E0 370 Statistical Learning Theory **Lecture 4 (Aug 23, 2011)**

VC-Dimension and Sauer's Lemma

Lecturer: Shivani Agarwal *Scribe: Achintya Kundu*

1 Introduction

In the previous lecture we saw the technique of uniform convergence for obtaining confidence bounds on the generalization error $\mathrm{er}_D[h_S]$ when the function h_S is selected from a function class $\mathcal{H}$ of sufficiently limited 'capacity'. Specifically, we saw that if h_S is selected from $\mathcal{H} \subseteq \{-1,1\}^{\mathcal{X}}$, then for any $0 < \delta < 1$, with probability at least $1-\delta$ (over the draw of $S \sim D^m$),

$$\mathrm{er}_D[h_S] \;\leq\; \mathrm{er}_S[h_S] + \sqrt{\frac{8\left(\ln \Pi_{\mathcal{H}}(2m) + \ln(\frac{4}{\delta})\right)}{m}}\,. \tag{1}$$

In order for this result to be meaningful (i.e. to give a non-trivial bound), the growth function $\Pi_{\mathcal{H}}(2m)$ needs to grow 'slowly' in m; in particular, it needs to have sub-exponential growth in m.

In this lecture we meet the *VC-dimension*, a fundamental combinatorial parameter associated with a class of binary-valued functions $\mathcal{H}$. We will see that if $\mathcal{H}$ has finite VC-dimension, then the growth function of $\mathcal{H}$ grows polynomially in m, yielding a meaningful uniform convergence result and generalization error bound.

2 Vapnik-Chervonenkis Dimension

Definition. Let $\mathcal{H} \subseteq \{-1,1\}^{\mathcal{X}}$ and let $x_1^m = (x_1, \ldots, x_m) \in \mathcal{X}^m$. We say x_1^m is *shattered* by $\mathcal{H}$ if $\left|\mathcal{H}_{|x_1^m}\right| = 2^m$; i.e. if $\forall \mathbf{b} \in \{-1,1\}^m$, $\exists h_{\mathbf{b}} \in \mathcal{H}$ s.t. $(h_{\mathbf{b}}(x_1), \ldots, h_{\mathbf{b}}(x_m)) = (b_1, \ldots, b_m)$. The *Vapnik-Chervonenkis (VC) dimension* of $\mathcal{H}$, denoted by $\mathrm{VCdim}(\mathcal{H})$, is the cardinality of the largest set of points in $\mathcal{X}$ that can be shattered by $\mathcal{H}$:

$$\mathrm{VCdim}(\mathcal{H}) = \max\left\{m \in \mathbb{N} \;\middle|\; \Pi_{\mathcal{H}}(m) = 2^m\right\}.$$

If $\mathcal{H}$ shatters arbitrarily large sets of points in $\mathcal{X}$, then $\mathrm{VCdim}(\mathcal{H}) = \infty$.

Example 1 (Intervals on the real line)**.** Let $\mathcal{X} = \mathbb{R}$, and let $\mathcal{H}$ be the class of all binary-valued functions on $\mathbb{R}$ that label all points within some closed interval as $+1$ and all points outside the interval as -1:

$$\mathcal{H} = \left\{h : \mathcal{X} \to \{-1,1\} \;\middle|\; h(x) = 1 \text{ if } x \in [a,b] \text{ and } -1 \text{ otherwise, for some } a \leq b\right\}$$

Clearly, any set of 2 points $x_1 < x_2$ in $\mathbb{R}$ can be shattered by $\mathcal{H}$: consider the functions in $\mathcal{H}$ corresponding to the intervals $[x_1-2, x_1-1]$, $[x_1-1, \frac{x_1+x_2}{2}]$, $[\frac{x_1+x_2}{2}, x_2+1]$, and $[x_1-1, x_2+1]$; these functions realize all possible binary labelings of the 2 points. Moreover, no set of 3 points $x_1 < x_2 < x_3$ in $\mathbb{R}$ can be shattered by $\mathcal{H}$: no function in $\mathcal{H}$ can label x_2 as negative and x_1, x_3 as positive. Therefore $\mathrm{VCdim}(\mathcal{H}) = 2$.

Example 2 (Axis-parallel rectangles in the plane)**.** Let $\mathcal{X} = \mathbb{R}^2$, and let $\mathcal{H}$ be the class of all binary-valued functions on $\mathcal{X}$ that label all points within some axis-parallel rectangle as $+1$ and all points outside the rectangle as -1:

$$\mathcal{H} = \left\{h : \mathcal{X} \to \{-1,1\} \;\middle|\; h(x) = 1 \text{ if } x \in [a,b] \times [c,d] \text{ and } -1 \text{ otherwise, for some } a \leq b, c \leq d\right\}$$

Figure 1 shows a set of 4 points in $\mathbb{R}^2$ that are shattered by $\mathcal{H}$; therefore $\mathrm{VCdim}(\mathcal{H}) \geq 4$ (note that there are also sets of 4 points that are *not* shattered by $\mathcal{H}$, such as any set that includes 3 points on a line; however this does not matter: to show $\mathrm{VCdim}(\mathcal{H}) \geq d$, we just need *some* set of d points to be shattered by $\mathcal{H}$). Moreover, it can be verified that *no* set of 5 points in $\mathbb{R}^2$ can be shattered by $\mathcal{H}$ (in any set of 5 points, there is at least one point which is neither the sole extreme left/right nor the sole extreme top/bottom; there is no function in $\mathcal{H}$ that can label this point negative and all others positive). Therefore $\mathrm{VCdim}(\mathcal{H}) = 4$.

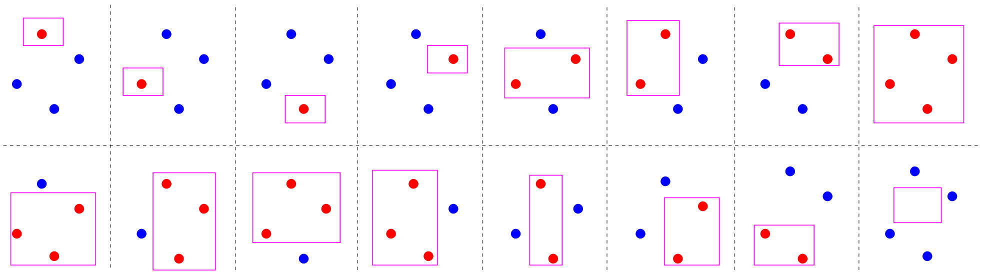

Figure 1: Four points in $\mathbb{R}^2$ that can be shattered using axis-parallel rectangles.

Example 3 (Linear classifiers in $\mathbb{R}^n$)**.** Let $\mathcal{X} = \mathbb{R}^n$, and let $\mathcal{H}$ be the class of linear classifiers in $\mathbb{R}^n$:

$$\mathcal{H} = \left\{ h : \mathcal{X} \to \{-1, 1\} \;\middle|\; h(x) = \mathrm{sign}(\mathbf{w} \cdot \mathbf{x} + b) \text{ for some } \mathbf{w} \in \mathbb{R}^n, b \in \mathbb{R} \right\}.$$

Consider first the case $n = 2$. Figure 2 shows a set of 3 points in $\mathbb{R}^2$ that are shattered by $\mathcal{H}$. Moreover, it is easy to verify that no set of 4 points in $\mathbb{R}^2$ can be shattered by $\mathcal{H}$. Therefore $\mathrm{VCdim}(\mathcal{H})$ in this case is 3. In general, for linear classifiers in $\mathbb{R}^n$, $\mathrm{VCdim}(\mathcal{H}) = n + 1$.

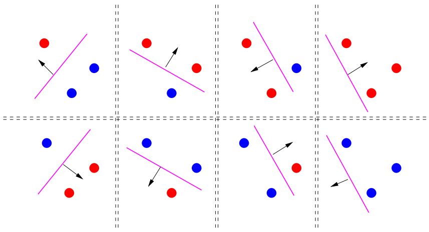

Figure 2: Three points in $\mathbb{R}^2$ that can be shattered using linear classifiers.

Example 4 (Finite unions of intervals on the real line)**.** Let $\mathcal{X} = \mathbb{R}$, and let $\mathcal{H}$ be the class of all functions on $\mathbb{R}$ that label all points within some finite union of closed intervals as 1 and all other points as -1:

$$\mathcal{H} = \left\{ h : \mathcal{X} \to \{-1, 1\} \;\middle|\; h(x) = 1 \text{ if } x \in \textstyle\bigcup_{i=1}^{k} [a_i, b_i] \text{ and } -1 \text{ otherwise, for some } k \in \mathbb{N}, a_i \leq b_i \ \forall i \in [k] \right\}.$$

Then $\mathrm{VCdim}(\mathcal{H}) = \infty$.

3 Sauer's Lemma

Theorem 3.1 (Sauer's Lemma; Sauer, 1972; Shelah, 1972; Vapnik & Chervonenkis, 1971)**.** Let $\mathcal{H} \subseteq \{-1, 1\}^{\mathcal{X}}$ with $\mathrm{VCdim}(\mathcal{H}) = d < \infty$. Then for all $m \in \mathbb{N}$,

$$\Pi_{\mathcal{H}}(m) \leq \sum_{i=0}^{d} \binom{m}{i}.$$

Proof. By induction on both m and d.

Base case. There are two base cases to consider:

(i) $d = 0, m \geq 1$. In this case $\mathcal{H}$ can contain only a single function, and we have $\Pi_{\mathcal{H}}(m) = 1$; moreover, $\sum_{i=0}^{d} \binom{m}{i} = \binom{m}{0} = 1$. Therefore the hypothesis holds.

(ii) $m = 1, d \geq 1$. In this case $\Pi_{\mathcal{H}}(m) = 2$; moreover, $\sum_{i=0}^{d} \binom{m}{i} = \binom{1}{0} + \binom{1}{1} = 2$. Therefore the hypothesis holds in this case as well.

Induction step. Let $m > 1, d > 0$. Assume the hypothesis for all (m', d') such that $m' < m$ or $d' < d$ (in particular, we will need only the cases $(m-1, d-1)$ and $(m-1, d)$). We will show the hypothesis is true for (m, d).

Let $x_1^m = (x_1, \ldots, x_m) \in \mathcal{X}^m$. Consider

$$\mathcal{H}_{|x_1^m} = \Big\{ (h(x_1), \ldots, h(x_m)) \;\Big|\; h \in \mathcal{H} \Big\}$$

and

$$\mathcal{H}_{|x_1^{m-1}} = \Big\{ (h(x_1), \ldots, h(x_{m-1})) \;\Big|\; h \in \mathcal{H} \Big\} \,.$$

All sequences in $\mathcal{H}_{|x_1^{m-1}}$ appear either once or twice in $\mathcal{H}_{|x_1^m}$ (followed by either 1 or -1 in the m-th position, or both). Let H_3 be the set of all sequences in $\mathcal{H}_{|x_1^{m-1}}$ that appear twice in $\mathcal{H}_{|x_1^m}$:

$$\mathcal{H}_3 = \Big\{ (y_1, \ldots, y_{m-1}) \in \mathcal{H}_{|x_1^{m-1}} \;\Big|\; \exists h, h' \in \mathcal{H} \text{ s.t. } h(x_i) = h'(x_i) = y_i \, \forall i \in [m-1] \text{ and } h(x_m) \neq h'(x_m) \Big\} \,.$$

Then clearly

$$\left|\mathcal{H}_{|x_1^m}\right| \;=\; \left|\mathcal{H}_{|x_1^{m-1}}\right| + |\mathcal{H}_3| \,. \tag{2}$$

Now $\mathcal{H}_{|x_1^{m-1}}$ is a restriction to $m-1$ points of functions from the class $\mathcal{H}$ of VC-dimension d. Moreover, $\mathcal{H}_3$ can be viewed as a restriction to $m-1$ points of functions from a class of VC-dimension at most $d-1$ (to see this, note that to any subset of the $m-1$ points $x_1, \ldots, x_{m-1}$ that is shattered in $\mathcal{H}_3$, we can add x_m to obtain a strictly larger set of points that is shattered by $\mathcal{H}$; the claim follows since $\text{VCdim}(\mathcal{H}) = d$). Applying the inductive hypothesis then gives

$$\begin{aligned}
\left|\mathcal{H}_{|x_1^m}\right| &\leq \sum_{i=0}^{d} \binom{m-1}{i} + \sum_{i=0}^{d-1} \binom{m-1}{i} && (3)\\
&= \sum_{i=0}^{d} \left(\binom{m-1}{i} + \binom{m-1}{i-1} \right) \quad \left(\text{where } \tbinom{m-1}{-1} = 0\right) && (4)\\
&= \sum_{i=0}^{d} \binom{m}{i} . && (5)
\end{aligned}$$

Since $x_1^m \in \mathcal{X}^m$ was arbitrary, the result follows. □

Corollary 3.2. Let $\mathcal{H} \subseteq \{-1, 1\}^{\mathcal{X}}$ with $\text{VCdim}(\mathcal{H}) = d < \infty$. For all $m \geq d$,

$$\Pi_{\mathcal{H}}(m) \leq \left(\frac{em}{d}\right)^d .$$

Proof. We have

$$\begin{aligned}
\Pi_{\mathcal{H}}(m) &\leq \sum_{i=0}^{d} \binom{m}{i} && (6)\\
&\leq \sum_{i=0}^{d} \binom{m}{i} \left(\frac{m}{d}\right)^{d-i} \quad \text{(since } m \geq d \text{ and } d \geq i) && (7)\\
&= \left(\frac{m}{d}\right)^{d} \sum_{i=0}^{d} \binom{m}{i} \left(\frac{d}{m}\right)^{i} 1^{m-i} && (8)\\
&\leq \left(\frac{m}{d}\right)^{d} \left(1+\frac{d}{m}\right)^{m} && (9)\\
&\leq \left(\frac{em}{d}\right)^{d}. && (10)
\end{aligned}$$

□

This yields the following sharp dichotomy: if $\text{VCdim}(\mathcal{H}) = \infty$, then $\Pi_{\mathcal{H}}(m) = 2^m \; \forall m \in \mathbb{N}$; on the other hand, if $\mathcal{H}$ has finite VC-dimension d, then $\Pi_{\mathcal{H}}(m) = O(m^d)$ (specifically, $\Pi_{\mathcal{H}}(m) = 2^m$ for $m \leq d$, and $\Pi_{\mathcal{H}}(m) \leq (em/d)^d$ for $m > d$). Thus the VC-dimension can be viewed as providing a one-number summary of the capacity of a class of binary-valued functions.

Combining the above with the uniform convergence result we obtained earlier in terms of the growth function, we thus get the following high confidence bound on the generalization error of a function learned from a function class $\mathcal{H}$ of finite VC-dimension:

Corollary 3.3. Let $\mathcal{H} \subseteq \{-1,1\}^{\mathcal{X}}$ with $\text{VCdim}(\mathcal{H}) = d < \infty$. Let D be any distribution on $\mathcal{X} \times \{-1,1\}$, and let $0 < \delta < 1$. For any algorithm that given a training sample S returns a function $h_S \in \mathcal{H}$, we have with probability at least $1-\delta$ (over the draw of $S \sim D^m$):

$$\text{er}_D[h_S] \;\leq\; \text{er}_S[h_S] + \sqrt{\frac{8(d(\ln(2m)+1)+\ln(\frac{4}{\delta}))}{m}}\,. \tag{11}$$

When comparing generalization error bounds obtained using different techniques, we will often ignore constant factors and will focus more on how the confidence bound/interval depends on various quantities of interest, such as the capacity/complexity of the class (here the VC-dimension d), the number of training examples m, and the confidence parameter δ. Thus we will often write the above bound as

$$\text{er}_D[h_S] \;\leq\; \text{er}_S[h_S] + c\sqrt{\frac{d\ln m + \ln(\frac{1}{\delta})}{m}}\,, \tag{12}$$

where $c > 0$ is some constant.

Exercise. Say you have learned a linear classifier using the SVM algorithm on a training sample $S \in (\mathbb{R}^n \times \{-1,1\})^m$ of size $m = 1000$, which you believe can be assumed to be drawn according to D^m for some fixed (but unknown) distribution D (on $\mathbb{R}^n \times \{-1,1\}$). Suppose the data is of dimensionality $n = 10$. You measure the training error $\text{er}_S[h_S] = \frac{1}{m}\sum_{i=1}^{m} \mathbf{1}(h_S(\mathbf{x}_i) \neq y_i)$ of the classifier h_S you have learned, and this turns out to be 0.217. Using the results we have seen so far, what can you say about the expected error of this classifier on a new example drawn according to D, $\text{er}_D[h_S]$? Can you derive a 95% confidence bound/interval for the expected error? What if $n = 1000$? What if $n = 10$, but you learn a cubic polynomial classifier using a degree-q polynomial kernel for $q = 3$? What if you use a Gaussian kernel?

4 Next Lecture

In the next lecture, we will consider obtaining bounds on the generalization error when learning real-valued functions, and will see how one can extend the notion of capacity to real-valued function classes. This will include concepts like covering numbers, pseudo-dimension, and fat-shattering dimension.