- If $s(\mathcal{A}, n) = 2^n$, then there is a set of points $\{z_1, \ldots, z_n\}$ such that $N_{\mathcal{A}}(z_1, \ldots, z_n) = 2^n$, and we say that $\mathcal{A}$ *shatters* $\{z_1, \ldots, z_n\}$.
- On the other hand, if $s(\mathcal{A}, n) < 2^n$, then any set of points $\{x_1, \ldots, x_n\}$ contains at least one subset that cannot be picked out by any member of $\mathcal{A}$. In addition, we must have $s(\mathcal{A}, m) < 2^m$, for all $m > n$.
- The VC dimension $V_{\mathcal{A}}$ of $\mathcal{A}$ (assuming $|\mathcal{A}| \geq 2$) is the largest integer $k \geq 1$ such that $s(\mathcal{A}, k) = 2^k$. If $s(\mathcal{A}, n) = 2^n$ for all n, then $V_{\mathcal{A}} = \infty$.
- The VC dimension of $\mathcal{A}$ is the *maximal number of points in R^d that can be shattered by $\mathcal{A}$*.
- Clearly, $V_{\mathcal{A}}$ measures the complexity of $\mathcal{A}$!

Some Simple Examples

1. Let $\mathcal{A}$ be the class of half-lines: $\mathcal{A} = \{(-\infty, a] \,|\, a \in R\}$, then
$$s(\mathcal{A}, n) = n + 1 \text{ and } V_{\mathcal{A}} = 1$$
2. Let $\mathcal{A}$ be the class of intervals: $\mathcal{A} = \{[a, b] \,|\, a, b \in R\}$, then
$$s(\mathcal{A}, n) = \frac{n(n+1)}{2} + 1 \text{ and } V_{\mathcal{A}} = 2$$
3. Let $\mathcal{A}_d$ be the class of "half-rectangles" in R^d:
$\mathcal{A}_d = \{(-\infty, a_1] \times \cdots \times (-\infty, a_d] \,|\, (a_1, \ldots, a_d) \in R^d\}$,
then $V_{\mathcal{A}_d} = d$.
4. Let $\mathcal{A}_d$ be the class of rectangles in R^d:
$\mathcal{A}_d = \{[a_1, b_1] \times \cdots \times [a_d, b_d] \,|\, (a_1, \ldots, a_d, b_1, \ldots, b_d]) \in R^{2d}\}$,
then $V_{\mathcal{A}_d} = 2d$.

Some Observations

- In all the examples, the VC dimension is equal to the number of parameters. While this is intuitive, it is *not* true in general. In fact, one can find a one-parameter family $\mathcal{A}$ for which $V_{\mathcal{A}} = \infty$. So be careful about naively attributing complexity to the number of parameters!
- Examples 3 and 4 generalize examples 1 and 2, respectively, by means of cartesian products. It can be shown that in such cases:

$$s(\mathcal{A}_d, n) \leq s(\mathcal{A}, n)^d, \quad \text{for all } n$$

- A general bound for shatter coefficients is

$$s(\mathcal{A}, n) \leq \sum_{i=0}^{V_{\mathcal{A}}} \binom{n}{i}, \quad \text{for all } n$$

Examples 1 and 2 achieve this bound, so it is tight.

Oriented Hyperplanes

- An *oriented hyperplane* is a hyperplane plus a direction that picks one of the two half-spaces created by it.
- Any oriented hyperplane in R^d can be uniquely specified by the associated half-space $\{x \in R^d \,|\, ax \geq b\}$ for some $a \in R^d, b \in R$.
- Oriented hyperplanes correspond to linear classifiers and thus the following result is very important.
- Let $\mathcal{A}_d$ be the class of half-spaces in R^d: $\mathcal{A}_d = \{x \in R^d \,|\, ax \geq b, a \in R^d, b \in R\}$. Then

$$s(\mathcal{A}, n) = 2\sum_{i=0}^{d} \binom{n-1}{i}$$

and $V_{\mathcal{A}_d} = d + 1$.

Example

- For $d = 2$,

$$s(\mathcal{A}_2, 1) = 2\binom{0}{0} = 2 = 2^1$$

$$s(\mathcal{A}_2, 2) = 2\left[\binom{1}{0} + \binom{1}{1}\right] = 4 = 2^2$$

$$s(\mathcal{A}_2, 3) = 2\left[\binom{2}{0} + \binom{2}{1} + \binom{2}{2}\right] = 8 = 2^3$$

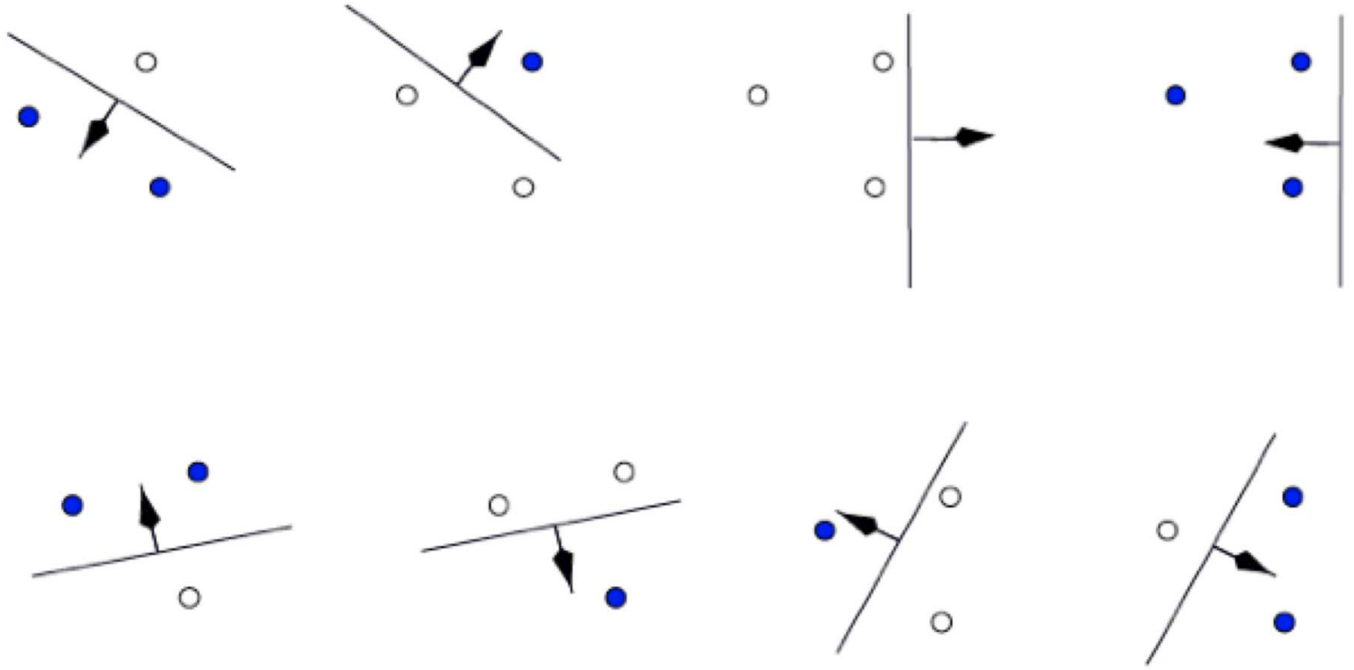
$$s(\mathcal{A}_2, 4) = 2\left[\binom{3}{0} + \binom{3}{1} + \binom{3}{2}\right] = 14 < 16 = 2^4$$

  Therefore, $V_{\mathcal{A}_2} = 3$.

- The above results imply that there is a set of 3 points that can be shattered by oriented hyperplanes in R^2, but no set of 4 points can be shattered — there are at least 2 subsets out of the $2^4 = 16$ that cannot be picked out.

Example - II

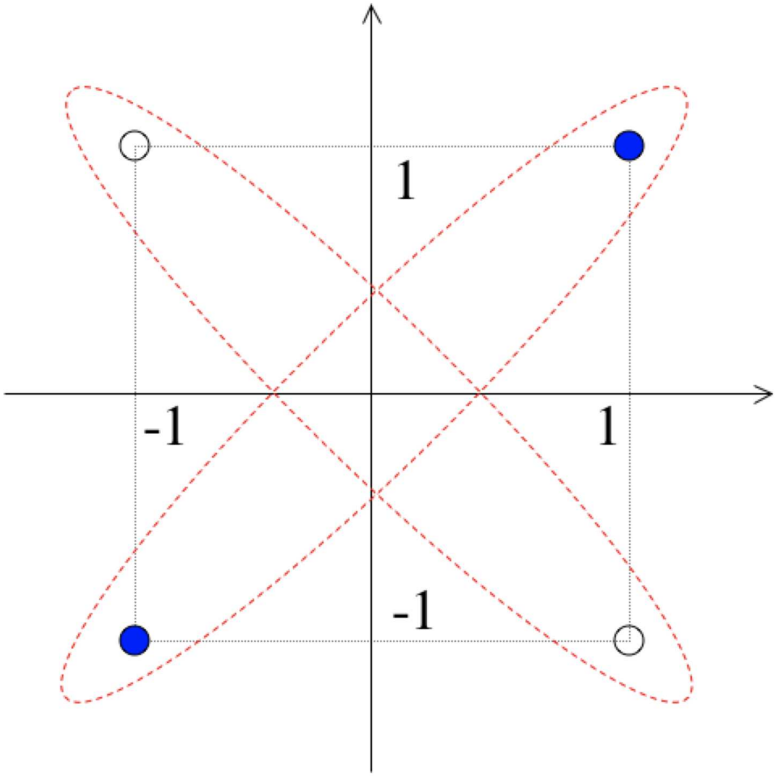
In fact, any set of 3 points (in general position) can be shattered by oriented hyperplanes in R^2.



(figure from Burges' tutorial paper)

Example - III

However, for the familiar XOR problem, each of the 2 indicated subsets of points cannot be picked out by any oriented hyperplane.



VC Theory of Classification

- The preceding concepts can be applied to define the complexity of a class $\mathcal{C}$ of classifiers (i.e., a classification rule).
- Given a classifier $\psi \in \mathcal{C}$, let us define the set $A_\psi = \{x \in R^d \,|\, \psi(x) = 1\}$, that is, the 1-decision region (this specifies the classifier completely, since the 0-decision region is simply A_ψ^c).
- Let $\mathcal{A}_\mathcal{C} = \{A_\psi \,|\, \psi \in \mathcal{C}\}$, that is, the family of all 1-decision regions produced by $\mathcal{C}$. We define the shatter coefficients $\mathcal{S}(\mathcal{C}, n)$ and VC dimension $V_\mathcal{C}$ for $\mathcal{C}$ as

$$\mathcal{S}(\mathcal{C}, n) = s(\mathcal{A}_\mathcal{C}, n)$$

$$V_\mathcal{C} = V_{\mathcal{A}_\mathcal{C}}$$

VC Parameters for Linear Classifiers

- This includes NMC, LDA, Perceptrons, and Linear SVMs.
- The results for oriented hyperplanes apply, therefore

$$\mathcal{S}(\mathcal{C}, n) = 2 \sum_{i=0}^{d} \binom{n-1}{i}$$

  with $V_{\mathcal{C}} = d + 1$.
- The VC dimension thus increases linearly with the number of variables.
- Note: the fact that all linear classification rules have the same VC dimension does not mean they will necessarily perform the same, particularly in the small-sample case.

Distribution-Free VC Bound

- We will see how to use $\mathcal{S}(\mathcal{C}, n)$ and $V_{\mathcal{C}}$ to bound

$$P(|\hat{\epsilon}[\psi] - \epsilon[\psi]| > \tau), \quad \text{for all } \tau > 0$$

for any classifier $\psi \in \mathcal{C}$, and *any* distribution of (X, Y), where $\hat{\epsilon}_n[\psi]$ is the *empirical error*, given data S_n:

$$\hat{\epsilon}_n[\psi] = \frac{1}{n} \sum_{i=1}^{n} |y_i - \psi(x_i)|$$

- If S_n could be assumed independent of every $\psi \in \mathcal{C}$, then $\hat{\epsilon}_n[\psi]$ would be an independent test-set error, and we could use Hoeffding's Inequality (see HW 4) to get

$$P(|\hat{\epsilon}[\psi] - \epsilon[\psi]| > \tau) \leq 2e^{-2n\tau^2}, \quad \text{for all } \tau > 0$$

Distribution-Free VC Bound - II

- *That is not sufficient.* If we want to study $|\hat{\epsilon}[\psi] - \epsilon[\psi]|$ for any distribution, and any classifier $\psi \in \mathcal{C}$, in particular a designed classifier ψ_n, we cannot assume independence from S_n.

- The solution is to bound $P(|\hat{\epsilon}[\psi] - \epsilon[\psi]| > \tau)$ *uniformly* for all possible $\psi \in \mathcal{C}$, that is, to find a (distribution-free) bound for the probability of the worst-case scenario:

$$P\left(\sup_{\psi \in \mathcal{C}} |\hat{\epsilon}[\psi] - \epsilon[\psi]| > \tau\right), \quad \text{for all } \tau > 0$$

Vapnik-Chervonenkis Theorem

- VC Theorem: Regardless of the distribution of (X, Y),

$$P\left(\sup_{\psi \in \mathcal{C}} |\hat{\epsilon}[\psi] - \epsilon[\psi]| > \tau\right) \leq 8\mathcal{S}(\mathcal{C}, n)e^{-n\tau^2/32}, \quad \text{for all } \tau > 0$$

- If $V_{\mathcal{C}}$ is finite, we can use the inequality $\mathcal{S}(\mathcal{C}, n) \leq (n+1)^{V_{\mathcal{C}}}$ to write the bound in terms of $V_{\mathcal{C}}$:

$$P\left(\sup_{\psi \in \mathcal{C}} |\hat{\epsilon}[\psi] - \epsilon[\psi]| > \tau\right) \leq 8(n+1)^{V_{\mathcal{C}}}e^{-n\tau^2/32}, \quad \text{for all } \tau > 0$$

- Therefore, if $V_{\mathcal{C}}$ is finite, the term $e^{-n\tau^2/32}$ dominates, and the bound decreases *exponentially* fast as $n \to \infty$.

No-Free-Lunch Theorem

- The following negative result shows that, independently of how to pick ψ_n from $\mathcal{C}$, the worst-case scenario demands one to have $n \gg V_{\mathcal{C}}$.
- Let $2 < V_{\mathcal{C}} < \infty$, and let Ω be the set of all r.v.'s (X, Y) corresponding to a given $\epsilon_{\mathcal{C}} = \inf_{\psi \in \mathcal{C}} \epsilon[\psi]$. Then for every classification rule associated with $\mathcal{C}$,

$$\sup_{(X,Y)\in\Omega} E[\epsilon_{n,\mathcal{C}} - \epsilon_{\mathcal{C}}] \geq e^{-8} \sqrt{\frac{\epsilon_{\mathcal{C}}(V_{\mathcal{C}} - 1)}{24n}}$$

for $n \geq \frac{V_{\mathcal{C}}-1}{2\epsilon_{\mathcal{C}}} \max\{9, 1/(1-2\epsilon_{\mathcal{C}})^2\}$.

No-Free-Lunch with Infinite VC

- If $V_{\mathcal{C}} = \infty$, the lower bound from the previous slide does not hold. We get instead a worse result.
- Here, we cannot make $n \gg V_{\mathcal{C}}$, and a worst-case bound can be found that is independent of n (this means that there exists a situation where the design error cannot be reduced no matter how large n may be). This is shown by the following result.
- If $V_{\mathcal{C}} = \infty$, then for every $\delta > 0$, and every classification rule associated with $\mathcal{C}$, there is a distribution for (X, Y) with $\epsilon_{\mathcal{C}} = 0$ but

$$E[\epsilon_{n,\mathcal{C}} - \epsilon_{\mathcal{C}}] = E[\epsilon_{n,\mathcal{C}}] > \frac{1}{2e} - \delta, \quad \text{for all } n > 1$$