Stat 6390.001: High Dimensional Probability and Statistics, Fall 2026 Topic (October 14)

VC Dimension and Covering Numbers

Lecturer: Yun Wei *Scribe: Carolyne Chepkemboi*

Disclaimer: *These notes have not been subjected to the usual scrutiny reserved for formal publications. They may be distributed outside this class only with the permission of the instructor.*

1 Vapnik–Chervonenkis (VC) Dimension

Definition 1.1 (Shattering and VC Dimension)**.** Let $\mathcal{F}$ be a class of binary-valued functions $f : \mathcal{X} \to \{0, 1\}$.
A collection of points $x_1^n = (x_1, \ldots, x_n)$ is said to be *shattered* by $\mathcal{F}$ if $(\mathcal{F}(x_1^n)) = 2^n$.
That is, every possible binary labeling of the n points can be realized by some function in $\mathcal{F}$. The VC dimension $\nu(\mathcal{F})$, is the largest integer n for which there exists a collection of n points shattered by $\mathcal{F}$.

The VC dimension of a set class is denoted by $\nu(\mathcal{S})$.

Proposition 1.2. *Let $\mathcal{G}$ be a vector space of functions $g : \mathbb{R}^d \to \mathbb{R}$, with $\dim(\mathcal{G}) < \infty$. Then the subgraph class $\mathcal{S}(\mathcal{G})$ has VC dimension at most $\dim(\mathcal{G})$:*

$$\nu(\mathcal{S}(\mathcal{G})) \leq \dim(\mathcal{G}).$$

To illustrate this proposition, consider the following examples:

Example 1.3 (Linear Functions in $\mathbb{R}^d$)**.** Consider the function

$$f_{a,b}(x) = \langle a, x\rangle + b, \qquad a \in \mathbb{R}^d, \quad b \in \mathbb{R}.$$

Define the family of all such linear functions:

$$\mathcal{L}^d = \{f_{a,b}(\cdot)|(a, b) \in \mathbb{R}^d \times \mathbb{R}\}.$$

The class $\mathcal{L}^d$ forms a vector space since it is closed under addition and scalar multiplication.
Let e_i denote the ith standard basis vector in $\mathbb{R}^d$. Then

$$f_{e_i,0}(x) = x_i, \qquad f_{0,1}(x) = 1.$$

Every function in $\mathcal{L}^d$ can be expressed as $f_{a,b}(x) = \sum_{i=1}^d a_i x_i + b$.
The functions $\{f_{e_1,0}, \ldots, f_{e_d,0}, f_{0,1}\}$ form a basis for $\mathcal{L}^d$, giving $\dim(\mathcal{L}^d) = d + 1$.
For each $f_{a,b}$, the corresponding subgraph set is

$$H_{a,b} = S_{f_{a,b}} = \{x \in \mathbb{R}^d|\langle a, x\rangle + b \leq 0\}.$$

These sets describe half-spaces in $\mathbb{R}^d$ when $a \neq 0$.
The associated subgraph class is

$$\mathcal{S}(\mathcal{L}^d) = \{H_{a,b}|(a, b) \in \mathbb{R}^d \times \mathbb{R}\}.$$

Since $\mathcal{L}^d$ is a vector space of dimension $d + 1$, the proposition gives

$$\boxed{\nu(\mathcal{S}(\mathcal{L}^d)) \leq d + 1.}$$

To illustrate this result, **consider the special case $d = 1$.**
Here,

$$f_{a,b}(x) = ax + b,$$

and

$$H_{a,b} = \{x \in \mathbb{R} | ax + b \leq 0\}.$$

For $a > 0$,

$$H_{a,b} = \{x \in \mathbb{R} | x \leq \frac{-b}{a}\} = \left(-\infty, -\frac{b}{a}\right],$$

and for $a < 0$,

$$H_{a,b} = \{x \in \mathbb{R} | x \geq \frac{-b}{a}\} = \left[-\frac{b}{a}, \infty\right).$$

Therefore, apart from the empty set and $\mathbb{R}$, the class consists of left-sided and right-sided intervals:

$$\mathcal{S}(\mathcal{L}^1) = \{(-\infty, t] | t \in \mathbb{R}\} \cup \{[s, \infty) | s \in \mathbb{R}\} \cup \{\varnothing, \mathbb{R}\}.$$

Next, consider two distinct points $x_1 < x_2$. Define the subsets:

1. If $t < x_1$, $(-\infty, t]$, then
$$(-\infty, t] \cap \{x_1, x_2\} = \varnothing.$$
2. If $x_1 < t < x_2$,$(-\infty, t]$, then
$$(-\infty, t] \cap \{x_1, x_2\} = \{x_1\}.$$
3. If $t > x_2$, $(-\infty, t]$, then
$$(-\infty, t] \cap \{x_1, x_2\} = \{x_1, x_2\}.$$
4. If $x_1 < s < x_2$, $[s, \infty)$, then
$$[s, \infty) \cap \{x_1, x_2\} = \{x_2\}.$$

All four subsets $\varnothing$, $\{x_1\}$, $\{x_2\}$, $\{x_1, x_2\}$ can be obtained using sets from $\mathcal{S}(\mathcal{L}^1)$.
Hence, two points can be shattered.
Next, **consider three distinct points**

$$x_1 < x_2 < x_3.$$

Any left-sided interval containing x_2 must also contain x_1. Similarly, any right-sided interval containing x_2 must contain x_3. Thus, the subset $\{x_2\}$ cannot be selected without including another point.

Consequently, three points cannot be shattered, and we conclude that the largest n that can be shattered is $n = 2$:

$$\boxed{\nu(\mathcal{S}(\mathcal{L}^1)) = 2.}$$

This agrees with the upper bound obtained from the proposition.
Generally, the VC dimension of affine half-spaces in $\mathbb{R}^d$ is

$$\boxed{\nu(\mathcal{S}(\mathcal{L}^d)) = d + 1.}$$

Example 1.4 (Spheres in $\mathbb{R}^d$)**.** Consider the collection of Euclidean balls

$$B_{a,b} = \{x \in \mathbb{R}^d | \|x - a\|_2 \leq b\},$$

where $a \in \mathbb{R}^d$ is the center and $b \in \mathbb{R}_+$ is the radius.

Define the function

$$f_{a,b}(x) = \|x - a\|_2^2 - b^2.$$

Expanding the squared Euclidean norm gives

$$\begin{aligned} f_{a,b}(x) &= \langle x - a, x - a \rangle - b^2 \\ &= \|x\|_2^2 - 2\langle a, x \rangle + \|a\|_2^2 - b^2 \\ &= \|x\|_2^2 - 2\sum_{i=1}^{d} a_i x_i + \|a\|_2^2 - b^2. \end{aligned}$$

Since $\|x - a\|_2 \leq b$ is equivalent to $f_{a,b}(x) \leq 0$, we have

$$B_{a,b} = S_{f_{a,b}} = \{x \in \mathbb{R}^d | f_{a,b}(x) \leq 0\}.$$

Thus, each Euclidean ball is a subgraph set associated with the function $f_{a,b}$.
However, the function class

$$\mathcal{F} = \{f_{a,b}(\cdot) | (a, b) \in \mathbb{R}^d \times \mathbb{R}_+\}$$

is not a vector space, since it is not closed under addition or scalar multiplication.

To apply the preceding proposition, we construct a larger vector space containing $\mathcal{F}$.
Define the feature map

$$\phi(x) = (1, x_1, \ldots, x_d, \|x\|_2^2) \in \mathbb{R}^{d+2}.$$

For any $c \in \mathbb{R}^{d+2}$, define

$$g_c(x) = \langle c, \phi(x) \rangle.$$

Expanding the inner product,

$$g_c(x) = c_1 + \sum_{i=2}^{d+1} c_i x_{i-1} + c_{d+2}\|x\|_2^2.$$

Now consider the function class

$$\mathcal{G} = \{g_c(\cdot) | c \in \mathbb{R}^{d+2}\}.$$

The class $\mathcal{G}$ is a vector space because it is closed under addition and scalar multiplication:

$$g_{c_1} + g_{c_2} = g_{c_1 + c_2},$$

and

$$\lambda g_c = g_{\lambda c}.$$

The functions

$$1, x_1, \ldots, x_d, \|x\|_2^2$$

form a basis for $\mathcal{G}$, so

$$\dim(\mathcal{G}) = d + 2.$$

Moreover, every $f_{a,b}$ can be expressed as g_c by choosing

$$c = (\|a\|_2^2 - b^2, -2a_1, \ldots, -2a_d, 1).$$

Therefore,

$$\mathcal{F} \subseteq \mathcal{G}.$$

Applying the proposition gives

$$\nu(\mathcal{S}(\mathcal{G})) \leq \dim(\mathcal{G}) = d + 2.$$

Let

$$\mathcal{B}^d = \{B_{a,b} | (a,b) \in \mathbb{R}^d \times \mathbb{R}_+\}.$$

Since $\mathcal{F} \subseteq \mathcal{G}$, the corresponding subgraph classes satisfy

$$\mathcal{B}^d = \mathcal{S}(\mathcal{F}) \subseteq \mathcal{S}(\mathcal{G}).$$

Hence,

$$\begin{aligned} \nu(\mathcal{B}^d) &= \nu(\mathcal{S}(\mathcal{F})) \\ &\leq \nu(\mathcal{S}(\mathcal{G})) \\ &\leq d + 2. \end{aligned}$$

Therefore,

$$\boxed{\nu(\mathcal{B}^d) \leq d + 2.}$$

This upper bound is not sharp, but it is sufficient for many applications. A more refined argument shows that the VC dimension of Euclidean balls in $\mathbb{R}^d$ is

$$\boxed{\nu(\mathcal{B}^d) = d + 1.}$$

The case $d = 2$ is considered in Exercise 4.13 of Wainwright (2019) **(part of the next homework)**.

2 A Remark on Chapter 4: Empirical Processes

Recall that

$$\|P_n - P\|_{\mathcal{F}} = \sup_{f \in \mathcal{F}} |P_n f - Pf|.$$

In general, this supremum need not be measurable when $\mathcal{F}$ is uncountable. If $\mathcal{F}$ is separable, the supremum can be taken over a countable dense subset. Otherwise, one may use outer probability, as discussed by van der Vaart and Wellner (1996).

For any finite class $\mathcal{G} \subseteq \mathcal{F}$, the random variable $\|P_n - P\|_{\mathcal{G}}$ is well defined, so it is sensible to define

$$\mathbb{E}\left[\phi\left(\|P_n - P\|_{\mathcal{F}}\right)\right] := \sup\left\{\mathbb{E}\left[\phi\left(\|P_n - P\|_{\mathcal{G}}\right)\right] \,|\, \mathcal{G} \subset \mathcal{F},\ \mathcal{G} \text{ has finite cardinality}\right\}.$$

Using this definition, we can work with expectations expressed as suprema over finite sets.

3 Chapter 5: Metric Entropy and its uses

3.1 Covering Numbers

Lemma 3.1. *Let T be a subset of a complete metric space. Then T is precompact (that is, its closure is compact) if and only if*

$$N(\delta, T, \rho) < \infty \qquad \text{for every } \delta > 0.$$

The covering number $N(\delta, T, \rho)$ is the smallest number of balls of radius δ needed to cover T. $N(\delta, T, \rho)$ is a quantitative way of measuring "how compact a set is".

As the radius $\delta \to 0^+$, the number of balls required to cover T is non-decreasing and typically diverges.

The logarithm of the covering number,

$$\boxed{\log N(\delta, T, \rho),}$$

is called the *metric entropy.*

References

[1] Martin J. Wainwright, *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*, Cambridge University Press, 2019.