Theoretical Statistics. Lecture 10.

Peter Bartlett

Uniform laws of large numbers: Bounding Rademacher complexity.

1. Growth function.
2. Vapnik-Chervonenkis dimension.

Recall: Uniform laws and Rademacher complexity

Definition: The **Rademacher complexity** of F is $\mathbf{E}\|R_n\|_F$, where the empirical process R_n is defined as

$$R_n(f) = \left| \frac{1}{n} \sum_{i=1}^{n} \epsilon_i f(X_i) \right|,$$

where the $\epsilon_1, \ldots, \epsilon_n$ are Rademacher random variables: i.i.d. uniform on $\{\pm 1\}$.

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 $\|P - P_n\|_F \stackrel{as}{\to} 0$.

Controlling Rademacher complexity: Growth function

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

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

[where R_n is the Rademacher process:

$$R_n(f) = \frac{1}{n}\sum_{i=1}^{n} \epsilon_i f(X_i).$$

and $F(X_1^n)$ is the set of restrictions of functions in F to $X_1, \ldots, X_n$.]

Controlling Rademacher complexity: Growth function

Proof: For $A \subseteq \mathbb{R}^n$ with $R = \max_{a \in A} \|a\|_2$, we saw that

$$\mathbf{E} \sup_{a \in A} \left| \frac{1}{n} \sum_{i=1}^{n} \epsilon_i a_i \right| \leq \frac{R}{n} \sqrt{2 \log(|A \cup -A|)}.$$

Here, we have $A = F(X_1^n)$, so $R \leq \sqrt{n}$, and we get

$$\begin{aligned}
\mathbf{E}\|R_n\|_F &= \mathbf{E}\mathbf{E}\left[\|R_n\|_{F(X_1^n)} | X_1, \ldots, X_n\right] \\
&\leq \mathbf{E}\sqrt{\frac{2\log(2|F(X_1^n)|)}{n}} \\
&\leq \sqrt{\frac{2\mathbf{E}\log(2|F(X_1^n)|)}{n}} \\
&\leq \sqrt{\frac{2\log(2\mathbf{E}|F(X_1^n)|)}{n}}.
\end{aligned}$$

Controlling Rademacher complexity: Growth function

e.g. For the class of distribution functions, $G = \{x \mapsto 1[x \leq \alpha] : \alpha \in \mathbb{R}\}$, we saw that $|G(x_1^n)| \leq n + 1$. So $\mathbf{E}\|R_n\|_F \leq \sqrt{\frac{2 \log 2(n+1)}{n}}$.

e.g. F parameterized by k bits:

If g maps to $[0, 1]$,

$$F = \left\{x \mapsto g(x, \theta) : \theta \in \{0, 1\}^k\right\},$$

$$|F(x_1^n)| \leq 2^k,$$

$$\mathbf{E}\|R_n\|_F \leq \sqrt{\frac{2(k+1) \log 2}{n}}.$$

Notice that $\mathbf{E}\|R_n\|_F \to 0$.

Growth function

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

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

- $\mathbf{E}\|R_n\|_F \leq \sqrt{\frac{2\log(2\Pi_F(n))}{n}}$.
- $\Pi_F(n) \leq |F|$, $\lim_{n\to\infty} \Pi_F(n) = |F|$.
- $\Pi_F(n) \leq 2^n$. (But then this gives no useful bound on $\mathbf{E}\|R_n\|_F$.)
- Notice that $\log \Pi_F(n) = o(n)$ 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] $d_{VC}(F) \leq d$ implies

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

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

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.

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

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

Thresholded linear functions:

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

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

Let $d = dim(G)$. To see that $d_{VC}(F) \geq d$, suppose that $g_1, \ldots, g_d \in G$ is a set of linearly independent functions. Then a fundamental result of linear algebra (row rank=column rank) implies that there are d points $x_1, \ldots, x_d$ such that the vectors $g_1(x_1^d), \ldots, g_d(x_1^d)$ are linearly independent. Let M be the $d \times d$ matrix of these values. Since G is linear, any linear combination of these functions is also in G. For coefficients v, this function's value on these d points is given by Mv. Since M is full rank, for any y, we can find a v so that $Mv = y$. In particular, y can have any sequence of signs, so $x_1, \ldots, x_d$ are shattered by G.

Vapnik-Chervonenkis dimension: Example

To see that $d_{VC}(F) \leq d$, consider any $x_1, \ldots, x_{d+1}$. Then

$$\{(g(x_1), \ldots, g(x_{d+1})) : g \in G\}$$

is a linear subspace of dimension d. So there must be a non-zero $v \in \mathbb{R}^{d+1}$ for which $\sum_i v_i g(x_i) = 0$ for all $g \in G$. Suppose that G shatters this set of $d + 1$ points. Wlog, suppose some $v_i > 0$. Consider a g for which $g(x_i) < 0$ for exactly those i with $v_i > 0$. Then

$$0 = \sum_i v_i g(x_i) = \underbrace{\sum_{i:v_i \leq 0} v_i g(x_i)}_{\leq 0} + \underbrace{\sum_{i:v_i > 0} v_i g(x_i)}_{< 0} < 0,$$

which is a contradiction.

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	0	1
f_3	0	0	1	0	1
f_4	0	0	1	0	0
f_5	0	0	0	0	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. So the newly shifted 0 plays no role in the shattering.)
- 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 $n \geq d$,

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

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