Statistics 210B Lecture 10 Notes

Daniel Raban

February 17, 2022

1 VC Dimension, Covering, and Packing

1.1 VC dimension

Last time we were discussing function classes with polynomial discrimination. Recall that a function class $\mathcal{F}$ has PD(ν) if for all n and $X_{1:n}$, $|\mathcal{F}(X_{1:n})| \leq (n+1)^\nu$. If $\mathcal{F}$ has PD(ν), then $\mathcal{R}_n(\mathcal{F}) \leq D\sqrt{\frac{\nu \log(n+1)}{n}}$. This gives the bound $\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}} \lesssim D\sqrt{\frac{\nu \log(n+1)}{n}}$.

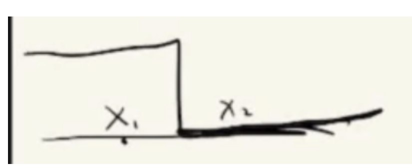
What function classes have polynomial discrimination? This question is answered by VC theory, named for Vapnik and Chervonenkis. If a function class has "VC dimenion ν," then $\mathcal{F}$ has PD(ν), which means that $\mathcal{R}_n(\mathcal{F}) \leq D\sqrt{\frac{\nu \log(n+1)}{n}}$.

Definition 1.1. Suppose $\mathcal{F} \subseteq \{F : \mathcal{X} \to \{0,1\}\}$ is binary valued. We say that $x_{1:n}$ is **shattered** by $\mathcal{F}$ if $|\mathcal{F}(x_{1:n})| = 2^n$. The **VC dimension**, $\nu(\mathcal{F})$, is the largest n such that there exists $x_{1:n}$ shattered by $\mathcal{F}$.

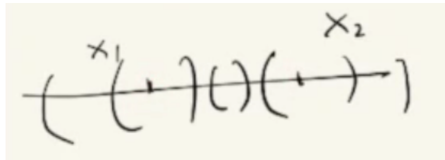
Note that $|\mathcal{F}(X_{1:n})| \leq 2^n$ always. So we want $\mathcal{F}$ to be able to distinguish between points in a maximal sense.

Example 1.1. Let $\mathcal{F} = \{\mathbb{1}_{\{x \leq t\}} : t \in \mathbb{R}\}$. We claim that $\nu(\mathcal{F}) = 1$. Recall that $\mathcal{R}_n(\mathcal{F}) \leq 4\sqrt{\frac{\log(n+1)}{n}}$; this will also be implied by the VC-dimension. We have to show that there is some x_1 that is shattered by $\mathcal{F}$, and we have to show that no x_1, x_2 can be shattered by $\mathcal{F}$.

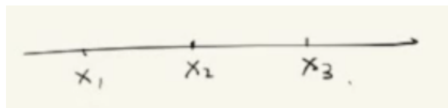
For $n = 1$, $\mathcal{F}(\{x_1\}) = \{0,1\}$, so $\{x_1\}$ is shattered by $\mathcal{F}$. For $n = 2$, we want to show thta $\mathcal{F}(\{x_1, x_2\}) \leq 2^2 - 1$. If we assume, without loss of generality, that $x_2 > x_1$, this is because $\mathcal{F}(\{x_1, x_2\}) = \{(0,0), (1,1), (1,0)\}$. Why does this not contain $(0,1)$? This is because if one of these indicators gives 1 to x_2, then it must give 1 to x_1.



Example 1.2. Let $\mathcal{F} = \{\mathbb{1}_{\{s \le x \le t\}} : s < t \in \mathbb{R}\}$. We claim that $\nu(\mathcal{F}) = 2$. When $n = 2$, we want to find x_1, x_2 such that $|\mathcal{F}((x_1, x_2))| = 2^2$. Here is how we can construct intervals to shatter a two point set:



Now suppose $x_1 < x_2 < x_3$. Then we cannot have $(1, 0, 1)$, since if an interval contains x_1, x_3 then it must contain x_2



Here is an example we will not prove.

Example 1.3. Let $\phi_1, \dots, \phi_p : \mathcal{X} \to \mathbb{R}$ be linear (which you can think of as feature maps), and consider $\mathcal{F} = \{\mathbb{1}_{\{\sum_{i=1}^p a_i \phi_i(x) \le b\}} : a_i, b \in \mathbb{R}\}$. Then $\nu(\mathcal{F}) \le p + 1$.

By definition, for all $n > \nu(\mathcal{F})$,

$$\sup_{x_{1:n}} |\mathcal{F}(x_{1:n})| \le 2^n - 1.$$

Proposition 1.1 (Vapnik-Chervonenkis, Sauer-Shelah[1])**.** *For $\mathcal{F}$ with VC dimension ν,*

$$\sup_{x_{1:n}} |\mathcal{F}(x_{1:n})| \le \sum_{i=1}^{\nu} \binom{n}{i} \le \min\left\{(n+1)^\nu, \left(\frac{ne}{\nu}\right)^\nu\right\}.$$

By this proposition, we immediately have

$$\mathcal{R}_n(\mathcal{F}) \le D\sqrt{\frac{\nu \log(n+1)}{n}}.$$

Her is an end-to-end result: If $\mathcal{F} = \{\mathbb{1}_{\{\sum_{i=1}^p a_i \phi_i(x) \le b\}} : a_i, b \in \mathbb{R}\}$ and $(X_i)_{i \in [n]} \overset{\text{iid}}{\sim} \mathbb{P}$, then

$$\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n f(X_i) - \mathbb{E}[f(X)] \right| \lesssim \sqrt{\frac{(p+1)\log n}{n}}.$$

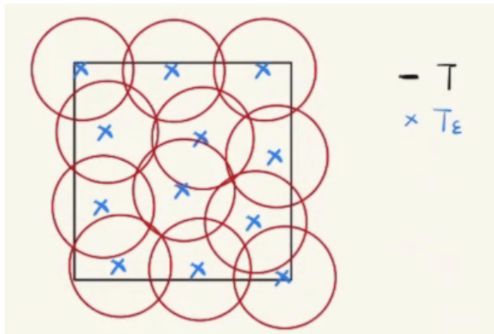
This $\log n$ factor can be eliminated later by the *chaining method*.

The proof of this proposition is a combinatorial argument; since the argument will not show up again, we will omit the proof, but you can look at the proof in the textbook.

[1] This proposition was proven independently by Vapnik and Chervonenkis in 1971, by Sauer in 1972, and by Shelah by 1972.

1.2 The metric entropy method

Given a sub-Gaussian X_θ for all $\theta \in T$, we hope to upper bound $\mathbb{E}[\sup_{\theta \in T} X_\theta]$. How do we do this when $|T| = \infty$? The idea is to approximate T by a finite set T_ε as follows:



This gives

$$\mathbb{E}\left[\sup_{\theta \in T} X_\theta\right] \leq \mathbb{E}\left[\sup_{\widetilde{\theta} \in T_\varepsilon} X_{\widetilde{\theta}}\right] + \mathbb{E}\left[\sup_{\theta \in T, \widetilde{\theta} \in T_\varepsilon} (X_\theta - X_{\widetilde{\theta}})\right].$$

We hope that

1. $|T_\varepsilon|$ is small.
2. $\mathbb{E}[\sup_{\theta \in T, \widetilde{\theta} \in T_\varepsilon} (X_\theta - X_{\widetilde{\theta}})]$ is small.

Given T and ρ, how can we find T_ε and bound $|T_\varepsilon|$?

1.3 Covering and packing

Definition 1.2. A **metric space** is a pair (T, ρ), where $\rho : T \times T \to \mathbb{R}$ such that

1. $\rho(\theta, \theta') \geq 0$ for all $\theta, \theta' \in T$, with equality holding iff $\theta = \theta'$.
2. $\rho(\theta, \theta') = \rho(\theta', \theta)$.
3. $\rho(\theta, \theta') \leq \rho(\theta, \theta'') + \rho(\theta'', \theta')$.

Example 1.4. If $T = \mathbb{R}^d$, here are a few useful metrics:

$$\rho(\theta, \theta') = \|\theta - \theta'\|_2, \qquad \rho(\theta, \theta') = \frac{1}{d} \sum_{i=1}^{d} \mathbb{1}_{\{\theta_i - \theta_i'\}}$$

The set T can be a function space, rather than a parameter space.

Example 1.5. Let $T = L^2(\mathcal{X}, \mu)$. Here are two metrics on T:

$$\rho(f,g) = \left(\int (f(x) - g(x))^2 \, d\mu(x) \right)^{1/2}, \qquad \rho(f,g) = \|f - g\|_\infty.$$

Definition 1.3. $T_\varepsilon = \{\theta^1, \dots, \theta^N\}$ is an ε**-covering** of a set T if for all $\theta \in T$, there exists a $\theta^i \in T_\varepsilon$ such that $\rho(\theta, \theta^i) \leq \varepsilon$. The ε**-covering number** of T with respect to ρ is defined as

$$N(\varepsilon, T.\rho) := \inf\{N : |T_\varepsilon| = N, T_\varepsilon \text{ is an } \varepsilon\text{-covering of } T.$$

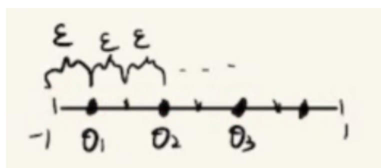
The maximal inequality gives

$$\mathbb{E}\left[\max_{\theta \in T_\varepsilon} X_\theta \right] \lesssim \sqrt{\log |T_\varepsilon|} \approx \sqrt{\log N(\varepsilon; T, \rho)}.$$

Definition 1.4. The function $\varepsilon \mapsto \log N(\varepsilon; T, \rho)$ for fixed (T, ρ) is caleld the **metric entropy of the set** T.

We will see examples that range from parametric families with $\log N(\varepsilon) \approx d \log(1 + 1/\varepsilon)$ to nonparametric families with $\log N(\varepsilon \approx (1/\varepsilon)^\alpha$, where $\alpha \geq 0$.

Example 1.6. Let $T = [-1, 1]$ with $\rho(\theta, \theta') = |\theta - \theta'|$. Then $N(\varepsilon; T, \rho \leq \frac{1}{\varepsilon} + 1$.



Example 1.7. If $T = [-1, 1]^d$ with $\rho(\theta, \theta') = \|\theta - \theta'\|_\infty$, then $N(\varepsilon; T, \rho) \leq (\frac{1}{\varepsilon} + 1)^d$.

Up to some constant, this bound is tight.

How about with other metrics? We may not be able to figure out a cover/packing. We can take a volume approach: We should expect

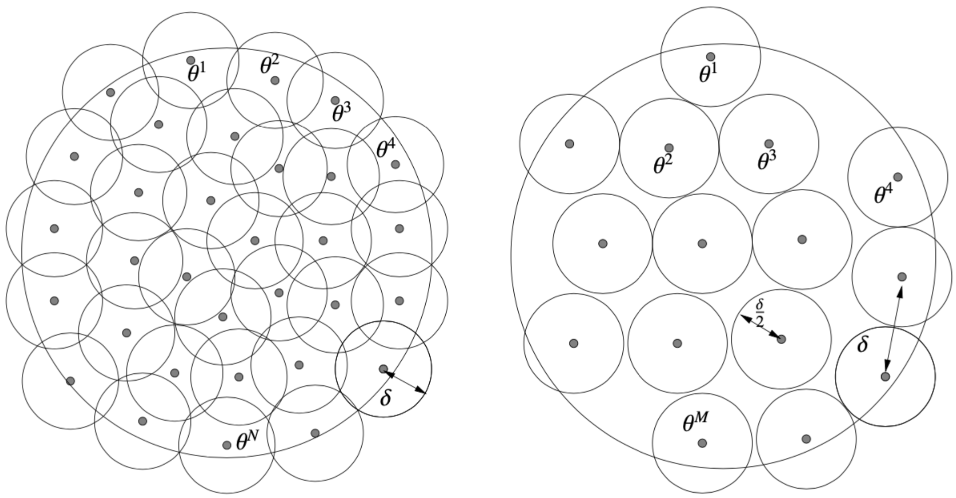
$$\log N(\varepsilon; T, \rho) \approx \log \left(\frac{\mathrm{Vol}(T)}{\mathrm{Vol}(B_\rho(\varepsilon))} \right).$$

To make this statement precise, we can introduce the idea of packing:

Definition 1.5. A set $\widetilde{T}_\varepsilon = \{\theta^1, \ldots, \theta^M\} \subseteq T$ is an ε-**packing** if for all $\theta^i, \theta^j \in \widetilde{T}_\varepsilon$ with $i \neq j$, $\rho(\theta^i, \theta^j) > \varepsilon$. The ε-**packing number** is

$$M(\varepsilon; T, \rho) = \sup\{M : |\widetilde{T}_\varepsilon| = M, \widetilde{T}_\varepsilon \text{ is an } \varepsilon\text{-packing of } T\}.$$

This means that $B_\rho(\theta^i, \varepsilon/2) \cap B_\rho(\theta^j, \varepsilon/2) = \varnothing$. Here is a picture from Wainwright's textbook comparing packings and coverings:

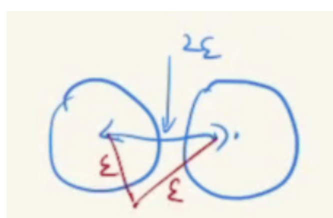


Lemma 1.1. *For all $\varepsilon > 0$, we have*

$$M(2\varepsilon; T, \rho) \leq N(\varepsilon; T, \rho) \leq M\varepsilon; T, \rho).$$

Proof. A maximal ε-packing gives an ε-covering. Suppose we have a maximal packing; then we cannot put another point into the packing, so the entire set T must be covered by the balls determined by the packing.

For a 2ε-packing with size M, all ε-coverings should have size at least M.



Otherwise, we would have a contradiction. $\square$