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

1 Introduction

So far we have seen how to obtain high confidence bounds on the generalization error $\er(h_S, \mathcal{H}, D)$ of a binary classifier $h_S$ learned by an algorithm from a function class $\mathcal{H} \subseteq \{-1, 1\}^X$ of limited capacity, using the ideas of uniform convergence. We saw the use of the growth function $\Pi_{\mathcal{H}}$ to measure the capacity of the class $\mathcal{H}$, as well as the VC-dimension $\text{VCdim}(\mathcal{H})$, 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 $S = \{(x_1, y_1), \ldots, (x_m, y_m)\} \in (X \times \mathcal{Y})^m$ for some $\mathcal{Y} \subseteq \mathbb{R}$, assumed now to be drawn from $D^m$ for some distribution $D$ on $X \times \mathcal{Y}$, and the goal is to learn from this sample a real-valued function $f_S : X \rightarrow \mathbb{R}$ that has low generalization error $\er_D[f_S]$ w.r.t. some appropriate loss function $\ell : \mathcal{Y} \times \mathbb{R} \rightarrow [0, \infty)$. We have seen how to obtain high confidence bounds on the generalization error of the learned function, $\er_D[f_S]$. Again, we will consider learning algorithms that learn $f_S$ from a function class $\mathcal{F} \subseteq \mathbb{R}^X$ of limited capacity, and then bound generalization error bounds for such algorithms via a uniform convergence result that upper bounds the probability

$$P_{S \sim D^m} \left( \sup_{f \in \mathcal{F}} \left| \er_D[f] - \er_S[f] \right| \geq \epsilon \right).$$

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) $\epsilon$-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 $\epsilon$-cover of $W$ w.r.t. $d$ if the union of (open) $d$-balls of radius $\epsilon$ centered at points in $C$ contains $W$:³

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

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

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

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

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 \( \epsilon \)-cover w.r.t. \( d \), we take \( N(\epsilon, W, d) = \infty \). Thus the covering numbers of \( W \) can be viewed as measuring the ‘extent’ of \( W \) in \((A, d)\) at ‘granularity’ or ‘scale’ \( \epsilon \).

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:

\[
d_1(x, x') = \frac{1}{n} \sum_{i=1}^{n} |x_i - x'_i|
\]

(3)

\[
d_2(x, x') = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (x_i - x'_i)^2}
\]

(4)

\[
d_{\infty}(x, x') = \max_{i} |x_i - x'_i|.
\]

(5)

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

\[
N(\epsilon, W, d_1) \leq N(\epsilon, W, d_2) \leq N(\epsilon, W, d_{\infty}).
\]

(6)

2.3 Uniform Covering Numbers for a Real-Valued Function Class

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

\[
N_p(\epsilon, F, m) = \max_{x_1^m \in \mathcal{X}^m} N(\epsilon, F_{|x_1^m}, d_p)
\]

(7)

if \( N(\epsilon, F_{|x_1^m}, d_p) \) is finite for all \( x_1^m \in \mathcal{X}^m \), and \( N_p(\epsilon, F, m) = \infty \) otherwise. This should be compared with the definition of growth function for a class of binary-valued functions \( H \), which also involved a maximum over \( x_1^m \in \mathcal{X}^m \): in that case, \( H_{|x_1^m} \) was finite, and the maximum was over the cardinality of \( H_{|x_1^m} \); here, \( F_{|x_1^m} \) may in general be infinite, and the maximum is over the ‘extent’ of \( F_{|x_1^m} \) in \( \mathbb{R}^m \) at scale \( \epsilon \), as measured using the metric \( d_p \). In particular, for \( H \subseteq \{-1, 1\}^X \), we have that for any \( \epsilon \leq 2 \), \( N(\epsilon, H_{|x_1^m}, d_{\infty}) = |H_{|x_1^m}| \), and therefore \( N_p(\epsilon, H, m) = \Pi_{H}(m) \). 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 \( F_{|x_1^m} \) in \( \mathbb{R}^m \)), and is unrelated to the use of the term ‘uniform’ in ‘uniform convergence’, which refers to the supremum over functions \( f \in F \). Unless otherwise stated, in what follows we will refer to the uniform covering numbers of a function class \( F \) as simply the covering numbers of \( F \).

3 Uniform Convergence in a Real-Valued Function class \( F \)

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

\[
\ell_F = \left\{ \ell_f : X \times \mathcal{Y} \rightarrow [0, B] \bigg| \ell_f(x, y) = \ell(y, f(x)) \text{ for some } f \in F \right\}.
\]

(8)

We will first prove a uniform convergence result for general losses \( \ell \) as above in terms of the \( d_1 \) covering numbers of the loss function class \( \ell_F \), and will then show that for many losses \( \ell \), including the squared loss when \( \mathcal{Y} \) and \( \hat{\mathcal{Y}} \) are bounded, the \( d_1 \) covering numbers of \( \ell_F \) can further be bounded in terms of the \( d_1 \) covering numbers of \( F \).
Theorem 3.1. Let $\mathcal{Y}, \mathcal{Y}^\prime \subseteq \mathbb{R}^4$. Let $\mathcal{F} \subseteq \mathcal{Y}^\times$, and let $\ell: \mathcal{Y} \times \mathcal{Y}^\prime \rightarrow [0, B]$. Let $D$ be any distribution on $\mathcal{X} \times \mathcal{Y}$. For any $\epsilon > 0$,

$$P_{S \sim D^m} \left( \sup_{f \in \mathcal{F}} \left| \epsilon_r f_D[f] - \epsilon_r f_S[f] \right| \geq \epsilon \right) \leq 4N_1(\epsilon/2, \ell, 2m) e^{-m\epsilon^2/2B^2}.$$ (9)

**Proof.** The proof uses similar techniques as in the proof of uniform convergence for the $\ell_{0,1}$ 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 $m\epsilon^2 \geq 8B^2$,

$$P_{S \sim D^m} \left( \sup_{f \in \mathcal{F}} \left| \epsilon_r f_D[h] - \epsilon_r f_S[h] \right| \geq \epsilon \right) \leq 2P_{(S, \tilde{S}) \sim D^m \times D^m} \left( \sup_{f \in \mathcal{F}} \left| \epsilon_r f_S[f] - \epsilon_r f_{\tilde{S}}[f] \right| \geq \frac{\epsilon}{2} \right).$$ (10)

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

$$P_{(S, \tilde{S}) \sim D^m \times D^m} \left( \sup_{f \in \mathcal{F}} \left| \epsilon_r f_S[f] - \epsilon_r f_{\tilde{S}}[f] \right| \geq \frac{\epsilon}{2} \right) \leq \sup_{(S, \tilde{S}) \in (\mathcal{X} \times \mathcal{Y})^{2m}} \left[ P_{\sigma \in \Gamma_{2m}} \left( \sup_{f \in \mathcal{F}} \left| \epsilon_r f_{\sigma[S]}[f] - \epsilon_r f_{\sigma[\tilde{S}]}[f] \right| \geq \frac{\epsilon}{2} \right) \right].$$ (11)

**Step 3: Reduction to a finite class.** Fix any $(S, \tilde{S}) \in (\mathcal{X} \times \mathcal{Y})^{2m}$, and for simplicity, define $(x_{m+1}, y_{m+1}) = (\tilde{x}_i, \tilde{y}_i) \forall i \in [m]$. Now consider $(\ell_x)(S, \tilde{S}) \in [0, B]^{2m}$. Let $\mathcal{G} \subseteq \mathcal{F}$ be such that $(\ell_x)(S, \tilde{S})$ is an $\epsilon/8$-cover of $(\ell_x)(S, \tilde{S})$ w.r.t. $d_1$. Clearly, we can take $|\mathcal{G}| = N(\epsilon/8, (\ell_x)(S, \tilde{S}), d_1) \leq N_1(\epsilon/8, \ell_x, 2m)$; since $\ell_x$ maps to a bounded interval, this is a finite number. Now consider any $\sigma \in \Gamma_{2m}$. We claim that if $f \in \mathcal{F}$ such that

$$\left| \epsilon_r f_{\sigma[S]}[f] - \epsilon_r f_{\sigma[\tilde{S}]}[f] \right| \geq \frac{\epsilon}{2},$$ (12)

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

$$\left| \epsilon_r f_{\sigma[S]}[g] - \epsilon_r f_{\sigma[\tilde{S}]}[g] \right| \geq \frac{\epsilon}{4}.$$ (13)

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

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

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

$$\begin{align*}
\frac{\epsilon}{4} &\leq \left| \epsilon_r f_{\sigma[S]}[g] - \epsilon_r f_{\sigma[\tilde{S}]}[g] \right| \\
&\leq \left| \left( \epsilon_r f_{\sigma[S]}[f] - \epsilon_r f_{\sigma[\tilde{S}]}[g] \right) - \left( \epsilon_r f_{\sigma[S]}[f] - \epsilon_r f_{\sigma[\tilde{S}]}[g] \right) + \left( \epsilon_r f_{\sigma[S]}[g] - \epsilon_r f_{\sigma[\tilde{S}]}[g] \right) \right| \\
&\leq \left| \epsilon_r f_{\sigma[S]}[f] - \epsilon_r f_{\sigma[\tilde{S}]}[g] \right| + \left| \epsilon_r f_{\sigma[S]}[f] - \epsilon_r f_{\sigma[\tilde{S}]}[g] \right| + \left| \epsilon_r f_{\sigma[S]}[g] - \epsilon_r f_{\sigma[\tilde{S}]}[g] \right| \\
&= \left| \epsilon_r f_{\sigma[S]}[g] - \epsilon_r f_{\sigma[\tilde{S}]}[g] \right| + \frac{1}{m} \sum_{i=1}^{m} \left( \ell_f(x_i, y_i) - \ell_g(x_i, y_i) \right) + \frac{1}{m} \sum_{i=m+1}^{2m} \left( \ell_f(x_{i+1}, y_{i+1}) - \ell_g(x_{i+1}, y_{i+1}) \right) \\
&\leq \left| \epsilon_r f_{\sigma[S]}[g] - \epsilon_r f_{\sigma[\tilde{S}]}[g] \right| + \frac{1}{m} \sum_{i=1}^{m} \left( \ell_f(x_{i(i)}, y_{i(i)}) - \ell_g(x_{i(i)}, y_{i(i)}) \right) \\
&\leq \left| \epsilon_r f_{\sigma[S]}[g] - \epsilon_r f_{\sigma[\tilde{S}]}[g] \right| + \frac{1}{4} \epsilon (by \ (14)).
\end{align*}$$ (20)

$\mathcal{Y}$ can be thought of as the space of ‘true labels’, and $\mathcal{Y}^\prime$ the space of ‘predicted labels’.
The claim follows. Thus we have

\[
P_{\sigma \in \Gamma_{2m}} \left( \sup_{f \in \mathcal{F}} \left| \mathcal{E}_{\sigma}(S)[f] - \mathcal{E}_{\sigma}(\hat{S})[f] \right| \geq \frac{\epsilon}{2} \right) \leq \mathcal{N}_1(\epsilon/8, \ell_x, 2m) \max_{g \in \mathcal{G}} P_{\sigma \in \Gamma_{2m}} \left( \left| \mathcal{E}_{\sigma}(S)[g] - \mathcal{E}_{\sigma}(\hat{S})[g] \right| \geq \frac{\epsilon}{4} \right) \tag{21}\]

which yields the following corollary:

\[
P_{\sigma \in \Gamma_{2m}} \left( \sup_{f \in \mathcal{F}} \left| \mathcal{E}_{\sigma}(S)[f] - \mathcal{E}_{\sigma}(\hat{S})[f] \right| \geq \frac{\epsilon}{2} \right) \leq \mathcal{N}_1(\epsilon/4, \ell_x, m) \max_{g \in \mathcal{G}} P_{\sigma \in \Gamma_{2m}} \left( \left| \mathcal{E}_{\sigma}(S)[g] - \mathcal{E}_{\sigma}(\hat{S})[g] \right| \geq \frac{\epsilon}{4} \right) \tag{22}\]

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

\[
P_{\sigma \in \Gamma_{2m}} \left( \left| \mathcal{E}_{\sigma}(S)[g] - \mathcal{E}_{\sigma}(\hat{S})[g] \right| \geq \frac{\epsilon}{4} \right) \leq 2e^{-m \epsilon^2/32B^2} \tag{23}\]

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 \( \ell_x \). For ‘well-behaved’ loss functions \( \ell \), these can be further bounded in terms of covering numbers of \( \mathcal{F} \):

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

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

then for any \( m \in \mathbb{N} \),

\[
\mathcal{N}_1(\epsilon, \ell_x, m) \leq \mathcal{N}_1(\epsilon/L, \mathcal{F}, m).
\]

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

\[
\frac{1}{m} \sum_{i=1}^m |\ell_f(x_i, y_i) - \ell_g(x_i, y_i)| = \frac{1}{m} \sum_{i=1}^m |\ell(y_i, f(x_i)) - \ell(y_i, g(x_i))| \leq \frac{L}{m} \sum_{i=1}^m |f(x_i) - g(x_i)|.
\]

Thus any \( d_1 \epsilon/L \)-cover for \( \mathcal{F}_{x_1} \) is a \( d_1 \epsilon \)-cover for \( (\ell_x)_S \), which implies the result.

This yields the following corollary:

Corollary 3.3. Let \( \mathcal{Y}, \hat{\mathcal{Y}} \subseteq \mathbb{R} \), \( \mathcal{F} \subseteq \hat{\mathcal{Y}}^X \), and \( \ell : \mathcal{Y} \times \hat{\mathcal{Y}} \rightarrow [0, B] \) such that \( \ell \) 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 \( \epsilon > 0 \):

\[
P_{\mathcal{S} \sim D^m} \left( \sup_{f \in \mathcal{F}} \left| \mathcal{E}_{\mathcal{S}}[f] - \mathcal{E}_{\mathcal{S}}[\hat{f}] \right| \geq \epsilon \right) \leq 4\mathcal{N}_1(\epsilon/8L, \mathcal{F}, 2m) e^{-m \epsilon^2/32B^2} \tag{28}\]

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

$$
\left| \ell_{sq}(y, \hat{y}_1) - \ell_{sq}(y, \hat{y}_2) \right| = \left| (y - \hat{y}_1)^2 - (y - \hat{y}_2)^2 \right| = \left| \hat{y}_1^2 - \hat{y}_2^2 - 2\hat{y}_1(y - \hat{y}_2) \right| \leq |\hat{y}_1 + \hat{y}_2| |\hat{y}_1 - \hat{y}_2| + 2|y| |\hat{y}_1 - \hat{y}_2| \leq 4|\hat{y}_1 - \hat{y}_2| \quad (\text{since } y, \hat{y}_1, \hat{y}_2 \in [-1, 1]).
$$

(29) (30) (31) (32)

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

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

$$
P_{S \sim D} = \left( \sup_{f \in \mathcal{F}} \left| \mathbb{E}_{D}^S[f] - \mathbb{E}_{D}^{\mathcal{F}}[f] \right| \geq \epsilon \right) \leq 4N_1(\epsilon/32, \mathcal{F}, 2m) e^{-m\epsilon^2/512}.
$$

(33)

**Exercise.** Show that for $\mathcal{Y} = \mathcal{Y} = [-1, 1]$, the absolute loss given by $\ell_{abs}(y, \hat{y}) = |\hat{y} - y|$ $\forall y \in \mathcal{Y}, \hat{y} \in \mathcal{Y}$ 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 $N_1(\epsilon/8, \ell_{F}, 2m)$ or $N_1(\epsilon/8L, \mathcal{F}, 2m)$ 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 $\mathcal{F} \subseteq \mathcal{Y}$ 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 there exist $m$ real numbers $r_1, \ldots, r_m \in \mathbb{R}$ such that $\forall b = (b_1, \ldots, b_m) \in \{-1, 1\}^m$, $\exists f_b \in \mathcal{F}$ such that $\text{sign}(f_b(x_i) - r_i) = b_i \forall i \in [m]$. The pseudo-dimension of $\mathcal{F}$ is the cardinality of the largest set of points in $\mathcal{X}$ that can be pseudo-shattered by $\mathcal{F}$:

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

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 $\mathcal{F} = \{ f : \mathbb{R}^n \to \mathbb{R} \mid f(x) = w \cdot x + b \text{ for some } w \in \mathbb{R}^n, b \in \mathbb{R} \}$, 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 $\mathcal{F} \subseteq \mathcal{Y}$ and let $x_1^m = (x_1, \ldots, x_m) \in \mathcal{X}^m$. We say $x_1^m$ is fat-shattered by $\mathcal{F}$ if there exist $m$ real numbers $r_1, \ldots, r_m \in \mathbb{R}$ such that $\forall b = (b_1, \ldots, b_m) \in \{-1, 1\}^m$, $\exists f_b \in \mathcal{F}$ such that $b_i(f_b(x_i) - r_i) \geq \gamma \forall i \in [m]$. The fat-dimension of $\mathcal{F}$ or the fat-shattering dimension of $\mathcal{F}$ at scale $\gamma$ is the cardinality of the largest set of points in $\mathcal{X}$ that can be fat-shattered by $\mathcal{F}$:

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

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

Clearly, $\text{fat}_{\mathcal{F}}(\gamma) \leq \text{Pdim}(\mathcal{F}) \forall \gamma > 0$. The fat-shattering dimension is often called a \emph{scale-sensitive} dimension since it depends on the scale $\gamma$. 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 $\mathcal{F} \subseteq [a, b]^X$ for some $a \leq b$. Let $0 < \epsilon \leq b - a$, and let $\text{fat}_{\mathcal{F}}(\epsilon/8) = d < \infty$. 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 $\mathcal{F} \subseteq [a, b]^X$ for some $a \leq b$. Let $\text{Pdim}(\mathcal{F}) = d < \infty$. 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 $h(x) = \text{sign}(f(x))$ for some real-valued function $f$, and will see how the quantities considered in this lecture can be useful in such situations.

References