Introduction to Statistical Learning Theory

Lecture 5

So far we characterised learning for binary classification. Next we will show a similar result for regression.

The loss functions are continuous but not in general bounded. To overcome this we will assume that the output is bounded in some $[0, B]$.

The main difference from the binary classification proof - $\mathcal{H}|_C$ is not finite even for finite C.

We will need to measure size in a different manor.

Definition (Covering number)

Let (A, d) be a metric space. A set C is an ϵ-cover of A if every point of $x \in A$ there is a point $y \in C$ such that $d(x, y) < \epsilon$. The covering number $\mathcal{N}(\epsilon, A, d)$ is the size of the smallest ϵ-cover (or ∞).

$$\mathcal{N}(\epsilon, A, d) = \min\{|C| \;\; s.t.\; C \; is \; an \; \epsilon - cover\} \tag{1}$$

We will use $d_2(x, y) = \sqrt{\frac{1}{d}\sum_{i=1}^{d}(x_i - y_i)^2}$, $d_1(x, y) = \frac{1}{d}\sum_{i=1}^{d}|x_i - y_i|$ and $d_\infty(x, y) = \max_{i \in [d]}|x_i - y_i|$.

We note that $d_1(x, y) \leq d_2(x, y) \leq d_\infty(x_y)$ and therefore $\mathcal{N}(\epsilon, A, d_1) \leq \mathcal{N}(\epsilon, A, d_2) \leq \mathcal{N}(\epsilon, A, d_\infty)$.

Examples:

The hypercube $I_n = [0,1]^n$ with d_∞: The volume of I_n is 1, while the volume of the d_∞ ball of radius ϵ (hypercubes) is ϵ^n. Using a regular grid we can cover using $\lfloor \frac{1}{\epsilon} + 1 \rfloor^n$, and $\left(\frac{1}{\epsilon}\right)^n$ is a lower bound. This means that the $\mathcal{N}(\epsilon, I_n, d_\infty) = \Theta\left[\left(\frac{1}{\epsilon}\right)^n\right]$.

The hypercube $I_n = [0,1]^n$ with d_2: The volume of I_n is 1, while the volume of the d_2 ball of radius ϵ is $V_n(\epsilon) = C_n \cdot \epsilon^n$.

$A \subset \{\pm 1\}^n$ with d_∞: If $\epsilon > 2$ then $\mathcal{N}(\epsilon, \{\pm 1\}^n, d_\infty) = 1$. If $\epsilon \leq 2$ then $\mathcal{N}(\epsilon, \{\pm 1\}^n, d_\infty) = |A|$

Definition (Uniform covering number)

Let $\mathcal{H}$ be a hypothesis space of real functions. For any $\epsilon > 0$, and m the d_p unifrom covering number $\mathcal{N}_p(\epsilon, \mathcal{H}, m)$ is defined as

$$\mathcal{N}_p(\epsilon, \mathcal{H}, m) = \max_{C:|C|=m} \mathcal{N}(\epsilon, \mathcal{H}|_C, d_p) \tag{2}$$

For $C = \{x_1, ..., x_m\}$ we have $\mathcal{H}_C \subset \mathbb{R}^m$. $\mathcal{N}_p(\epsilon, \mathcal{H}, m)$ is the maximal "size" of $\mathcal{H}|_C$ for finite C. This is a continues version of the growth function.

For $\epsilon < 2$ and binary functions, $\mathcal{N}_\infty(\epsilon, \mathcal{H}, m) = \Pi_{\mathcal{H}}(m)$.

We define for convenience $\mathcal{F} = \ell \circ \mathcal{H}$ a set of functions from $\mathcal{X} \times \mathcal{Y}$ to $\mathbb{R}$.
$\mathcal{F} = \{\ell_h(x, y) = \ell(h(x), y) : \ for \ h \in \mathcal{H}\}$

Theorem

For $\mathcal{H}$ space of real functions, and loss ℓ bounded in $[0, B]$ for any distribution $\mathcal{D}$ and $\epsilon > 0$ we have

$$P_{S\sim\mathcal{D}^m}\left(\sup_h |L_S(h) - L_{\mathcal{D}}(h)| \geq \epsilon\right) \leq 4\mathcal{N}_1(\epsilon/8, \mathcal{F}, 2m)\exp\left(-\frac{m\epsilon^2}{32B^2}\right) \tag{3}$$

Proof sketch: The idea is similar to binary classification. First symmetrization - replace $L_{\mathcal{D}}$ with $L_{\tilde{S}}$

Second - Fix some sample (S_1, S_2) and use random permutations $\sigma \in \Gamma_m$.

Last step is discretization: Let G be an $\epsilon/8$ cover of $\mathcal{F}|_{(S_1,S_2)}$

Proof cont: If for some $\ell_h = \ell \circ h \in \mathcal{F}$ we have

$$|L_{S_1}(h) - L_{S_2}(h)| = \left| \frac{1}{m} \sum_{i=1}^{m} \ell_h(x_i, y_i) - \frac{1}{m} \sum_{i=m+1}^{2m} \ell_h(x_i, y_i) \right| \geq \frac{\epsilon}{2} \tag{4}$$

then for some $\ell_g \in G$ we have

$$\left| \frac{1}{m} \sum_{i=1}^{m} \ell_g(x_i, y_i) - \frac{1}{m} \sum_{i=m+1}^{2m} \ell_g(x_i, y_i) \right| \geq \frac{\epsilon}{4} \tag{5}$$

from the triangle inequality.

We can now use the Hoeffding inequality on the finite set G. □

Lemma

If the loss function ℓ is Lipschitz in the prediction with constant $L > 0$ then $\mathcal{N}_1(\epsilon, \mathcal{F}, m) \leq \mathcal{N}_1(\epsilon/L, \mathcal{H}, m)$

Proof:

$$\frac{1}{m}\sum_{i=1}^{m} |\ell_h(x_i, y_i) - \ell_g(x_i, y_i)| = \frac{1}{m}\sum_{i=1}^{m} |\ell(h(x_i), y_i) - \ell(g(x_i), y_i)|$$

$$\leq \frac{L}{m}\sum_{i=1}^{m} |h(x_i) - g(x_i)| = L \cdot d_1(h|_C, g|_C)$$

So if we have an ϵ/L cover of $\mathcal{H}|_C$ we have a ϵ cover of $\mathcal{F}|_C$.

We have seen we can generalize the growth function $\prod_{\mathcal{H}}(m)$ to uniform covering number $\mathcal{N}_1(\epsilon, \mathcal{H}, m)$.

We now will define continuous versions of the VC dimension, pseudo-dimension and fat-shattering dimension.

For binary classification, shattering means we can generate any output we want. This is too strong to generalize as is. The useful generalization is that binarize the output any way we want.

Definition (pseudo-shattering)

Let $\mathcal{H}$ be a set of real valued functions from input space $\mathcal{X}$. We say $C = (x_1, ..., x_m)$ is pseudo-shattered by $\mathcal{H}$ if there exists a vector $r = (r_1, ..., r_m)$ (called "witness") such that for all $b \in \{\pm 1\}^m = (b_1, ..., b_m)$ there exists $h_b \in \mathcal{H}$ such that $sign(h_b(x_i) - r_i) = b_i$

Shattering means you can find thresholds r_i such that you can get any combination of above/below.

Definition (pseudo-dimension)

Let $\mathcal{H}$ be a set of real valued functions from input space $\mathcal{X}$. The pseudo-dimension $Pdim(\mathcal{H})$ is the cardinality of the largest set pseudo-shattered by $\mathcal{H}$.

We can connect the Pseudo-dimension to the VC dimension:

Theorem

For every $h \in \mathcal{H}$ define the binary function $B_h(x, r) = sign(h(x) - r)$. Define $B_{\mathcal{H}} = \{B_h : h \in \mathcal{H}\}$ then $VC(B_{\mathcal{H}}) = Pdim(\mathcal{H})$.

The proof is direct from the definition of pseudo-shattering.

An alternative to $Pdim(\mathcal{H})$ is the fat-shattering dimension or scale-sensitive dimension.

Definition (γ-shattering)

Let $\mathcal{H}$ be a set of real valued functions from input space $\mathcal{X}$. We say $C = (x_1, ..., x_m)$ is γ-shattered by $\mathcal{H}$ if there exists a vector $r = (r_1, ..., r_m)$ such that for all $b \in \{\pm 1\}^m = (b_1, ..., b_m)$ there exists $h_b \in \mathcal{H}$ such that $b_i(h_b(x_i) - r_i) \geq \gamma$

γ-Shattering means you can find thresholds r_i such that you can get any combination of above/below with a *margin* of γ.

Definition (Fat-shattering dimension)

Let $\mathcal{H}$ be a set of real valued functions from input space $\mathcal{X}$. The fat-shattering dimension at scale γ, $fat_{\mathcal{H}}(\gamma)$ is the cardinality of the largest γ-shattered by $\mathcal{H}$.

Example 1: If $\mathcal{H}$ is a vector space of real-valued functions , then $Pdim(\mathcal{H}) = fat_{\mathcal{H}}(\gamma) = dim(\mathcal{H})$.

Proof: Using $B_{\mathcal{H}}$ and scale invariance.

Example 2: If $\mathcal{H}$ is the set of all functions from $[0,1]$ to $[0,1]$ with total variation at most V , then $fat_{\mathcal{H}}(\gamma) = 1 + \left\lfloor \frac{V}{2\gamma} \right\rfloor$ and $Pdim = \infty$.
Proof - HW.

Theorem

Let $\mathcal{H}$ be a set of real-valued functions

1. *For all γ, $fat_{\mathcal{H}}(\gamma) \leq Pdim(\mathcal{H})$.*
2. *The function $fat_{\mathcal{H}}$ is non-increasing with γ.*
3. *If a finite set S is pseudo-shattered, then there is some $\gamma_0 > 0$ such that for all $\gamma < \gamma_0$ the set S is γ-shattered.*
4. $\lim_{\gamma \searrow 0} fat_{\mathcal{H}}(\gamma) = Pdim(\mathcal{H})$

Proof: For (1) we note that if a set is γ-shattered it is pseudo-shattered as well.
For (2) we note that if $\gamma' < \gamma$ and a set is γ-shattered, it is also γ'-shattered.

Proof continued: (3) Assume $S = \{x_1, ..., x_n\}$ is pseudo-shattered, i.e. there exists a witness $r_1, ..., r_n$. For each $b \in \{\pm 1\}^m$ there exists $h_b \in \mathcal{H}$ such that $sign(h_b(x_i) - r_i) = b_i$.

If for all b, x_i we have $|h_b(x_i) - r_i| > 0$ then the set is γ-shattered for $\gamma \leq \gamma_0 = \min |h_b(x_i) - r_i|$ and witnessed by r.

Otherwise define $\gamma_0 = \frac{1}{2} \min\{r_i - h_b(x_i) : r_i > h_b(x_i)\}$ and the witness to the shattering is $r - \gamma_0/2$.

(4) is a conclusion from (1)-(3).

Theorem

Let $\mathcal{H}$ be a set of real-valued functions from $\mathcal{X}$ into $[0,1]$. Let $d = fat_{\mathcal{H}}(\epsilon/8)$ for $\epsilon \in (0,1]$. Then for $m \geq d$,

$$\mathcal{N}_1(\epsilon, \mathcal{H}, m) < 2\left(\frac{4}{\epsilon}\right)^{3d\log_2(16em/d\epsilon)} \tag{6}$$

Theorem

Let $\mathcal{H}$ be a set of real-valued functions from $\mathcal{X}$ into $[0,1]$. Let $d = Pdim(\mathcal{H})$ then,

$$\mathcal{N}_1(\epsilon, \mathcal{H}, m) < e(d+1)\left(\frac{2e}{\epsilon}\right)^{d} \tag{7}$$

Notice that the bound using $Pdim$ does not depend on m.

If $Pdim$ is finite, then that bound will be better but the fat-shattering one.

Theorem

Let $\mathcal{H}$ be a hypothesis space of real valued functions with $Pdim(\mathcal{H}) < \infty$ then $\mathcal{H}$ has uniform convergence with $\mathfrak{M}(\epsilon, \delta) = \mathcal{O}\left(\frac{Pdim(\mathcal{H}) \ln\left(\frac{1}{\epsilon}\right) + \ln\left(\frac{1}{\delta}\right)}{\epsilon^2}\right)$

Theorem

Let $\mathcal{H}$ be a hypothesis space of real valued functions with finite fat-shattering dimension then $\mathcal{H}$ has uniform convergence with $\mathfrak{M}(\epsilon, \delta) = \mathcal{O}\left(\frac{fat_{\mathcal{H}}(\epsilon/256) \ln\left(\frac{1}{\epsilon}\right) + \ln\left(\frac{1}{\delta}\right)}{\epsilon^2}\right)$

We will sketch the proof of the fat-shattering bound.

If $d = fat_{\mathcal{H}}(\epsilon/8)$ for $\epsilon \in (0, 1]$. We need to show that for $m \geq d$, $\mathcal{N}_1(\epsilon, \mathcal{H}, m) < 2\left(\frac{4}{\epsilon}\right)^{3d\log_2(16em/d\epsilon)}$.

The first step is to use packing numbers. Instead of measuring the size by how many points you need to cover, we measure by how many separate points can we pack.

Definition

A set W is ϵ-separated (regarding d metric) if $\forall x, y \in W : d(x, y) > \epsilon$. The ϵ packing number $\mathcal{M}(\epsilon, W, d)$ is the maximal criminality of an ϵ-separated set. The uniform packing number $\mathcal{M}_p(\epsilon, \mathcal{H}, k) = \max\{\mathcal{M}(\epsilon, \mathcal{H}|_C, d_p) : |C| = k\}$

Theorem

Let (A, d) be a metric space. For all positive ϵ, and for every subset $W \subset A$ the covering and packing number satisfy

$$\mathcal{M}(2\epsilon, W, d) \leq \mathcal{N}(\epsilon, W, d) \leq \mathcal{M}(\epsilon, W, d) \tag{8}$$

Left inequality if from the pigeonhole principle. The right inequality is from the fact that a maximal packing is also a cover.

The main work is to prove the following lemma

Lemma

Let $\mathcal{H}$ be a set of real-valued functions from $\mathcal{X}$ into $[0,1]$ and that $0 < \epsilon \leq 1$ then

$$\mathcal{M}_1(\epsilon, \mathcal{H}, m) < 2b^{3(\lceil \log_2 y \rceil + 1)} \tag{9}$$

where $b = \lfloor 4/\epsilon \rfloor$, $d = fat_{\mathcal{H}}(\epsilon/8)$ and $y = \sum_{i=1}^{d} \binom{m}{i} b^i$

Proving the theorem given the lemma is easy:
$\mathcal{N}_1(\epsilon, \mathcal{H}, m) \leq \mathcal{M}_1(\epsilon, \mathcal{H}, m) \leq 2\left(\frac{4}{\epsilon}\right)^{3(\lceil \log_2 y \rceil + 1)}$ and we have
$y \leq \sum_{i=1}^{d} \binom{m}{i} \left(\frac{4}{\epsilon}\right)^d \leq \left(\frac{4em}{\epsilon d}\right)^d$
From this we can show $\lceil \log_2(y) \rceil + 1 \leq d \log_2\left(\frac{16em}{d\epsilon}\right)$ to finish the proof.

We new need to prove the lemma. The next step is to replace each function h with a quantized version.

For all $h \in \mathcal{H}$ define the quantized version $Q_\alpha(h)$ as $Q_\alpha(h)(x) = \alpha\lfloor h(x)/\alpha \rfloor$. If $h(x) = k\alpha + r$ where $k \in \mathbb{Z}$ and $0 \leq r < \alpha$ then $Q_\alpha(h)(x) = k\alpha$. We define $Q_\alpha(\mathcal{H})$ as the set of quantized functions.

It is not hard to show that for all $h_1, h_2 \in \mathcal{H}$ we have

$$d_1(h_1|_C, h_2|_C) = \frac{1}{n}\sum_{i=1}^{n} |h_1(x_i) - h_2(x_i)| \leq d_1(Q_\alpha(h_1)|_C, Q_\alpha(h_2)|_C) + \alpha \tag{10}$$

From this we can see that $\mathcal{M}_1(\epsilon, \mathcal{H}, m) \leq \mathcal{M}_1(\epsilon - \alpha, Q_\alpha(\mathcal{H}), m)$

We have $\mathcal{M}_1(\epsilon, \mathcal{H}, m) \leq \mathcal{M}_1(\epsilon - \alpha, Q_\alpha(\mathcal{H}), m)$ and one can show $fat_{Q_\alpha(\mathcal{H})}(\epsilon) \leq fat_{\mathcal{H}}(\epsilon - \alpha/2)$.

Pick $\alpha = \epsilon/4$ and denote $F = Q_{\epsilon/4}(\mathcal{H})$, $d = fat_F(\epsilon/4) \leq fat_{\mathcal{H}}(\epsilon/8)$
It is enough to prove that $\mathcal{M}_1(3\epsilon/4, F, m) \leq 2b^{3(\lceil \log_2(y) \rceil + 1)}$. We can rescale everything by $\epsilon/4$

Lemma

Let $\mathcal{Y} = \{0, 1, ..., b\}$ with $b \geq 3$ and suppose $|\mathcal{X}| = m$ and $F \subset \mathcal{Y}^{\mathcal{X}}$ has $fat_F(1) = d \geq 1$, then

$$\mathcal{M}(3, F, d_1) \leq 2b^{3(\lceil \log_2(y) \rceil + 1)} \tag{11}$$

for $y = \sum_{i=1}^{d} \binom{m}{i} b^i$

Lemma

Proof sketch: The proof is based on the following definition

$$t(k,m) = \min\{|\{(A,r) : G\ 1-shatters\, A \subset \mathcal{X},\ witnessed\, by\, r,\ A \neq \emptyset\}| :$$

$$|\mathcal{X}| = m, G \subset \mathcal{Y}^{\mathcal{X}}\ |G| = k,\ and\, G\, is\, 3-separated\}$$

The minimum is infinity if it is the empty set.

The key observation is that if the fat-shattering dimension is d, there are at most $y = \sum_{i=1}^{d} \binom{m}{i} b^i$ possible (A, r) pairs.

If $t(k,m) > y$ that means that every 3-separated set of size k must 1-shatter a set of size greater then d. Since this is not possible it proves $\mathcal{M}(3, F, d_1) < k$

It is therefore enough to show $t(2b^{3(\lceil \log_2(y) \rceil + 1)}, m) \geq y$

Lemma

Need to show $t(2b^{3(\lceil \log_2(y) \rceil + 1)}, m) \geq y$

Let G be 3-separated set of size $k = 2b^{3(\lceil \log_2(y) \rceil + 1)}$, then we can split it into $k/2$ arbitrary pairs.

One can show (pigeonhole) that there x_0, i, j with $j > i + 2$ and at least k/b^3 pairs such that $g_1(x_0) = i$ and $g_2(x_0) = j$.

We can define $G_1, G_2 \subset G$ as all the functions such that $g(x_0) = i$ and $g(x_0) = j$. Both G_1 and G_2 are 3-separated.

If G_1 or G_2 shatters (A, r) $A \subset X - \{x_0\}$ then so does G. If both then so G shatters $A \cup \{x_0\}$. We can conclude that $t(k, m) \geq 2t(\lfloor \frac{k}{b^3} \rfloor, m - 1)$ and the proof follows by induction.