CS281B/Stat241B. Statistical Learning Theory. Lecture 8.

Peter Bartlett

Uniform laws of large numbers.

1. Recall: Rademacher complexity.
   (a) Concentration: whp, $\|P - P_n\|_F \leq \mathbf{E}\|P - P_n\|_F + \epsilon$.
   (b) Symmetrization: $\mathbf{E}\|P - P_n\|_F \leq 2\mathbf{E}\|R_n\|_F$.
   (c) Control $\mathbf{E}\|R_n\|_F$.
2. Bounding Rademacher complexity:
   (a) Structural results.
   (b) Growth function.
   (c) Vapnik-Chervonenkis dimension, Sauer's lemma.

Recall: Uniform laws and Rademacher complexity

Theorem: For $F \subset [0,1]^{\mathcal{X}}$, with probability at least $1 - 2\exp(-2\epsilon^2 n)$,

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

and

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

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

Rademacher complexity: structural results

Theorem:

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 $\phi : \mathbb{R} \times \mathcal{Z}$ has $\alpha \mapsto \phi(\alpha, z)$ 1-Lipschitz for all z and $\phi(0, z) = 0$, then for $\phi(F) = \{z \mapsto \phi(f(z), z)\}$, $\mathbf{E}\|R_n\|_{\phi(F)} \leq 2\mathbf{E}\|R_n\|_F$.

Rademacher complexity: structural results

Proofs:

Recall: Uniform laws and Rademacher complexity

Lemma: **[Finite Class Lemma]** For $f \in F$ satisfying $|f(x)| \leq 1$,

$$\mathbf{E}\|R_n\|_F \leq \mathbf{E}\sqrt{\frac{2\log(|F(X_1^n) \cup -F(X_1^n)|)}{n}} \leq \sqrt{\frac{2\log(2\Pi_F(n))}{n}}.$$

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

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

- $\Pi_F(n) \leq |F|$, $\lim_{n\to\infty} \Pi_F(n) = |F|$.
- $\Pi_F(n) \leq 2^n$. (But then this gives no useful bound on $\mathbf{E}\|R_n\|_F$.)
- $\log \Pi_F(n) = o(n)$ implies $\mathbf{E}\|R_n\|_F \to 0$.

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

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

Vapnik-Chervonenkis dimension: “Sauer’s Lemma”

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

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

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

So the VC-dimension is a single integer summary of the growth function: either it is finite, and $\Pi_F(n) = O(n^d)$, or $\Pi_F(n) = 2^n$. No other growth is possible.

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

Vapnik-Chervonenkis dimension: "Sauer's Lemma"

Thus, for $d_{VC}(F) \leq d$ and $n \geq d$, we have

$$\mathbf{E}\|R_n\|_F \leq \sqrt{\frac{2\log(2\Pi_F(n))}{n}} \leq \sqrt{\frac{2\log 2 + 2d\log(en/d)}{n}}.$$

Vapnik-Chervonenkis dimension: Examples

e.g.: $F = \{x \mapsto 1[x \leq \alpha] : \alpha \in \mathbb{R}\}$.
$d_{VC}(F) = 1$.

e.g.: $F = \{x \mapsto 1[x$ below and to left of $y] : y \in \mathbb{R}^2\}$.
$d_{VC}(F) = 2$. [PICTURE]

e.g.: $F = \{x \mapsto 1[x \in H] : H$ halfspace$\}$.
For $d = 2$, $d_{VC}(F) = 3$. [PICTURE]

Vapnik-Chervonenkis dimension: Example

Theorem: For the class of thresholded linear functions,

$$F = \{x \mapsto 1[g(x) \geq 0] : g \in G\}, \qquad \text{where } G \text{ is a linear space,}$$

$d_{VC}(F) = dim(G)$.

Proof:

Vapnik-Chervonenkis Lemma: Proof

Fix $x_1, \ldots, x_n$ and consider the table of values of $F(x_1^n)$:

	x_1	x_2	x_3	x_4	x_5
f_1	0	1	0	1	1
f_2	1	0	0	1	1
f_3	1	1	1	0	1
f_4	0	1	1	0	0
f_5	0	0	0	1	0

The cardinality of $F(x_1^n)$ is the number of distinct rows.

Vapnik-Chervonenkis Lemma: Proof

Consider the following shifting transformation of the table: For a column i, change each 1 to a 0, unless it would lead to a row that is already in the table.

Shifting the columns from left to right gives:

	x_1	x_2	x_3	x_4	x_5
f_1	0	1	0	0	0
f_2	0	0	0	1	1
f_3	0	0	0	0	1
f_4	0	0	0	0	0
f_5	0	0	0	1	0

Vapnik-Chervonenkis Lemma: Proof

Suppose this shifting operation is performed column-by-column until it leads to no change of the table. Then:

- The number of rows does not change.
- Consider a row with any 1s. Every row with some of those 1s changed to 0s is in the table.

Vapnik-Chervonenkis Lemma: Proof

- The VC-dimension never increases. (Consider a set that is shattered after shifting a column. If the set does not include the column, it was certainly shattered before shifting. If it does include the column, we need to show that the set was shattered before. Suppose that an entry was shifted down to a zero. The 1s that remain in the column are there because there was a row before shifting that is identical but for a 0 in that column. Those 0s suffice for the shattering, and the newly shifted 0 is not needed for the shattering. But those 0s were present before shifting, so the set was shattered before.)
- So no row has more than d 1s.

Vapnik-Chervonenkis Lemma: Proof

Thus, the number of rows is no more than $\sum_{i=0}^{d} \binom{n}{i}$.

Finally, for $n \geq d$,

$$\begin{aligned}
\sum_{i=0}^{d} \binom{n}{i} &\leq \left(\frac{n}{d}\right)^d \sum_{i=0}^{d} \binom{n}{i} \left(\frac{d}{n}\right)^i \\
&= \left(\frac{n}{d}\right)^d \left(1 + \frac{d}{n}\right)^n \qquad \text{(binomial theorem)} \\
&\leq \left(\frac{en}{d}\right)^d .
\end{aligned}$$

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 $k \leq 2^t$, 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, $\Pi_F(n) \leq$ 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 $f(x,\theta) = \mathrm{sign}(w^T x - w_0)$, 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 $w^T x_i - w_0 = 0$. 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 $\mathbb{R}^d$ (and so $C(n, d)$ cells), and we add a plane, the number of cells that we split in two is precisely $C(n, d-1)$, the number of cells in the new plane (a $d-1$-subspace) that the first n planes define. Then an inductive argument shows that

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