18.657: Mathematics of Machine Learning

Lecturer: Philippe Rigollet
Scribe: Vira Semenova and Philippe Rigollet

Lecture 5
Sep. 23, 2015

In this lecture, we complete the analysis of the performance of the empirical risk minimizer under a constraint on the VC dimension of the family of classifiers. To that end, we will see how to control Rademacher complexities using shatter coefficients. Moreover, we will see how the problem of controlling uniform deviations of the empirical measure μ_n from the true measure μ as done by Vapnik and Chervonenkis relates to our original classification problem.

4.1 Shattering

Recall from the previous lecture that we are interested in sets of the form

$$T(z) := \big\{(\mathbb{1}(z_1 \in A), \dots, \mathbb{1}(z_n \in A)), A \in \mathcal{A}\big\}\,, \quad z = (z_1, \dots, z_n)\,. \tag{4.1}$$

In particular, the cardinality of $T(z)$, i.e., the number of binary patterns these vectors can replicate as A ranges over $\mathcal{A}$, will be of critical importance, as it will arise when controlling the Rademacher complexity. Although the cardinality of $\mathcal{A}$ may be infinite, the cardinality of $T(z)$ is always at most 2^n. When it is of the size 2^n, we say that $\mathcal{A}$ *shatters* the set $z_1, \dots, z_n$. Formally, we have the following definition.

Definition: A collection of sets $\mathcal{A}$ *shatters* the set of points $\{z_1, z_2, ..., z_n\}$

$$\text{card}\{(\mathbb{1}(z_1 \in A), \dots, \mathbb{1}(z_n \in A)), A \in \mathcal{A}\} = 2^n\,.$$

The sets of points $\{z_1, z_2, ..., z_n\}$ that we are interested are realizations of the pairs $Z_1 = (X_1, Y_1), \dots, Z_n = (X_n, Y_n)$ and may, in principle take any value over the sample space. Therefore, we define the *shatter coefficient* to be the largest cardinality that we may obtain.

Definition: The *shatter coefficients* of a class of sets $\mathcal{A}$ is the sequence of numbers $\{\mathcal{S}_{\mathcal{A}}(n)\}_{n\ge 1}$, where for any $n \ge 1$,

$$\mathcal{S}_{\mathcal{A}}(n) = \sup_{z_1,\dots,z_n} \text{card}\{(\mathbb{1}(z_1 \in A), \dots, \mathbb{1}(z_n \in A)), A \in \mathcal{A}\}$$

and the suprema are taken over the whole sample space.

By definition, the nth shatter coefficient $\mathcal{S}_{\mathcal{A}}(n)$ is equal to 2^n if there exists a set $\{z_1, z_2, ..., z_n\}$ that $\mathcal{A}$ shatters. The largest of such sets is precisely the Vapnik-Chervonenkis or VC dimension.

Definition: The Vapnik-Chervonenkis dimension, or *VC-dimension* of $\mathcal{A}$ is the largest integer d such that $\mathcal{S}_{\mathcal{A}}(d) = 2^d$. We write $\mathsf{VC}(\mathcal{A}) = d$.

If $\mathcal{S}_{\mathcal{A}}(n) = 2^n$ for all positive integers n, then $\mathsf{VC}(\mathcal{A}) := \infty$

In words, $\mathcal{A}$ shatters *some* set of points of cardinality d but shatters *no* set of points of cardinality $d+1$. In particular, $\mathcal{A}$ also shatters no set of points of cardinality $d' > d$ so that the VC dimension is well defined.

In the sequel, we will see that the VC dimension will play the role similar to of cardinality, but on an exponential scale. For interesting classes $\mathcal{A}$ such that $\mathrm{card}(\mathcal{A}) = \infty$, we also may have $\mathsf{VC}(\mathcal{A}) < \infty$. For example, assume that $\mathcal{A}$ is the class of *half-lines*, $\mathcal{A} = \{(-\infty, a], a \in \mathbb{R}\} \cup \{[a, \infty), a \in \mathbb{R}\}$, which is clearly infinite. Then, we can clearly shatter a set of size 2 but we for three points $z_1, z_2, z_3, \in \mathbb{R}$, if for example $z_1 < z_2 < z_3$, we cannot create the pattern $(0,1,0)$ (see Figure 4.1). Indeed, half lines can can only create patterns with zeros followed by ones or with ones followed by zeros but not an alternating pattern like $(0,1,0)$.

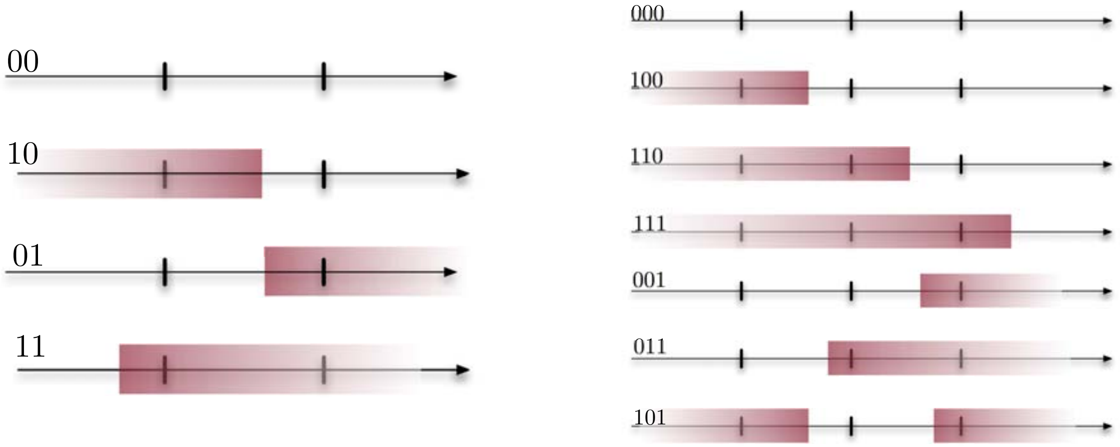

Figure 1: If $\mathcal{A} = \{\text{halflines}\}$, then any set of size $n = 2$ is shattered because we can create all $2^n = 4$ 0/1 patterns (left); if $n = 3$ the pattern $(0,1,0)$ cannot be reconstructed: $\mathcal{S}_A(3) = 7 < 2^3$ (right). Therefore, $\mathsf{VC}(\mathcal{A}) = 2$.

4.2 The VC inequality

We have now introduced all the ingredients necessary to state the main result of this section: the VC inequality.

Theorem (VC inequality): For any family of sets $\mathcal{A}$ with VC dimension $\mathsf{VC}(\mathcal{A}) = d$, it holds

$$\mathbb{E} \sup_{A \in \mathcal{A}} |\mu_n(A) - \mu(A)| \le 2\sqrt{\frac{2d\log(2en/d)}{n}}$$

Note that this result holds even if $\mathcal{A}$ is infinite as long as its VC dimension is finite. Moreover, observe that $\log(|\mathcal{A}|)$ has been replaced by a term of order $d \log\big(2en/d\big)$.

To prove the VC inequality, we proceed in three steps:

1. Symmetrization, to bound the quantity of interest by the Rademacher complexity:

$$\mathbb{E}[\sup_{A\in\mathcal{A}} |\mu_n(A) - \mu(A)|] \le 2\mathcal{R}_n(\mathcal{A}).$$

   We have already done this step in the previous lecture.

2. Control of the Rademacher complexity using shatter coefficients. We are going to show that

$$\mathcal{R}_n(\mathcal{A}) \le \sqrt{\frac{2\log\big(2\mathcal{S}_\mathcal{A}(n)\big)}{n}}$$

3. We are going to need the *Sauer-Shelah* lemma to bound the shatter coefficients by the VC dimension. It will yield

$$\mathcal{S}_\mathcal{A}(n) \le \Big(\frac{en}{d}\Big)^d, \quad d = \mathsf{VC}(\mathcal{A})\,.$$

Put together, these three steps yield the VC inequality.

Step 2: Control of the Rademacher complexity

We prove the following Lemma.

Lemma: For any $B \subset \mathbb{R}^n$, such that $|B| < \infty$:, it holds

$$\mathcal{R}_n(B) = \mathbb{E}\big[\max_{b\in B}\big|\frac{1}{n}\sum_{i=1}^n \sigma_i b_i\big|\big] \le \max_{b\in B} |b|_2 \frac{\sqrt{2\log(2|B|)}}{n}$$

where $|\cdot|_2$ denotes the Euclidean norm.

Proof. Note that

$$\mathcal{R}_n(B) = \frac{1}{n}\mathbb{E}\big[\max_{b\in B}|Z_b|\big]\,,$$

where $Z_b = \sum_{i=1}^n \sigma_i b_i$. In particular, since $-|b_i| \le \sigma_i|b_i| \le |b_i|$, a.s., Hoeffding's lemma implies that the moment generating function of Z_b is controlled by

$$\mathbb{E}\big[\exp(sZ_b)\big] = \prod_{i=1}^n \mathbb{E}\big[\exp(s\sigma_i b_i)\big] \le \prod_{i=1}^n \exp(s^2 b_i^2/2) = \exp(s^2|b|_2^2/2) \tag{4.2}$$

Next, to control $\mathbb{E}\big[\max_{b\in B}|Z_b|\big]$, we use the same technique as in Lecture 3, section 1.5. To that end, define $\bar{B} = B \cup \{-B\}$ and observe that for any $s > 0$,

$$\mathbb{E}\left[\max_{b\in B}|Z_b|\right] = \mathbb{E}\left[\max_{b\in\bar{B}} Z_b\right] = \frac{1}{s}\log\exp\left(s\mathbb{E}\left[\max_{b\in\bar{B}} Z_b\right]\right) \le \frac{1}{s}\log\mathbb{E}\left[\exp\left(s\max_{b\in\bar{B}} Z_b\right)\right],$$

where the last inequality follows from Jensen's inequality. Now we bound the max by a sum to get

$$\mathbb{E}\left[\max_{b\in B}|Z_b|\right] \le \frac{1}{s}\log\sum_{b\in\bar{B}}\mathbb{E}\left[\exp(sZ_b)\right] \le \frac{\log|\bar{B}|}{s} + \frac{s|b|_2^2}{2n},$$

where in the last inequality, we used (4.2). Optimizing over $s > 0$ yields the desired result. □

We apply this result to our problem by observing that

$$\mathcal{R}_n(\mathcal{A}) = \sup_{z_1,\ldots,z_n} \mathcal{R}_n(T(z))$$

where $T(z)$ is defined in (4.1). In particular, since $T(z) \subset \{0,1\}$, we have $|b|_2 \le \sqrt{n}$ for all $b \in T(z)$. Moreover, by definition of the shatter coefficients, if $B = T(z)$, then $|\bar{B}| \le 2|T(z)| \le 2\mathcal{S}_{\mathcal{A}}(n)$. Together with the above lemma, it yields the desired inequality:

$$\mathcal{R}_n(\mathcal{A}) \le \sqrt{\frac{2\log(2\mathcal{S}_{\mathcal{A}}(n))}{n}}\,.$$

Step 3: Sauer-Shelah Lemma

We need to use a lemma from combinatorics to relate the shatter coefficients to the VC dimension. A priori, it is not clear from its definition that the VC dimension may be at all useful to get better bounds. Recall that steps 1 and 2 put together yield the following bound

$$\mathbb{E}[\sup_{A\in\mathcal{A}} |\mu_n(A) - \mu(A)|] \le 2\sqrt{\frac{2\log(2\mathcal{S}_{\mathcal{A}}(n))}{n}} \tag{4.3}$$

In particular, if $\mathcal{S}_{\mathcal{A}}(n)$ is exponential in n, the bound (4.3) is not informative, i.e., it does not imply that the uniform deviations go to zero as the sample size n goes to infinity. The VC inequality suggest that this is not the case as soon as $\mathsf{VC}(\mathcal{A}) < \infty$ but it is not clear a priori. Indeed, it may be the case that $\mathcal{S}_{\mathcal{A}}(n) = 2^n$ for $n \le d$ and $\mathcal{S}_{\mathcal{A}}(n) = 2^n - 1$ for $n > d$, which would imply that $\mathsf{VC}(\mathcal{A}) = d < \infty$ but that the right-hand side in (4.3) is larger than 2 for all n. It turns our that this can never be the case: if the VC dimension is finite, then the shatter coefficients are at most *polynomial* in n. This result is captured by the Sauer-Shelah lemma, whose proof is omitted. The reading section of the course contains pointers to various proofs, specifically the one based on *shifting* which is an important technique in enumerative combinatorics.

Lemma (Sauer-Shelah): If $\mathsf{VC}(\mathcal{A}) = d$, then $\forall n \ge 1$,

$$\mathcal{S}_{\mathcal{A}}(n) \le \sum_{k=0}^{d} \binom{n}{k} \le \left(\frac{en}{d}\right)^d.$$

Together with (4.3), it clearly yields the VC inequality. By applying the bounded difference inequality, we also obtain the following VC inequality that holds with high probability. This is often the preferred from for this inequality in the literature.

Corollary (VC inequality): For any family of sets $\mathcal{A}$ such that $\mathsf{VC}(\mathcal{A}) = d$ and any $\delta \in (0,1)$, it holds with probability at least $1-\delta$,

$$\sup_{A\in\mathcal{A}} |\mu_n(A) - \mu(A)| \le 2\sqrt{\frac{2d\log(2en/d)}{n}} + \sqrt{\frac{\log(2/\delta)}{2n}}\,.$$

Note that the logarithmic term $\log(2en/d)$ is actually superfluous and can be replaced by a numerical constant using a more careful bounding technique. This is beyond the scope of this class and the interested reader should take a look at the recommending readings.

4.3 Application to ERM

The VC inequality provides an upper bound for $\sup_{A\in\mathcal{A}} |\mu_n(A) - \mu(A)|$ in terms of the VC dimension of the class of sets $\mathcal{A}$. This result translates directly to our quantity of interest:

$$\sup_{h\in H} |\hat{R}_n(h) - R(h)| \leq 2\sqrt{\frac{2\mathsf{VC}(\mathcal{A})\log\left(\frac{2en}{\mathsf{VC}(\mathcal{A})}\right)}{n}} + \sqrt{\frac{\log(2/\delta)}{2n}} \tag{4.4}$$

where $\mathcal{A} = \{A_h \,:\, h \in \mathcal{H}\}$ and $A_h = \{(x,y) \in \mathcal{X} \times \{0,1\} : h(x) \neq y\}$. Unfortunately, the VC dimension of this class of subsets of $\mathcal{X} \times \{0,1\}$ is not very natural. Since, a classifier h is a $\{0,1\}$ valued function, it is more natural to consider the VC dimension of the family $\mathcal{A} = \big\{\{h = 1\} \,:\, h \in \mathcal{H}\big\}$.

Definition: Let $\mathcal{H}$ be a collection of classifiers and define

$$\bar{\mathcal{A}} = \big\{\{h=1\} \,:\, h \in \mathcal{H}\big\} = \big\{A : \exists\, h \in \mathcal{H}, h(\cdot) = \mathbb{1}(\cdot \in A)\big\}\,.$$

We define the VC dimension $\mathsf{VC}(\mathcal{H})$ of $\mathcal{H}$ to be the VC dimension of $\bar{\mathcal{A}}$.

It is not clear how $\mathsf{VC}(\bar{\mathcal{A}})$ relates to the quantity $\mathsf{VC}(\bar{\mathcal{A}})$, where $\mathcal{A} = \{A_h \,:\, h \in \mathcal{H}\}$ and $A_h = \{(x,y) \in \mathcal{X} \times \{0,1\} : h(x) \neq y\}$ that appears in the VC inequality. Fortunately, these two are actually equal as indicated in the following lemma.

Lemma: Define the two families for sets: $\mathcal{A} = \{A_h \,:\, h \in \mathcal{H}\} \in 2^{\mathcal{X}\times\{0,1\}}$ where $A_h = \{(x,y) \in \mathcal{X} \times \{0,1\} : h(x) \neq y\}$ and $\bar{\mathcal{A}} = \big\{\{h=1\} \,:\, h \in \mathcal{H}\big\} \in 2^{\mathcal{X}}$.
Then, $\mathcal{S}_{\bar{\mathcal{A}}}(n) = \mathcal{S}_{\bar{\mathcal{A}}}(n)$ for all $n \geq 1$. It implies $\mathsf{VC}(\bar{\mathcal{A}}) = \mathsf{VC}(\mathcal{A})$.

Proof. Fix $x = (x_1, ..., x_n) \in \mathcal{X}^n$ and $y = (y_1, y_2, ..., y_n) \in \{0,1\}^n$ and define

$$T(x,y) = \{(\mathbb{1}(h(x_1) \neq y_1), \ldots, \mathbb{1}(h(x_n) \neq y_n)), h \in \mathcal{H}\}$$

and

$$\bar{T}(x) = \{(\mathbb{1}(h(x_1) = 1), \ldots, \mathbb{1}(h(x_n) = 1)), h \in \mathcal{H}\}$$

To that end, fix $v \in \{0,1\}$ and recall the XOR (exclusive OR) boolean function from $\{0,1\}$ to $\{0,1\}$ defined by $u \oplus v = \mathbb{1}(u \neq v)$. It is clearly[1] a bijection since $(u \oplus v) \oplus v = u$.

[1] One way to see that is to introduce the "spinned" variables $\tilde{u} = 2u - 1$ and $\tilde{v} = 2v - 1$ that live in $\{-1,1\}$. Then $\widetilde{u \oplus v} = \tilde{u} \cdot \tilde{v}$, and the claim follows by observing that $(\tilde{u} \cdot \tilde{v}) \cdot \tilde{v} = \tilde{u}$. Another way is to simply write a truth table.

When applying XOR componentwise, we have

$$\begin{pmatrix} \mathbb{1}(h(x_1) \neq y_1) \\ \vdots \\ \mathbb{1}(h(x_i) \neq y_i) \\ \vdots \\ \mathbb{1}(h(x_n) \neq y_n) \end{pmatrix} = \begin{pmatrix} \mathbb{1}(h(x_1) = 1) \\ \vdots \\ \mathbb{1}(h(x_i) = 1) \\ \vdots \\ \mathbb{1}(h(x_n) = 1) \end{pmatrix} \oplus \begin{pmatrix} y_1 \\ \vdots \\ y_i \\ \vdots \\ y_n \end{pmatrix}$$

Since XOR is a bijection, we must have $\mathrm{card}[T(x,y)] = \mathrm{card}[\bar{T}(x)]$. The lemma follows by taking the supremum on each side of the equality. □

It yields the following corollary to the VC inequality.

Corollary: Let $\mathcal{H}$ be a family of classifiers with VC dimension d. Then the empirical risk classifier $\hat{h}^{\mathrm{erm}}$ over $\mathcal{H}$ satisfies

$$R(\hat{h}^{\mathrm{erm}}) \leq \min_{h \in \mathcal{H}} R(h) + 4\sqrt{\frac{2d \log(2en/d)}{n}} + \sqrt{\frac{\log(2/\delta)}{2n}}$$

with probability $1 - \delta$.

Proof. Recall from Lecture 3 that

$$R(\hat{h}^{\mathrm{erm}}) - \min_{h \in \mathcal{H}} R(h) \leq 2 \sup_{h \in \mathcal{H}} \left| \hat{R}_n(h) - R(h) \right|$$

The proof follows directly by applying (4.4) and the above lemma. □