EECS 281B / STAT 241B: Advanced Topics in Statistical Learning **Spring 2009**

Lecture 14 — March 9

Lecturer: Martin Wainwright *Scribe: Nicholas Hay*

Note: These lecture notes are still rough, and have only have been mildly proofread.

14.1 More on shattering and VC dimension

Given a class $\mathcal{A}$ of subsets, its shattering coefficients are given by

$$\mathbf{s}(\mathcal{A}, n) = \max_{z_1,\ldots,z_n} \operatorname{card}\{A \cap \{z_1, \ldots, z_n\} \mid A \in \mathcal{A}\}$$

and its VC dimension by $V_{\mathcal{A}} = \sup\{n \mid \mathbf{s}(\mathcal{A}, n) = 2^n\}$.

Example: The class of one dimensional half spaces $\mathcal{A}_1 = \{(-\infty, a] \mid a \in \mathbb{R}\}$ has $\mathbf{s}(\mathcal{A}_1, n) = n + 1$ and so $V_{\mathcal{A}_1} = 1$. The class of half open intervals $\mathcal{A}_2 = \{(b, a] \mid b < a \in \mathbb{R}\}$ has $\mathbf{s}(\mathcal{A}_1, n) = \frac{n(n+1)}{2} + 1$ and so $V_{\mathcal{A}_2} = 2$.

Recall from previous lectures:

Theorem 14.1 (GC). *Given any class of sets $\mathcal{A}$*

$$\mathbb{P}\left(\sup_{A\in\mathcal{A}} |\hat{\mathbb{P}}_n(A) - \mathbb{P}(A)| > \epsilon\right) \leq 8\,\mathbf{s}(\mathcal{A}, n) \exp\left\{-\frac{n\epsilon^2}{32}\right\}$$

where $\hat{\mathbb{P}}_n(A) = \frac{1}{n}\sum_{i=1}^n \mathbb{I}(z^{(i)} \in A)$ for iid samples $Z^{(i)}$ for $i = 1, \ldots, n$.

VC dimension and shatter coefficients are closely connected:

1. If $V_{\mathcal{A}} = \infty$ then $\mathbf{s}(\mathcal{A}, n) = 2^n$ for all n,
2. If $V_{\mathcal{A}} < \infty$ then $\mathbf{s}(\mathcal{A}, n) \leq (n + 1)^{V_{\mathcal{A}}}$ for all n.

The first is by definition, the second as a corollary of the following lemma.

Lemma 14.2 (Sauer). *If $\mathcal{A}$ be a class with finite VC dimension $V_{\mathcal{A}}$, then*

$$\mathbf{s}(\mathcal{A}, n) \leq \sum_{i=0}^{V_{\mathcal{A}}} \binom{n}{i}.$$

Given this, we can derive the (weak) upper bound

$$\begin{aligned}
\mathbf{s}(\mathcal{A}, n) &\leq \sum_{i=0}^{V_{\mathcal{A}}} \frac{n!}{i!(n-i)!} \\
&\leq \sum_{i=0}^{V_{\mathcal{A}}} n^i \frac{1}{i!} \\
&\leq \sum_{i=0}^{V_{\mathcal{A}}} n^i \binom{V_{\mathcal{A}}}{i} \\
&= (n+1)^{V_{\mathcal{A}}}
\end{aligned}$$

So far we've computed the VC dimension of classes case-by-case. We want systematic ways to upper bound the VC dimension. The following proposition is the first.

Proposition 14.3. *Let $\mathcal{G}$ be a finite-dimensional vector space of functions on $\mathbb{R}^d$. Then the class of sets*

$$\mathcal{A}_{\mathcal{G}} = \Big\{ \{x \mid g(x) \geq 0\} \mid g \in \mathcal{G} \Big\}$$

has VC dimension at most $\dim \mathcal{G}$.

Proof: We will show that no subset of $\mathbb{R}^d$ of size $n = \dim \mathcal{G} + 1$ can be shattered by $\mathcal{A}_{\mathcal{G}}$. Fix n points $x^{(1)}, \ldots, x^{(n)} \in \mathbb{R}^d$. Consider the map $L \colon \mathcal{G} \to \mathbb{R}^n$ defined by

$$L(g) = (g(x^{(1)}), \ldots, g(x^{(n)})).$$

This map is linear, and so its range is a linear subspace of R^n of dimension at most $\dim \mathcal{G}$. Since $n > \dim \mathcal{G}$ there must exist a nonzero vector $\gamma \in \mathbb{R}^n$ orthogonal to this subspace, i.e. such that

$$\sum_{i=1}^{n} \gamma_i g(x^{(i)}) = 0 \tag{14.1}$$

for all $g \in \mathcal{G}$. Without loss of generality suppose $\gamma_i < 0$ for some i, and observe that equation (14.1) is equivalent to

$$\sum_{\{i \mid \gamma_i \geq 0\}} \gamma_i g(x^{(i)}) = \sum_{\{i \mid \gamma_i < 0\}} -\gamma_i g(x^{(i)}) \tag{14.2}$$

for all $g \in \mathcal{G}$.

Now proceed via proof by contradiction: suppose that $x^{(1)}, \ldots, x^{(n)}$ can be shattered by $\mathcal{A}$. Then there must exist $g \in \mathcal{G}$ such that

$$\{x \mid g(x) \geq 0\} = \{i \mid \gamma_i \geq 0\}.$$

But with this choice of g the LHS of equation 14.2 must be nonnegative, whilst the RHS must be negative (since $\gamma_i < 0$ for some i), which is a contradiction. So we conclude that no subset of size n of $\mathbb{R}^d$ can be shattered. $\square$

Example: Consider the set of half spaces

$$\mathcal{A} = \Big\{ \{x \in \mathbb{R}^d \mid a^T x \geq b\} \mid \text{for some } a \in \mathbb{R}^d \text{ and } b \in \mathbb{R} \Big\}.$$

This class is of the form required for proposition 14.3, we need only compute the dimension of the underlying vector space of functions. This is seen to be $d+1$ by the following basis:

$$\begin{aligned} g_0(x) &= 1 \\ g_i(x) &= x_i \qquad \text{for } i = 1, \dots, d \end{aligned}$$

So $V_{\mathcal{A}} \leq d + 1$.

14.2 Application to binary classification

Suppose we are learning binary classifiers $f : \mathbb{R}^d \to \{-1, +1\}$ of the form

$$\mathcal{F} = \left\{ f = \operatorname{sgn}(g) \mid g(x) = a_0 + \sum_{i=1}^{d} a_i x_i,\ a_i \in \mathbb{R} \right\}.$$

From the previous example we have $V_{\mathcal{F}} \leq d + 1$. Define the optimal linear risk to be

$$R^*_{\mathcal{F}} = \inf_{f \in \mathcal{F}} R(f) = \inf_{f \in \mathcal{F}} \mathbb{P}(Y \neq f(X)).$$

Suppose $\hat{f}_n$ is selected to minimize the empirical risk given iid samples $(x^{(i)}, y^{(i)})$ for $i = 1, \dots, n$:

$$\hat{f}_n \in \operatorname*{argmin}_f \hat{R}_n(f) = \operatorname*{argmin}_f \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}(y^{(i)} \neq f(x^{(i)})).$$

Corollary 14.4. *For all $n \in \mathbb{N}$, $\epsilon > 0$ with $n\epsilon^2 > 2$, the error probability of the empirically optimal classifier $\hat{f}_n$ satisfies*

$$\mathbb{P}\left[\left| R(\hat{f}_n) - R^*_{\mathcal{F}} \right| > \epsilon \right] \leq 8(n+1)^{d+1} \exp\left\{ -\frac{n\epsilon^2}{128} \right\}.$$

Note that $\hat{f}_n$ is a random classifier: it depends on the particular n iid samples used to train it. The mild condition $n\epsilon^2 > 2$ is required for the GC theorem that we use in proving this corollary (see Step 1 [symmetrization] in proof of GC theorem). This is no real restriction since we care about the behaviour of this bound as n tends to infinity for fixed ϵ.

Proof: Observe we can decompose the error into two terms

$$\begin{aligned} R(\hat{f}_n) - R^*_{\mathcal{F}} &= R(\hat{f}_n) - \inf_{f \in \mathcal{F}} R(f) \\ &= [R(\hat{f}_n) - \hat{R}_n(\hat{f}_n)] + [\hat{R}_n(\hat{f}_n) - \inf_{f \in \mathcal{F}} R(f)] \end{aligned}$$

The first term is easily bounded

$$R(\hat{f}_n) - \hat{R}_n(\hat{f}_n) \le \sup_{f\in\mathcal{F}} \left| R(f) - \hat{R}_n(f) \right|$$

For the second, observe that for any $f \in \mathcal{F}$ we can uniformly bound

$$\begin{aligned} \hat{R}_n(\hat{f}_n) - R(f) &\le \hat{R}_n(\hat{f}_n) - R(f) \\ &\le \sup_{f\in\mathcal{F}} \left| \hat{R}_n(f) - R(f) \right| \end{aligned}$$

This bounds the second term

$$\begin{aligned} \hat{R}_n(\hat{f}_n) - \inf_{f\in\mathcal{F}} R(f) &= \sup \hat{R}_n(\hat{f}_n) - R(f) \\ &\le \sup_{f\in\mathcal{F}} \left| \hat{R}_n(f) - R(f) \right| \end{aligned}$$

Combining the above with theorem 14.1, lemma 14.2, and the bound on the VC dimension of $\mathcal{F}$ we have

$$\begin{aligned} \mathbb{P}\left[\left| R(\hat{f}_n) - R^*_{\mathcal{F}} \right| > \epsilon \right] &\le \mathbb{P}\left[\sup_{f\in\mathcal{F}} \left| \hat{R}_n(f) - R(f) \right| > \epsilon/2 \right] \\ &\le 8\mathbf{s}(\mathcal{A}, n) \exp\left\{ -\frac{n\epsilon^2}{128} \right\} \\ &\le 8(n+1)^{V_{\mathcal{F}}} \exp\left\{ -\frac{n\epsilon^2}{128} \right\} \\ &\le 8(n+1)^{d+1} \exp\left\{ -\frac{n\epsilon^2}{128} \right\} \end{aligned}$$

□

The above result can equivalently stated in terms of bounds on expectations:

Corollary 14.5. *Under the same conditions as corollary 14.4*

$$\begin{aligned} \mathbb{E}\left[R(\hat{f}_n) - R^*_{\mathcal{F}} \right] &\le 16\sqrt{\frac{\log 8e\,\mathbf{s}(\mathcal{F}, n)}{2n}} \\ &= O\left(\sqrt{\frac{\log \mathbf{s}(\mathcal{F}, n)}{n}} \right) \end{aligned}$$

If the $V_{\mathcal{F}} < +\infty$ then

$$\mathbb{E}\left[R(\hat{f}_n) - R^*_{\mathcal{F}} \right] = O\left(\sqrt{\frac{V_{\mathcal{F}} \log n}{n}} \right).$$

This corollary follows by a careful integration of the tail bound, as we will discuss in the next lecture.