Theoretical Statistics. Lecture 11.

Peter Bartlett

Uniform laws of large numbers: Bounding Rademacher complexity.

1. Vapnik-Chervonenkis dimension.
2. Structural results for Rademacher complexity.
3. Metric entropy.

Recall: Uniform laws and Rademacher complexity

Theorem: For $F \subset [0,1]^{\mathcal{X}}$,

$$\frac{1}{2}\mathbf{E}\|R_n\|_F - \sqrt{\frac{\log 2}{2n}} \leq \mathbf{E}\|P - P_n\|_F \leq 2\mathbf{E}\|R_n\|_F,$$

and, with probability at least $1 - 2\exp(-2\epsilon^2 n)$,

$$\mathbf{E}\|P - P_n\|_F - \epsilon \leq \|P - P_n\|_F \leq \mathbf{E}\|P - P_n\|_F + \epsilon.$$

Thus, $\mathbf{E}\|R_n\|_F \to 0$ iff $\|P - P_n\|_F \stackrel{as}{\to} 0$.

Recall: Growth function

Definition: For a class $F \subseteq \{0,1\}^{\mathcal{X}}$, the **growth function** is

$$\Pi_F(n) = \max\{|F(x_1^n)| : x_1, \ldots, x_n \in \mathcal{X}\}.$$

$\mathbf{E}\|R_n\|_F \leq \sqrt{\frac{2\log(2\Pi_F(n))}{n}}$. Notice that $\log \Pi_F(n) = o(n)$ implies $\mathbf{E}\|R_n\|_F \to 0$.

Recall: Vapnik-Chervonenkis dimension

Definition: A class $F \subseteq \{0,1\}^{\mathcal{X}}$ **shatters** $\{x_1, \ldots, x_d\} \subseteq \mathcal{X}$ means that $|F(x_1^d)| = 2^d$.

The Vapnik-Chervonenkis dimension of F is

$$\begin{aligned} d_{VC}(F) &= \max\{d : \text{some } x_1, \ldots, x_d \in \mathcal{X} \text{ is shattered by } F\} \\ &= \max\left\{d : \Pi_F(d) = 2^d\right\}. \end{aligned}$$

Recall: "Sauer's Lemma"

Theorem: [Vapnik-Chervonenkis] $d_{VC}(F) \leq d$ implies

$$\Pi_F(n) \leq \sum_{i=0}^{d} \binom{n}{i}.$$

If $n \geq d$, the latter sum is no more than $\left(\frac{en}{d}\right)^d$.

$$\Pi_F(n) \begin{cases} = 2^n & \text{if } n \leq d, \\ \leq (e/d)^d \, n^d & \text{if } n > d. \end{cases}$$

VC-dimension bounds for parameterized families

Consider a parameterized class of binary-valued functions,

$$F = \{x \mapsto f(x,\theta) : \theta \in \mathbb{R}^p\},$$

where $f : \mathbb{R}^m \times \mathbb{R}^p \to \{\pm 1\}$.

Suppose that f can be computed using no more than t operations of the following kinds:

1. arithmetic ($+$, $-$, $\times$, $/$),
2. comparisons ($>$, $=$, $<$),
3. output ± 1.

Theorem: $d_{VC}(F) \leq 4p(t+2)$.

VC-dimension bounds for parameterized families

Proof idea:

Any f of this kind can be expressed as $f(x,\theta) = h(\mathrm{sign}(g_1(x,\theta)),\ldots,\mathrm{sign}(g_k(x,\theta)))$ for functions g_i that are polynomial in θ, and some boolean function h. (Notice that $k \le 2^t$, and the degree of any polynomial g_i is no more than 2^t.) Notice that a change of the value of f must be due to a change of the sign of one of the g_i. Hence, $\Pi_F(n) \le$ number of connected components in $\mathbb{R}^d$ after the sets $g_i(x_j) = 0$ are removed. We won't go through the proof of this (it can be found in *Neural Network Learning: Theoretical Foundations*). It is rather similar to the case of linear threshold functions, which we'll look at next.

VC-dimension bounds for linear threshold functions

Consider $f(x, \theta) = \text{sign}(w^T x - w_0)$, where $x \in \mathbb{R}^d$ and $\theta = (w^T, w_0)$. Then f can only change value on some $x_1, \ldots, x_n$ for θ such that $w^T - w_0 = 0$. Then (provided these zero sets satisfy some genericity condition), $|F(x_1^n)| = C(n, d+1)$, where $C(n, d+1)$ is the number of cells created in $\mathbb{R}^{d+1}$ when n hyperplanes are removed.

Inductive argument: $C(1, d) = 2$. And $C(n+1, d) = C(n, d) + C(n, d-1)$. To see this, notice that when we have n planes in $\mathbb{R}^p$, and we add a plane, the number of cells that we split in two is precisely the number of cells in the $d-1$-subspace of the new plane that the first n planes leave. Then an inductive argument shows that

$$\Pi_F(n) = C(n, d+1) = 2 \sum_{i=0}^{d} \binom{n-1}{i}. \qquad \text{[Schaffli, 1851.]}$$

Rademacher complexity: structural results

1. $F \subseteq G$ implies $\|R_n\|_F \leq \|R_n\|_G$.
2. $\|R_n\|_{cF} = |c|\|R_n\|_F$.
3. For $|g(X)| \leq 1$, $|\mathbf{E}\|R_n\|_{F+g} - \mathbf{E}\|R_n\|_F| \leq \sqrt{2\log 2/n}$.
4. $\|R_n\|_{\mathrm{co}\, F} = \|R_n\|_F$, where $\mathrm{co}\, F$ is the convex hull of F.
5. If $\phi : \mathcal{X} \times \mathbb{R}$ has $y \mapsto \phi(x, y)$ 1-Lipschitz for all x and $\phi(x, 0) = 0$, then for $\phi(F) = \{x \mapsto \phi(x, f(x))\}$, $\mathbf{E}\|R_n\|_{\phi(F)} \leq 2\mathbf{E}\|R_n\|_F$.

Rademacher complexity: structural results

Proofs:

(1) and (2) are immediate. For (3):

$$\|R_n\|_{F+g} = \sup_{f\in F}\left|\frac{1}{n}\sum_i \epsilon_i\left(f(X_i)+g(X_i)\right)\right|,$$

$$\text{so} \qquad \left|\mathbf{E}\|R_n\|_{F+g} - \mathbf{E}\|R_n\|_F\right| \le \mathbf{E}\left|R_n(g)\right| \le \sqrt{\frac{2\log 2}{n}}$$

for $|g(X)| \le 1$.

(4) follows from the fact that a linear criterion in a convex set is maximized at an extreme point.

(5) is a result due to Ledoux and Talagrand. See website for a link to a proof.

Covering and packing numbers

Definition: A **pseudometric space** (S, d) is a set S and a function $d : S \times S \to [0, \infty)$ satisfying

1. $d(x, x) = 0$,
2. $d(x, y) = d(y, x)$,
3. $d(x, z) \leq d(x, y) + d(y, z)$.

Examples:

1. Metric spaces like $(\mathbb{R}^d, \| \cdot \|_2)$.
2. A set F of functions with pseudometric $d(f, g) = \frac{1}{n} \sum_{i=1}^{n} |f(x_i) - g(x_i)|$.

Covering numbers

Definition: An ϵ-cover of a subset T of a pseudometric space (S, d) is a set $\hat{T} \subset T$ such that for each $t \in T$ there is a $\hat{t} \in \hat{T}$ such that $d(t, \hat{t}) \leq \epsilon$. The ϵ-covering number of T is

$$N(\epsilon, T, d) = \min\{|\hat{T}| : \hat{T} \text{ is an } \epsilon\text{-cover of } T\}.$$

A set T is **totally bounded** if, for all $\epsilon > 0$, $N(\epsilon, T, d) < \infty$.
The function $\epsilon \mapsto \log N(\epsilon, T, d)$ is the **metric entropy** of T.
If $\lim_{\epsilon \to 0} \log N(\epsilon)/\log(1/\epsilon)$ exists, it is called the **metric dimension**.

[PICTURE]

Intuition: A d-dimensional set has metric dimension d. ($N(\epsilon) = \Theta(1/\epsilon^d)$.)

Covering numbers

Example: $([0,1]^d, l_\infty)$ has $N(\epsilon) = \Theta(1/\epsilon^d)$.