E0 370 Statistical Learning Theory **Lecture 5 (Aug 25, 2011)**

Covering Numbers, Pseudo-Dimension, and Fat-Shattering Dimension

Lecturer: Shivani Agarwal *Scribe: Shivani Agarwal*

1 Introduction

So far we have seen how to obtain high confidence bounds on the generalization error $\mathrm{er}_D^{0\text{-}1}[h_S]$ of a binary classifier h_S learned by an algorithm from a function class $\mathcal{H} \subseteq \{-1,1\}^{\mathcal{X}}$ of limited capacity, using the ideas of uniform convergence. We saw the use of the growth function $\Pi_{\mathcal{H}}(m)$ to measure the capacity of the class $\mathcal{H}$, as well as the VC-dimension $\mathrm{VCdim}(\mathcal{H})$, which provides a one-number summary of the capacity of $\mathcal{H}$.

In this lecture we will consider the problem of *regression*, or learning of real-valued functions. Here we are given a training sample $S = ((x_1, y_1), \ldots, (x_m, y_m)) \in (\mathcal{X} \times \mathcal{Y})^m$ for some $\mathcal{Y} \subseteq \mathbb{R}$, assumed now to be drawn from D^m for some distribution D on $\mathcal{X} \times \mathcal{Y}$, and the goal is to learn from this sample a real-valued function $f_S : \mathcal{X} \to \mathbb{R}$ that has low generalization error $\mathrm{er}_D^{\ell}[f_S]$ w.r.t. some appropriate loss function $\ell : \mathcal{Y} \times \mathbb{R} \to [0, \infty)$, such as the squared loss ℓ_{sq} given by $\ell_{\mathrm{sq}}(y, \hat{y}) = (\hat{y} - y)^2$. As before, we will be interested in obtaining high confidence bounds on the generalization error of the learned function, $\mathrm{er}_D^{\ell}[f_S]$. Again, we will consider learning algorithms that learn f_S from a function class $\mathcal{F} \subseteq \mathbb{R}^{\mathcal{X}}$ of limited capacity, and obtain generalization error bounds for such algorithms via a uniform convergence result that upper bounds the probability

$$\mathbf{P}_{S \sim D^m}\Big(\sup_{f \in \mathcal{F}} \Big| \mathrm{er}_D^{\ell}[f] - \mathrm{er}_S^{\ell}[f] \Big| \geq \epsilon \Big) .$$

However in order to derive such a result, we will need a different notion of capacity for a class of real-valued functions $\mathcal{F}$. In particular, we will use the *covering numbers* of $\mathcal{F}$, which will play a role analogous to that played by the growth function in the case of binary-valued function classes.

2 Covering Numbers

We start by considering covering numbers of subsets of a general metric space. We will then specialize this to subsets of Euclidean space, and use this to define covering numbers for a real-valued function class.

2.1 Covering Numbers in a General Metric Space

Let (A, d) be a metric space.[1] Let $W \subseteq A$ and let $\epsilon > 0$. A set $C \subseteq W$ is said to be a (proper) ϵ-*cover* of W w.r.t. d if for every $w \in W$, $\exists\, c \in C$ such that $d(w, c) < \epsilon$.[2] In other words, $C \subseteq W$ is an ϵ-cover of W w.r.t. d if the union of (open) d-balls of radius ϵ centered at points in C contains W:[3]

$$\bigcup_{c \in C} B_{d,\epsilon}(c) \;\supseteq\; W . \tag{1}$$

If W has a finite ϵ-cover w.r.t. d, then we define the ϵ-*covering number* of W (w.r.t. d) to be the cardinality of the smallest ϵ-cover of W:

$$\mathcal{N}(\epsilon, W, d) = \min\Big\{ |C| \;\Big|\; C \text{ is an } \epsilon\text{-cover of } W \text{ w.r.t. } d \Big\} . \tag{2}$$

[1] Recall that a metric space (A, d) consists of a set A together with a metric $d : A \times A \to [0, \infty)$ that satisfies the following for all $x, y, z \in A$: (1) $d(x, y) = 0 \Leftrightarrow x = y$; (2) $d(x, y) = d(y, x)$; and (3) $d(x, z) \leq d(x, y) + d(x, z)$.

[2] Sometimes it is also convenient to consider *improper* covers $C \subseteq A$ which need not be contained in W.

[3] Recall that the open d-ball centered at $x \in A$ is defined as $B_{d,\epsilon}(x) = \{y \in A \mid d(x, y) < \epsilon\}$.

If W does not have a finite ϵ-cover w.r.t. d, we take $\mathcal{N}(\epsilon, W, d) = \infty$. Thus the covering numbers of W can be viewed as measuring the 'extent' of W in (A, d) at 'granularity' or 'scale' ϵ.

2.2 Covering Numbers in Euclidean Space

Consider now $A = \mathbb{R}^n$. We can define a number of different metrics on $\mathbb{R}^n$, including in particular the following:

$$d_1(\mathbf{x}, \mathbf{x}') = \frac{1}{n}\sum_{i=1}^{n} |x_i - x_i'| \tag{3}$$

$$d_2(\mathbf{x}, \mathbf{x}') = \sqrt{\frac{1}{n}\sum_{i=1}^{n} (x_i - x_i')^2} \tag{4}$$

$$d_\infty(\mathbf{x}, \mathbf{x}') = \max_i |x_i - x_i'| \,. \tag{5}$$

Accordingly, for any $W \subseteq \mathbb{R}^n$, we can define the corresponding covering numbers $\mathcal{N}(\epsilon, W, d_p)$ for $p = 1, 2, \infty$. It is easy to see that $d_1(\mathbf{x}, \mathbf{x}') \le d_2(\mathbf{x}, \mathbf{x}')$ (by Jensen's inequality) and that $d_2(\mathbf{x}, \mathbf{x}') \le d_\infty(\mathbf{x}, \mathbf{x}')$, from which it follows that the corresponding covering numbers satisfy the relation

$$\mathcal{N}(\epsilon, W, d_1) \;\le\; \mathcal{N}(\epsilon, W, d_2) \;\le\; \mathcal{N}(\epsilon, W, d_\infty)\,. \tag{6}$$

2.3 Uniform Covering Numbers for a Real-Valued Function Class

Now let $\mathcal{F}$ be a class of real-valued functions on $\mathcal{X}$, and let $x_1^m = (x_1, \ldots, x_m) \in \mathcal{X}^m$. Then $\mathcal{F}_{|x_1^m} \subseteq \mathbb{R}^m$. For any $\epsilon > 0$ and $m \in \mathbb{N}$, the *uniform d_p covering numbers* of $\mathcal{F}$ (for $p = 1, 2, \infty$) are defined as

$$\mathcal{N}_p(\epsilon, \mathcal{F}, m) \;=\; \max_{x_1^m \in \mathcal{X}^m} \mathcal{N}(\epsilon, \mathcal{F}_{|x_1^m}, d_p) \tag{7}$$

if $\mathcal{N}(\epsilon, \mathcal{F}_{|x_1^m}, d_p)$ is finite for all $x_1^m \in \mathcal{X}^m$, and $\mathcal{N}_p(\epsilon, \mathcal{F}, m) = \infty$ otherwise. This should be compared with the definition of growth function for a class of binary-valued functions $\mathcal{H}$, which also involved a maximum over $x_1^m \in \mathcal{X}^m$: in that case, $\mathcal{H}_{|x_1^m}$ was finite, and the maximum was over the *cardinality* of $\mathcal{H}_{|x_1^m}$; here, $\mathcal{F}_{|x_1^m}$ may in general be infinite, and the maximum is over the 'extent' of $\mathcal{F}_{|x_1^m}$ in $\mathbb{R}^m$ at scale ϵ, as measured using the metric d_p. In particular, for $\mathcal{H} \subseteq \{-1, 1\}^{\mathcal{X}}$, we have that for any $\epsilon \le 2$, $\mathcal{N}(\epsilon, \mathcal{H}_{|x_1^m}, d_\infty) = |\mathcal{H}_{|x_1^m}|$, and therefore $\mathcal{N}_\infty(\epsilon, \mathcal{H}, m) = \Pi_{\mathcal{H}}(m)$. Thus the uniform covering numbers can be viewed as generalizing the notion of growth function to classes of real-valued functions.

Note that the term 'uniform' here refers to the maximum over all $x_1^m \in \mathcal{X}^m$ (of the covering numbers of $\mathcal{F}_{|x_1^m}$ in $\mathbb{R}^m$), and is unrelated to the use of the term 'uniform' in 'uniform convergence', which refers to the supremum over functions $f \in \mathcal{F}$. Unless otherwise stated, in what follows we will refer to the uniform covering numbers of a function class $\mathcal{F}$ as simply the *covering numbers* of $\mathcal{F}$.

3 Uniform Convergence in a Real-Valued Function class $\mathcal{F}$

We can assume that functions in $\mathcal{F}$ take values in some set $\widehat{\mathcal{Y}} \subseteq \mathbb{R}$, so that $\mathcal{F} \subseteq \widehat{\mathcal{Y}}^{\mathcal{X}}$. We will require the loss function ℓ to be bounded, i.e. we will assume $\exists\, B > 0$ such that $0 \le \ell(y, \hat{y}) \le B \;\forall\; y \in \mathcal{Y}, \hat{y} \in \widehat{\mathcal{Y}}$. We will find it useful to define for any function class $\mathcal{F} \subseteq \widehat{\mathcal{Y}}^{\mathcal{X}}$ and loss $\ell : \mathcal{Y} \times \widehat{\mathcal{Y}} \to [0, B]$ the *loss function class* $\ell_{\mathcal{F}} \subseteq [0, B]^{\mathcal{X} \times \mathcal{Y}}$ given by

$$\ell_{\mathcal{F}} \;=\; \Big\{ \ell_f : \mathcal{X} \times \mathcal{Y} \to [0, B] \;\Big|\; \ell_f(x, y) = \ell(y, f(x)) \text{ for some } f \in \mathcal{F} \Big\}\,. \tag{8}$$

We will first prove a uniform convergence result for general losses ℓ as above in terms of the d_1 covering numbers of the loss function class $\ell_{\mathcal{F}}$, and will then show that for many losses ℓ, including the squared loss when $\mathcal{Y}$ and $\widehat{\mathcal{Y}}$ are bounded, the d_1 covering numbers of $\ell_{\mathcal{F}}$ can further be bounded in terms of the d_1 covering numbers of $\mathcal{F}$.

Theorem 3.1. Let $\mathcal{Y}, \widehat{\mathcal{Y}} \subseteq \mathbb{R}$.[4] Let $\mathcal{F} \subseteq \widehat{\mathcal{Y}}^{\mathcal{X}}$, and let $\ell : \mathcal{Y} \times \widehat{\mathcal{Y}} \to [0, B]$. Let D be any distribution on $\mathcal{X} \times \mathcal{Y}$. For any $\epsilon > 0$,

$$\mathbf{P}_{S \sim D^m}\left(\sup_{f \in \mathcal{F}}\left|\mathrm{er}_D^\ell[f] - \mathrm{er}_S^\ell[f]\right| \geq \epsilon\right) \;\leq\; 4\mathcal{N}_1\big(\epsilon/8, \ell_{\mathcal{F}}, 2m\big)\, e^{-m\epsilon^2/32B^2}\,. \tag{9}$$

Proof. The proof uses similar techniques as in the proof of uniform convergence for the $\ell_{0\text{-}1}$ loss in the binary case that we saw in Lecture 3, and has the same 4 broad steps. The key difference is in the reduction to a finite class (step 3).

Step 1: Symmetrization. Following the same steps as in Lecture 3, we can show that for $m\epsilon^2 \geq 8B^2$,

$$\mathbf{P}_{S \sim D^m}\left(\sup_{f \in \mathcal{F}}\left|\mathrm{er}_D^\ell[h] - \mathrm{er}_S^\ell[h]\right| \geq \epsilon\right) \;\leq\; 2\,\mathbf{P}_{(S,\tilde{S}) \sim D^m \times D^m}\left(\sup_{f \in \mathcal{F}}\left|\mathrm{er}_S^\ell[f] - \mathrm{er}_{\tilde{S}}^\ell[f]\right| \geq \frac{\epsilon}{2}\right). \tag{10}$$

Step 2: Swapping permutations. Again using the same argument as in Lecture 3, we can show that

$$\begin{aligned}
&\mathbf{P}_{(S,\tilde{S}) \sim D^m \times D^m}\left(\sup_{f \in \mathcal{F}}\left|\mathrm{er}_S^\ell[f] - \mathrm{er}_{\tilde{S}}^\ell[f]\right| \geq \frac{\epsilon}{2}\right) \\
&\leq \sup_{(S,\tilde{S}) \in (\mathcal{X} \times \mathcal{Y})^{2m}}\left[\mathbf{P}_{\sigma \in \Gamma_{2m}}\left(\sup_{f \in \mathcal{F}}\left|\mathrm{er}_{\sigma(S)}^\ell[f] - \mathrm{er}_{\sigma(\tilde{S})}^\ell[f]\right| \geq \frac{\epsilon}{2}\right)\right]. \qquad (11)
\end{aligned}$$

Step 3: Reduction to a finite class. Fix any $(S, \tilde{S}) \in (\mathcal{X} \times \mathcal{Y})^{2m}$, and for simplicity, define $(x_{m+i}, y_{m+i}) = (\tilde{x}_i, \tilde{y}_i)\ \forall\ i \in [m]$. Now consider $(\ell_{\mathcal{F}})_{|(S,\tilde{S})} \in [0, B]^{2m}$. Let $\mathcal{G} \subseteq \mathcal{F}$ be such that $(\ell_{\mathcal{G}})_{|(S,\tilde{S})}$ is an $\epsilon/8$-cover of $(\ell_{\mathcal{F}})_{|(S,\tilde{S})}$ w.r.t. d_1. Clearly, we can take $|\mathcal{G}| = \mathcal{N}\Big(\epsilon/8, (\ell_{\mathcal{F}})_{|(S,\tilde{S})}, d_1\Big) \leq \mathcal{N}_1(\epsilon/8, \ell_{\mathcal{F}}, 2m)$; since $\ell_{\mathcal{F}}$ maps to a bounded interval, this is a finite number. Now consider any $\sigma \in \Gamma_{2m}$. We claim that if $\exists\ f \in \mathcal{F}$ such that

$$\left|\mathrm{er}_{\sigma(S)}^\ell[f] - \mathrm{er}_{\sigma(\tilde{S})}^\ell[f]\right| \geq \frac{\epsilon}{2}\,, \tag{12}$$

then $\exists\ g \in \mathcal{G}$ such that

$$\left|\mathrm{er}_{\sigma(S)}^\ell[g] - \mathrm{er}_{\sigma(\tilde{S})}^\ell[g]\right| \geq \frac{\epsilon}{4}\,. \tag{13}$$

To see this, let $f \in \mathcal{F}$ be such that (12) holds. Take any $g \in \mathcal{G}$ for which

$$\frac{1}{2m}\sum_{i=1}^{2m}\left|\ell_f(x_i, y_i) - \ell_g(x_i, y_i)\right| < \frac{\epsilon}{8}\,. \tag{14}$$

Such a g exists since $(\ell_{\mathcal{G}})_{(S,\tilde{S})}$ is an $\epsilon/8$-cover of $(\ell_{\mathcal{F}})_{(S,\tilde{S})}$ w.r.t. d_1. We will show that g satisfies (13). In particular, we have

$$\begin{aligned}
\frac{\epsilon}{2} &\leq \left|\mathrm{er}_{\sigma(S)}^\ell[f] - \mathrm{er}_{\sigma(\tilde{S})}^\ell[f]\right| && (15)\\
&= \left|\Big(\mathrm{er}_{\sigma(S)}^\ell[f] - \mathrm{er}_{\sigma(S)}^\ell[g]\Big) - \Big(\mathrm{er}_{\sigma(\tilde{S})}^\ell[f] - \mathrm{er}_{\sigma(\tilde{S})}^\ell[g]\Big) + \Big(\mathrm{er}_{\sigma(S)}^\ell[g] - \mathrm{er}_{\sigma(\tilde{S})}^\ell[g]\Big)\right| && (16)\\
&\leq \left|\mathrm{er}_{\sigma(S)}^\ell[f] - \mathrm{er}_{\sigma(S)}^\ell[g]\right| + \left|\mathrm{er}_{\sigma(\tilde{S})}^\ell[f] - \mathrm{er}_{\sigma(\tilde{S})}^\ell[g]\right| + \left|\mathrm{er}_{\sigma(S)}^\ell[g] - \mathrm{er}_{\sigma(\tilde{S})}^\ell[g]\right| && (17)\\
&= \left|\mathrm{er}_{\sigma(S)}^\ell[g] - \mathrm{er}_{\sigma(\tilde{S})}^\ell[g]\right| + \left|\frac{1}{m}\sum_{i=1}^{m}\Big(\ell_f(x_i, y_i) - \ell_g(x_i, y_i)\Big)\right| + \left|\frac{1}{m}\sum_{i=m+1}^{2m}\Big(\ell_f(x_{\sigma(i)}, y_{\sigma(i)}) - \ell_g(x_{\sigma(i)}, y_{\sigma(i)})\Big)\right| && (18)\\
&\leq \left|\mathrm{er}_{\sigma(S)}^\ell[g] - \mathrm{er}_{\sigma(\tilde{S})}^\ell[g]\right| + \frac{1}{m}\sum_{i=1}^{2m}\left|\ell_f(x_{\sigma(i)}, y_{\sigma(i)}) - \ell_g(x_{\sigma(i)}, y_{\sigma(i)})\right| && (19)\\
&< \left|\mathrm{er}_{\sigma(S)}^\ell[g] - \mathrm{er}_{\sigma(\tilde{S})}^\ell[g]\right| + \frac{\epsilon}{4} \quad \text{(by (14))}. && (20)
\end{aligned}$$

[4]$\mathcal{Y}$ can be thought of as the space of 'true labels', and $\widehat{\mathcal{Y}}$ the space of 'predicted labels'.

The claim follows. Thus we have

$$
\begin{aligned}
&\mathbf{P}_{\sigma\in\Gamma_{2m}}\left(\sup_{f\in\mathcal{F}}\left|\mathrm{er}^{\ell}_{\sigma(S)}[f]-\mathrm{er}^{\ell}_{\sigma(\tilde{S})}[f]\right|\geq\frac{\epsilon}{2}\right) \\
&\leq\ \mathbf{P}_{\sigma\in\Gamma_{2m}}\left(\max_{g\in\mathcal{G}}\left|\mathrm{er}^{\ell}_{\sigma(S)}[g]-\mathrm{er}^{\ell}_{\sigma(\tilde{S})}[g]\right|\geq\frac{\epsilon}{4}\right) && (21)\\
&\leq\ \mathcal{N}_1(\epsilon/8,\ell_{\mathcal{F}},2m)\ \max_{g\in\mathcal{G}}\mathbf{P}_{\sigma\in\Gamma_{2m}}\left(\left|\mathrm{er}^{\ell}_{\sigma(S)}[g]-\mathrm{er}^{\ell}_{\sigma(\tilde{S})}[g]\right|\geq\frac{\epsilon}{4}\right)\quad\text{(by union bound).} && (22)
\end{aligned}
$$

Step 4: Hoeffding's inequality. As in Lecture 3, Hoeffding's inequality can now be used to show that for any $g\in\mathcal{G}$,

$$
\begin{aligned}
&\mathbf{P}_{\sigma\in\Gamma_{2m}}\left(\left|\mathrm{er}^{\ell}_{\sigma(S)}[g]-\mathrm{er}^{\ell}_{\sigma(\tilde{S})}[g]\right|\geq\frac{\epsilon}{4}\right) && (23)\\
&=\ \mathbf{P}_{\mathbf{r}\in\{-1,1\}^m}\left(\frac{1}{m}\left|\sum_{i=1}^{m}r_i\Big(\ell_g(x_i,y_i)-\ell_g(x_i,y_i)\Big)\right|\geq\frac{\epsilon}{4}\right) && (24)\\
&\leq\ 2\,e^{-m\epsilon^2/32B^2}\,. && (25)
\end{aligned}
$$

Putting everything together yields the desired result for $m\epsilon^2\geq 8B^2$; for $m\epsilon^2<8B^2$, the result holds trivially. □

The above result yields a high-confidence bound on the generalization error of a function learned from $\mathcal{F}$ in terms of covering numbers of $\ell_{\mathcal{F}}$. For 'well-behaved' loss functions ℓ, these can be further bounded in terms of covering numbers of $\mathcal{F}$:

Lemma 3.2. Let $\mathcal{Y},\widehat{\mathcal{Y}}\subseteq\mathbb{R}$. Let $\mathcal{F}\subseteq\widehat{\mathcal{Y}}^{\mathcal{X}}$, and let $\ell:\mathcal{Y}\times\widehat{\mathcal{Y}}\to[0,B]$. If ℓ is Lipschitz in its second argument with Lipschitz constant $L>0$, i.e.

$$
|\ell(y,\hat{y}_1)-\ell(y,\hat{y}_2)|\ \leq\ L\,|\hat{y}_1-\hat{y}_2|\quad\forall\ y\in\mathcal{Y},\ \hat{y}_1,\hat{y}_2\in\widehat{\mathcal{Y}}\,,
$$

then for any $m\in\mathbb{N}$,

$$
\mathcal{N}_1\big(\epsilon,\ell_{\mathcal{F}},m\big)\ \leq\ \mathcal{N}_1\big(\epsilon/L,\mathcal{F},m\big)\,.
$$

Proof. Let $S=((x_1,y_1),\ldots,(x_m,y_m))\in(\mathcal{X}\times\mathcal{Y})^m$, and let $f,g\in\mathcal{F}$. Then

$$
\begin{aligned}
\frac{1}{m}\sum_{i=1}^{m}\left|\ell_f(x_i,y_i)-\ell_g(x_i,y_i)\right| &=\ \frac{1}{m}\sum_{i=1}^{m}\left|\ell(y_i,f(x_i))-\ell(y_i,g(x_i))\right| && (26)\\
&\leq\ \frac{L}{m}\sum_{i=1}^{m}\left|f(x_i)-g(x_i)\right|. && (27)
\end{aligned}
$$

Thus any d_1 ϵ/L-cover for $\mathcal{F}_{|x_1^m}$ is a d_1 ϵ-cover for $(\ell_{\mathcal{F}})_{|S}$, which implies the result. □

This yields the following corollary:

Corollary 3.3. Let $\mathcal{Y},\widehat{\mathcal{Y}}\subseteq\mathbb{R}$, $\mathcal{F}\subseteq\widehat{\mathcal{Y}}^{\mathcal{X}}$, and $\ell:\mathcal{Y}\times\widehat{\mathcal{Y}}\to[0,B]$ such that ℓ is Lipschitz in its second argument with Lipschitz constant $L>0$. Let D be any distribution on $\mathcal{X}\times\mathcal{Y}$. Then for any $\epsilon>0$:

$$
\mathbf{P}_{S\sim D^m}\left(\sup_{f\in\mathcal{F}}\left|\mathrm{er}^{\ell}_{D}[f]-\mathrm{er}^{\ell}_{S}[f]\right|\geq\epsilon\right)\ \leq\ 4\mathcal{N}_1\big(\epsilon/8L,\mathcal{F},2m\big)\,e^{-m\epsilon^2/32B^2}\,. \qquad (28)
$$

As an example, consider the squared loss $\ell_{\mathrm{sq}}(y,\hat{y})=(\hat{y}-y)^2$. It can be shown that if $\mathcal{Y},\widehat{\mathcal{Y}}$ are bounded, then ℓ_{sq} is bounded and Lipschitz. In particular, for $\mathcal{Y}=\widehat{\mathcal{Y}}=[-1,1]$, we have $0\leq\ell_{\mathrm{sq}}(y,\hat{y})\leq 4\ \forall\ y\in\mathcal{Y},\hat{y}\in\widehat{\mathcal{Y}}$, and ℓ_{sq} is Lipschitz with Lipschitz constant $L = 4$:

$$\begin{aligned}
\left|\ell_{\text{sq}}(y, \hat{y}_1) - \ell_{\text{sq}}(y, \hat{y}_1)\right| &= \left|(y - \hat{y}_1)^2 - (y - \hat{y}_2)^2\right| && (29)\\
&= \left|\hat{y}_1^2 - \hat{y}_2^2 - 2y(\hat{y}_1 - \hat{y}_2)\right| && (30)\\
&\leq |\hat{y}_1 + \hat{y}_2|\,|\hat{y}_1 - \hat{y}_2| + 2|y|\,|\hat{y}_1 - \hat{y}_2| && (31)\\
&\leq 4\,|\hat{y}_1 - \hat{y}_2| \quad (\text{since } y, \hat{y}_1, \hat{y}_2 \in [-1, 1]). && (32)
\end{aligned}$$

Thus when both labels and predictions are in $[-1, 1]$, one gets for the squared loss:

Corollary 3.4. Let $\mathcal{Y} = \widehat{\mathcal{Y}} = [-1, 1]$, $\mathcal{F} \subseteq \widehat{\mathcal{Y}}^{\mathcal{X}}$, and $\ell_{\text{sq}} : \mathcal{Y} \times \widehat{\mathcal{Y}} \to [0, 4]$ be given by $\ell_{\text{sq}}(y, \hat{y}) = (\hat{y} - y)^2$. Let D be any distribution on $\mathcal{X} \times \mathcal{Y}$. Then for any $\epsilon > 0$:

$$\mathbf{P}_{S \sim D^m}\left(\sup_{f \in \mathcal{F}}\left|\text{er}_D^{\text{sq}}[f] - \text{er}_S^{\text{sq}}[f]\right| \geq \epsilon\right) \;\leq\; 4\mathcal{N}_1\big(\epsilon/32, \mathcal{F}, 2m\big)\, e^{-m\epsilon^2/512}. \tag{33}$$

Exercise. Show that for $\mathcal{Y} = \widehat{\mathcal{Y}} = [-1, 1]$, the absolute loss given by $\ell_{\text{abs}}(y, \hat{y}) = |\hat{y} - y| \; \forall\, y \in \mathcal{Y}, \hat{y} \in \widehat{\mathcal{Y}}$ is bounded and is Lipschitz in its second argument with Lipschitz constant $L = 1$.

4 Pseudo-Dimension and Fat-Shattering Dimension

Just as the growth function $\Pi_{\mathcal{H}}(2m)$ needed to be sub-exponential in m for the uniform convergence result in the binary classification case to be meaningful, the covering numbers $\mathcal{N}_1(\epsilon/8, \ell_{\mathcal{F}}, 2m)$ or $\mathcal{N}_1(\epsilon/8L, \mathcal{F}, 2m)$ need to be sub-exponential in m for the above result to be meaningful. In the binary case, we saw that if the VC-dimension of $\mathcal{H}$ is finite, then the growth function of $\mathcal{H}$ grows polynomially in m. Analogous results can be shown to hold for the covering numbers. We provide some basic definitions and results here; further details can be found for example in [1].

Definition (Pseudo-dimension)**.** Let $\mathcal{F} \subseteq \mathbb{R}^{\mathcal{X}}$ and let $x_1^m = (x_1, \ldots, x_m) \in \mathcal{X}^m$. We say x_1^m is *pseudo-shattered* by $\mathcal{F}$ if $\exists\; \mathbf{r} = (r_1, \ldots, r_m) \in \mathbb{R}^m$ such that $\forall\; \mathbf{b} = (b_1, \ldots, b_m) \in \{-1, 1\}^m$, $\exists\; f_{\mathbf{b}} \in \mathcal{F}$ such that $\text{sign}(f_{\mathbf{b}}(x_i) - r_i) = b_i \;\forall\; i \in [m]$. The *pseudo-dimension* of $\mathcal{F}$ is the cardinality of the largest set of points in $\mathcal{X}$ that can be pseudo-shattered by $\mathcal{F}$:

$$\text{Pdim}(\mathcal{F}) = \max\left\{m \in \mathbb{N} \;\middle|\; \exists\; x_1^m \in \mathcal{X}^m \text{ such that } x_1^m \text{ is pseudo-shattered by } \mathcal{F}\right\}.$$

If $\mathcal{F}$ pseudo-shatters arbitrarily large sets of points in $\mathcal{X}$, we say $\text{Pdim}(\mathcal{F}) = \infty$.

Fact. If $\mathcal{F}$ is a vector space of real-valued functions, then $\text{Pdim}(\mathcal{F}) = \dim(\mathcal{F})$. For example, for the class of all affine functions over $\mathbb{R}^n$ given by $\mathcal{F} = \{f : \mathbb{R}^n \to \mathbb{R} \mid f(\mathbf{x}) = \mathbf{w} \cdot \mathbf{x} + b \text{ for some } \mathbf{w} \in \mathbb{R}^n, b \in \mathbb{R}\}$, we have $\text{Pdim}(\mathcal{F}) = n + 1$. Clearly, if $\mathcal{F}' \subseteq \mathcal{F}$, $\text{Pdim}(\mathcal{F}') \leq \text{Pdim}(\mathcal{F})$.

Definition (Fat-shattering dimension)**.** Let $\mathcal{F} \subseteq \mathbb{R}^{\mathcal{X}}$ and let $x_1^m = (x_1, \ldots, x_m) \in \mathcal{X}^m$. Let $\gamma > 0$. We say x_1^m is *γ-shattered* by $\mathcal{F}$ if $\exists\; \mathbf{r} = (r_1, \ldots, r_m) \in \mathbb{R}^m$ such that $\forall\; \mathbf{b} = (b_1, \ldots, b_m) \in \{-1, 1\}^m$, $\exists\; f_{\mathbf{b}} \in \mathcal{F}$ such that $b_i(f_{\mathbf{b}}(x_i) - r_i) \geq \gamma \;\forall\; i \in [m]$. The *$\gamma$-dimension* of $\mathcal{F}$ or the *fat-shattering dimension* of $\mathcal{F}$ at scale γ is the cardinality of the largest set of points in $\mathcal{X}$ that can be γ-shattered by $\mathcal{F}$:

$$\text{fat}_{\mathcal{F}}(\gamma) = \max\left\{m \in \mathbb{N} \;\middle|\; \exists\; x_1^m \in \mathcal{X}^m \text{ such that } x_1^m \text{ is } \gamma\text{-shattered by } \mathcal{F}\right\}.$$

If $\mathcal{F}$ γ-shatters arbitrarily large sets of points in $\mathcal{X}$, we say $\text{fat}_{\mathcal{F}}(\gamma) = \infty$.

Clearly, $\text{fat}_F(\gamma) \leq \text{Pdim}(\mathcal{F}) \;\forall\; \gamma > 0$. The fat-shattering dimension is often called a *scale-sensitive* dimension since it depends on the scale γ. Both quantities, when finite, can be used to bound the covering numbers of a function class $\mathcal{F}$ whose functions take values in a bounded range:

Theorem 4.1. Let $\mathcal{F} \subseteq [a,b]^{\mathcal{X}}$ for some $a \leq b$. Let $0 < \epsilon \leq b - a$, and let $\mathrm{fat}_{\mathcal{F}}(\epsilon/8) = d < \infty$. Then for $m \geq d \geq 1$,

$$\mathcal{N}_1(\epsilon, \mathcal{F}, m) = O\left(\left(\frac{1}{\epsilon}\right)^{d \log_2(m/\epsilon d)}\right).$$

Theorem 4.2. Let $\mathcal{F} \subseteq [a,b]^{\mathcal{X}}$ for some $a \leq b$. Let $\mathrm{Pdim}(\mathcal{F}) = d < \infty$. Then for all $0 < \epsilon \leq b - a$ and $m \in \mathbb{N}$,

$$\mathcal{N}_1(\epsilon, \mathcal{F}, m) = O\left(\left(\frac{1}{\epsilon}\right)^{d}\right).$$

5 Next Lecture

In the next lecture, we will return to binary classification, but will focus on learning functions of the form $h(x) = \mathrm{sign}(f(x))$ for some real-valued function f, and will see how the quantities considered in this lecture can be useful in such situations.

References

[1] Martin Anthony and Peter L. Bartlett. *Learning in Neural Networks: Theoretical Foundations.* Cambridge University Press, 1999.