Introduction to Statistical Learning Theory

Lecture 3

Reminder:

Definition (Growth function)

The growth function of $\mathcal{H}$, $\Pi_{\mathcal{H}}(m)$ is the size of the largest restriction of $\mathcal{H}$ to a set of size m.

$$\Pi_{\mathcal{H}}(m) = \max\{|\mathcal{H}|_C| : C \subset \mathcal{X}, |C| = m\}$$

Theorem (Uniform convergence bound)

Let $\mathcal{H}$ be a hypothesis set of $\{\pm 1\}$ valued functions and ℓ be the $0-1$ loss, then for any distribution $\mathcal{D}$ on $\mathcal{X} \times \{\pm 1\}$, any $\epsilon > 0$ and positive integer m, we have

$$P_{S\sim\mathcal{D}^m}\left(\exists h \in \mathcal{H} : |L_S(h) - L_{\mathcal{D}}(h)| \geq \epsilon\right) \leq 4\Pi_{\mathcal{H}}(2m)\exp\left(-\frac{\epsilon^2 m}{8}\right)$$

In order to prove uniform convergence, and therefore PAC learnability, it is enough to show that the growth function is sub-exponential.

As we will see, the behavior $\Pi_{\mathcal{H}}(m)$ is greatly controlled by a single parameter - the VC dimension.

Definition (Shattering)

Let $\mathcal{H}$ be a set of functions from $\mathcal{X}$ to $\mathcal{Y} = \{\pm 1\}$. We say that $\mathcal{H}$ shatters $C \subset \mathcal{X}$ if $\mathcal{H}|_C = 2^C$.

Definition (VC-dimension)

Let $\mathcal{H}$ be a set of functions from $\mathcal{X}$ to $\mathcal{Y} = \{\pm 1\}$. The VC-dimension of $\mathcal{H}$ is the size of the largest finite set that $\mathcal{H}$ shatters (or ∞ if there is no maximum).

So $VC(\mathcal{H}) = d \Leftrightarrow \Pi_{\mathcal{H}}(d) = 2^d \wedge \Pi_{\mathcal{H}}(d+1) < 2^{d+1}$

Theorem (Sauer–Shelah)

Let $\mathcal{H}$ be a set of functions from $\mathcal{X}$ to $\mathcal{Y} = \{\pm 1\}$ with VC-dimension $d < \infty$, then $\Pi_{\mathcal{H}}(m) \leq \sum_{k=1}^{d} \binom{m}{k}$

Notice that $S(m, d) = \sum_{k=0}^{d} \binom{m}{k}$ is the number of subset of size smaller or equal to d of a set of size m.

Proof.

We will show a stronger claim

$$|\mathcal{H}|_C| \leq |\{B \subset C : \mathcal{H} \text{ shatters } B\}| \leq \sum_{k=0}^{d} \binom{m}{k}.$$

This is done by induction on m. For $m = 1$ the claim is trivial.

Proof (Cont.)

Let $C = \{x_1, ..., x_{m+1}\}$ and $\tilde{C} = \{x_1, ..., x_m\}$. Each function of $\mathcal{H}|_{\tilde{C}}$ corresponds to either one function in $\mathcal{H}|_C$ if it has a unique extension, or to two function if both extensions are possible.

Define $\mathcal{F} \subset \mathcal{H}|_{\tilde{C}}$ as all the function that correspond to two functions in $\mathcal{H}|_C$, then $|\mathcal{H}|_C| = |\mathcal{H}|_{\tilde{C}}| + |\mathcal{F}|$.

From our induction hypothesis $|\mathcal{H}|_{\tilde{C}}| \leq |\{B \subset \tilde{C} : \mathcal{H}\, shatters\, B\}| = |\{B \subset C : \mathcal{H}\, shatters\, B \wedge x_{m+1} \notin B\}|$.

For $\mathcal{F}$: $|\mathcal{F}| \leq |\{B \subset \tilde{C} : \mathcal{F}\, shatters\, B\}|$. For each such B shattered by $\mathcal{F}$, $B \cup \{x_{m+1}\}$ is shattered by $\mathcal{H}$, so $|\mathcal{F}| \leq |\{B \subset C : \mathcal{H}\, shatters\, B \wedge x_{m+1} \in B\}|$. □

Our next step is bound $\Pi_{\mathcal{H}}(m)$ with a simpler bound:

Theorem

Let $\mathcal{H}$ be a set of functions from $\mathcal{X}$ to $\mathcal{Y} = \{\pm 1\}$ with VC-dimension $d < \infty$, then $\Pi_{\mathcal{H}}(m) \leq \left(\frac{em}{d}\right)^d$ for $m > d$.

Proof.

We know from the Sauer-Shelah theorem that $\Pi_{\mathcal{H}}(m) \leq \sum_{k=0}^{d} \binom{m}{k}$

$$\sum_{k=0}^{d} \binom{m}{k} \leq \left(\frac{m}{d}\right)^d \sum_{k=0}^{d} \binom{m}{k} \left(\frac{d}{m}\right)^k \leq \left(\frac{m}{d}\right)^d \sum_{k=0}^{m} \binom{m}{k} \left(\frac{d}{m}\right)^k =$$

$$\overset{*}{=} \left(\frac{m}{d}\right)^d \left(1 + \frac{d}{m}\right)^m \overset{**}{<} \left(\frac{em}{d}\right)^d$$

Where (*) is the binomial theorem, and (**) follows from Euler's inequality $\left(1 + \frac{1}{x}\right)^x < e$ for all $x > 0$.

We can combine all our results to show that if $\mathcal{H}$ has VC-dimension $d < \infty$ then it is PAC learnable:

Theorem (PAC learnability of finite VC-dimension)

Let $\mathcal{H}$ be a set of functions from $\mathcal{X}$ to $\mathcal{Y} = \{\pm 1\}$ with VC-dimension $d < \infty$, then $\mathcal{H}$ has the uniform convergence property with $\mathfrak{M}(\epsilon, \delta) = \mathcal{O}\left(\frac{d\ln(\frac{1}{\epsilon}) + \ln(\frac{1}{\delta})}{\epsilon^2}\right)$ and is therefore PAC learnable with the ERM algorithm.

Note: One can get (with some extra effort) a better bound without the $\ln(\frac{1}{\epsilon})$ factor.

Proof (sketch).

To prove uniform convergence we need to show that

$$P_{S\sim\mathcal{D}^m}\left(\exists h\in\mathcal{H}:\ |L_S(h)-L_{\mathcal{D}}(h)|\geq\epsilon\right)\leq\delta\qquad\forall m\geq\mathfrak{M}(\epsilon,\delta)$$

We already showed that

$$P_{S\sim\mathcal{D}^m}\left(\exists h\in\mathcal{H}:\ |L_S(h)-L_{\mathcal{D}}(h)|\geq\epsilon\right)\leq 4\Pi_{\mathcal{H}}(2m)\exp\left(-\frac{\epsilon^2 m}{8}\right)$$

Using the inequality $\Pi_{\mathcal{H}}(m)\leq\left(\frac{em}{d}\right)^d$ and the inequality $\forall\alpha,x>0:\ \ln(x)\leq\alpha x-\ln(\alpha)$, we can show (with some algebra) that $4\Pi_{\mathcal{H}}(2m)\exp\left(-\frac{\epsilon^2 m}{8}\right)\leq\delta$. □

Some examples:
Example 1: $\mathcal{H} = 2^{\mathcal{X}}$ for infinite $\mathcal{X}$, $\Pi_{\mathcal{H}}(m) = 2^m$ Therefore $VC(\mathcal{H}) = \infty$.

Example 2: For finite $\mathcal{H}$, $\Pi_{\mathcal{H}}(m) \leq |\mathcal{H}| \Rightarrow VC(\mathcal{H}) \leq \log_2(|\mathcal{H}|)$.

Example 3: For $\mathcal{H} = \{h_a(x) = sign(x - a), a \in \mathbb{R}\}$,
$\Pi_{\mathcal{H}}(m) = m + 1 \Rightarrow VC(\mathcal{H}) = 1$.

Example 4: For $\mathcal{H} = \{h_a^{\pm}(x) = sign(\pm x - a), a \in \mathbb{R}\}$,
$\Pi_{\mathcal{H}}(m) = 2m \Rightarrow VC(\mathcal{H}) = 2$.

Example 5: The class of axis aligned rectangles $h_{(x_1,x_2,y_1,y_2)}(x,y) := 1 \Leftrightarrow x_1 < x < x_2 \wedge y_1 < y < y_2$ for $x_1 < x_2,\ y_1 < y_2$.

It is easy to find a 4 element set that $\mathcal{H}$ shatters. To show $VC(\mathcal{H}) = 4$, we need to show it cannot shatter any set of five elements. This can be done by observing that one point is always in the convex hall of the other points that cannot get zero if all others are one.

Example 6: The class of convex sets in the plan $h_C(x,y) := 1 \Leftrightarrow (x,y) \in C$ for a convex set $C \subset \mathbb{R}^2$. We can see that $VC(\mathcal{H}) = \infty$ by arranging points on the circle.

We will now compute the VC dimension on a simple, yet practical, hypothesis set - linear classifiers:
$\mathcal{H}_d = \{h_w(x) = sign(\langle w, x \rangle) : w \in \mathbb{R}^d\}$.

To show that $VC(\mathcal{H}_d) = d$ it is enough to prove the following lemma:

Lemma

The vectors $x_1, ..., x_k \in \mathbb{R}^d$ are shattered by $\mathcal{H}_d$ if and only if they are linearly independent.

Proof.

$\Rightarrow$ assume by contradiction that they are linearly dependent, so there exist some j such that $x_j = \sum_{i=1}^{j-1} \alpha_i x_i$. Any labeling y_i such that $\alpha_i y_i \geq 0$ has to have $y_j = 1$, therefore the set is not shattered - a contradiction.

$\Leftarrow$ Let X be the matrix with rows x_i^T, then the vector of labels given by any w is just $sign(X \cdot w)$. Our assumptions means that X has rank k and is therefore an onto mapping, and the set is shattered. $\square$

So far we have shown that if $VC(\mathcal{H}) < \infty$, then $\mathcal{H}$ is learnable by the ERM algorithm. We will show bounds on what is possible to learn.

We start with the "No-Free-Lunch theorem" that shows there is no universal learner, i.e. any learning algorithm will fail on some distributions. This means that any (useful) learning algorithm has to have some assumptions on the task being learned.

Theorem (No-Free-Lunch)

Let A be any learning algorithm for the task of binary classification with respect to the $0-1$ loss over a domain $\mathcal{X}$. Let m be any number smaller than $|\mathcal{X}|/2$, representing a training set size. Then, there exists a distribution $\mathcal{D}$ over $\mathcal{X} \times \{0,1\}$ such that:

1) There exists a function $f : \mathcal{X} \to \{0,1\}$ such that $L_{\mathcal{D}}(f) = 0$.
2) With probability at least $1/7$ over the choice of $S \sim \mathcal{D}^m$ we have that $L_{\mathcal{D}}(A(S)) \geq 1/8$.

Proof.

First we can assume w.l.o.g. that $|\mathcal{X}| = 2m$ (simply by giving probability zero to excess points). There are $T = 2^{2m}$ functions $f : \mathcal{X} \to \{0, 1\}$, we will label them as $f_1, ..., f_T$. For each f_i define $\mathcal{D}_i$ to be the distribution defined by f_i, i.e.

$$\mathcal{D}_i\left((x, y)\right) = \begin{cases} \frac{1}{2m} & if\ y = f_i(x) \\ 0 & otherwise. \end{cases} \tag{1}$$

The major step is to show that $\max_{i \in [T]} \mathbb{E}_{S \sim \mathcal{D}_i^m}\left[L_{\mathcal{D}_i}(A(S))\right] \geq 1/4$. To prove this we will use the probabilistic method - we will show that by picking a distribution $\mathcal{D}_i$ randomly, the expected loss is larger then $1/4$.

We will need this final notation: There are $k = (2m)^m$ possible samples $S = (x_1, ..., x_m)$ which we label $S_1, ..., S_k$. For each distribution $\mathcal{D}_i$ we define $S_j^i = ((x_1, f_i(x_1)), ..., (x_m, f_i(x_m))$.

Proof (cont.)

A simple changing of sums gives:

$$\mathbb{E}_{\mathcal{D}_i}\mathbb{E}_{S\sim\mathcal{D}_i^m}\left[L_{\mathcal{D}_i}(A(S))\right] = \frac{1}{T}\sum_{i=1}^{T}\frac{1}{k}\sum_{j=1}^{k}L_{\mathcal{D}_i}(A(S_j^i))$$

$$= \frac{1}{k}\sum_{j=1}^{k}\frac{1}{T}\sum_{i=1}^{T}L_{\mathcal{D}_i}(A(S_j^i)).$$

We will finish the proof by showing $\frac{1}{T}\sum_i L_{\mathcal{D}_i}(A(S_j^i)) \geq 1/4$.

This can be done by looking at all the m unseen point $v_1, ..., v_m$ and to notice that for each v_p we can split the distributions into $T/2$ pairs f_{i_1}, f_{i_2} that agree on all the labels except on v_p.

Proof (cont.)

As they agree on the sample, $S_j^{i_1} = S_j^{i_2}$ and the average loss between them is $1/2$. Since half the points are unseen this gives a $1/4$ loss bound. More rigorously:

$$\frac{1}{T}\sum_{i=1}^{T} L_{\mathcal{D}_i}(A(S_j^i)) \geq \frac{1}{2m}\sum_{p=1}^{m}\frac{1}{T}\sum_{i=1}^{T}\mathbb{1}[A(S_j^i)(v_p) \neq f_i(v_p)] =$$

$$\frac{1}{2m}\sum_{p=1}^{m}\frac{1}{T}\sum_{i=1}^{T/2}\mathbb{1}[A(S_j^{i_1})(v_p) \neq f_{i_1}(v_p)] + \mathbb{1}[A(S_j^{i_2})(v_p) \neq f_{i_2}(v_p)] =$$

$$\frac{1}{2m}\sum_{p=1}^{m}\frac{1}{T}\sum_{i=1}^{T/2} 1 = \frac{1}{4}.$$

This shows that $\max_{i\in[T]} \mathbb{E}_{S\sim\mathcal{D}_i}\left[L_{\mathcal{D}_i}(A(S)\right] \geq 1/4$.

Proof (final).

We have $\max_{i \in [T]} \mathbb{E}_{S \sim \mathcal{D}_i} [L_{\mathcal{D}_i}(A(S)] \geq 1/4$. Mark that distribution as $\mathcal{D}$. To finish the proof we need to show that: with probability at least $1/7$ over the choice of $S \sim \mathcal{D}^m$ we have that $L_{\mathcal{D}}(A(S)) \geq 1/8$. Using the law of total expectation we get:

$$\frac{1}{4} \leq \mathbb{E}_{S \sim \mathcal{D}} [L_{\mathcal{D}}(A(S)] \leq P\left(L_{\mathcal{D}}(A(S) \geq \frac{1}{8}\right) \cdot 1 + \left(1 - P\left(L_{\mathcal{D}}(A(S) \geq \frac{1}{8}\right)\right) \cdot \frac{1}{8} \Rightarrow P\left(L_{\mathcal{D}}(A(S) \geq \frac{1}{8}\right) \geq \frac{1}{7}$$

□

We will use the No-Free-Lunch theorem to show that any $\mathcal{H}$ with VC dimension is not PAC learnable.

Theorem

Let $\mathcal{H}$ be a hypothesis class of functions from a domain X to $\{0,1\}$ with $VC(\mathcal{H}) = \infty$ and let the loss function be the $0-1$ loss. The hypothesis class $\mathcal{H}$ is not PAC learnable.

Proof.

Assume by contradiction that $\mathcal{H}$ is PAC learnable. Then there exists some learning algorithm A (not necessarily ERM) such that for all $\epsilon, \delta > 0$ there exists $\mathcal{M}(\epsilon, \delta)$ so that if $m > \mathcal{M}(\epsilon, \delta)$ then for all distributions $\mathcal{D}$,
$P_{S\sim\mathcal{D}^m}(L_{\mathcal{D}}(A(S)) > L_{\mathcal{D}}(h^*) + \epsilon) < \delta$ where
$h^* = \arg\min_{h\in\mathcal{H}} L_{\mathcal{D}}(h)$

Proof.

Assume by contradiction that such algorithm exists. Pick some $\epsilon < 1/8$, $\delta < 1/7$ and $m > \mathcal{M}(\epsilon, \delta)$. Since $VC(\mathcal{H}) = \infty$ there exists some $x_1, ..., x_{2m} \in \mathcal{X}$ that $\mathcal{H}$ shatters.

From the No-Free-Lunch theorem there is a distribution $\mathcal{D}$ such that: There exists some $f : \mathcal{X} \to \{0, 1\}$ with $L_{\mathcal{D}}(f) = 0$ and $P_{S \sim \mathcal{D}^m}(L_{\mathcal{D}}(A(S)) > 1/8) > 1/7$.
If we remember the proof of the No-Free-Lunch, then we can recall that we can build such distribution supported only by $\{x_1, ..., x_{2m}\}$. Since this set is shattered by $\mathcal{H}$, this means that $L_{\mathcal{D}}(h^*) = 0$.

This finishes the proof as $P_{S \sim \mathcal{D}^m}(L_{\mathcal{D}}(A(S)) > L_{\mathcal{D}}(h^*) + \epsilon) \geq P_{S \sim \mathcal{D}^m}(L_{\mathcal{D}}(A(S)) > 1/8) > 1/7 > \delta$. □

We can combine everything we did so far and get the fundamental theorem of statistical learning (binary classification):

Theorem (Fundamental Theorem of Statistical Learning)

Let $\mathcal{H}$ be a hypothesis class of functions from a domain X to $\{0,1\}$ and let the loss function be the $0-1$ loss. The following are equivalent:

1. *$\mathcal{H}$ has uniform convergence.*
2. *The ERM is a PAC learning algorithm for $\mathcal{H}$.*
3. *$\mathcal{H}$ is PAC learnable.*
4. *$\mathcal{H}$ has finite VC dimension.*

Proof.

$1 \Rightarrow 2$ We have seen uniform convergence implies that ERM is PAC learnable in lecture 2.

$2 \Rightarrow 3$ Obvious.

$3 \Rightarrow 4$ We just proved that PAC learnability implies finite VC dimension.

$4 \Rightarrow 1$ We proved in lecture 3 that finite VC dimension implies uniform convergence.

Remarks:
We notice that the VC dimension fully determines learnability for *binary classification*.

We can extend to regression problem with a similar idea called fat shattering dimension.

The VC dimension doesn't just determine learnability, it also gives a bound on the sample complexity (which we will show is tight).