Introduction to Statistical Learning Theory

Lecture 4

Quick recap:

We have seen that if $\mathcal{H}$ has finite VC dimension then it has uniform convergence and therefore PAC learnable using the ERM algorithm.

We also have seen the No-Free-Lunch theorem that shows that any learning algorithm will fail on some tasks.

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 $L_{\mathcal{D}}(A(S)) \geq 1/8$.

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 $\mathcal{M}(\epsilon, \delta)$ such 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).

We have shown that if $VC(\mathcal{H}) = d$ then we can learn with $\mathcal{M}(\epsilon, \delta) = \mathcal{O}\left(\frac{d\ln(1/\epsilon)+\ln(1/\delta)}{\epsilon^2}\right)$ (and claimed the $\ln(1/\epsilon)$ can be removed).
We will show that this bound is tight (up to the $\ln(1/\epsilon)$).

Theorem (Complexity lower bound)

Let $\mathcal{H}$ be a hypothesis class of functions from a domain $\mathcal{X}$ to $\{0, 1\}$ with $VC(\mathcal{H}) > 0$ and let the loss function be the $0-1$ loss. Any PAC learning algorithm has sample complexity $\mathcal{M}(\epsilon, \delta) = \Omega\left(\frac{d+\ln(1/\delta)}{\epsilon^2}\right)$.

We will split the dependence in δ and d, starting with δ:

Lemma (1)

Under the previous conditions, $\mathcal{M}(\epsilon, \delta) \geq 0.5 \log(1/(4\delta))/\epsilon^2$ *for* $\epsilon < 1/\sqrt{2}$.

The idea of the proof is to pick 2 almost identical distributions (depending on ϵ) with different optimal solution, so that in order to differentiate with high probability a large number of samples is needed.

Proof: Choose some $c \in \mathcal{X}$ that $\mathcal{H}$ shatters. For each $b \in \{\pm 1\}$ we will define a distribution $\mathcal{D}_b$ that picks c with probability 1, and b with probability $\frac{1+\epsilon}{2}$. This means that $\mathcal{D}_b((c, y)) = \frac{1+by\epsilon}{2}$. It is also not hard to see that $L_{\mathcal{D}_b}(h) = \frac{1-bh(c)\epsilon}{2}$.

Since $L_{\mathcal{D}_b}(h) = \frac{1-bh(c)\epsilon}{2}$ the optimal hypothesis has $L_{\mathcal{D}_b}(h^*) = \frac{1-\epsilon}{2}$, so if $h(c) \neq b$ then $L_{\mathcal{D}_b}(h) = \frac{1+\epsilon}{2} = L_{\mathcal{D}_b}(h^*) + \epsilon$. This means that h is an ϵ approximation iff $h(c) = b$.

We will use the following notations: As x is irrelevant, we will only look at $Y = (y_1, ..., y_m)$. Also we will write $A(Y)$ for $A(Y)(c)$ (as this is what we care about). Lastly we will define
$N_+ = \{Y \in \{\pm 1\}^m : \sum y_i \geq 0\}$ and $N_- = \{\pm 1\}^m \backslash N_+$.

Notice that for $Y \in N_+$, we have $P_+(Y) \geq P_-(Y)$ and the opposite for $Y \in N_-$.
We will now show that optimal algorithm (considering the worst case out of $\mathcal{D}_+$ and $\mathcal{D}_-$) is the ERM.

$$\max_{b\in\{\pm\}} P_b\left(A(Y)\neq b\right) \geq \frac{1}{2}P_+(A(Y)=-1)+\frac{1}{2}P_-(A(Y)=1)$$

$$=\frac{1}{2}\sum_{Y\in N_+} P_+(Y)\mathbb{1}(A(Y)=-1)+\sum_{Y\in N_-} P_+(Y)\mathbb{1}(A(Y)=-1)+$$

$$\frac{1}{2}\sum_{Y\in N_+} P_-(Y)\mathbb{1}(A(Y)=1)+\sum_{Y\in N_-} P_-(Y)\mathbb{1}(A(Y)=1)=$$

$$\frac{1}{2}\sum_{Y\in N_+} P_+(Y)\mathbb{1}(A(Y)=-1)+P_-(Y)\mathbb{1}(A(Y)=1)+$$

$$\frac{1}{2}\sum_{Y\in N_-} P_+(Y)\mathbb{1}(A(Y)=-1)+P_-(Y)\mathbb{1}(A(Y)=1)\geq$$

$$\frac{1}{2}\sum_{Y\in N_+} P_-(Y)\mathbb{1}(A(Y)=-1)+P_-(Y)\mathbb{1}(A(Y)=1)+$$

$$\frac{1}{2}\sum_{Y\in N_+} P_+(Y)\mathbb{1}(A(Y)=-1)+P_+(Y)\mathbb{1}(A(Y)=1)=\frac{1}{2}\left(L_{\mathcal{D}_+}(ERM)+L_{\mathcal{D}_-}(ERM)\right)$$

For the ERM, $L_{\mathcal{D}_+}(ERM) = L_{\mathcal{D}_-}(ERM)$ (up to ties which we can exclude by having uneven m). Both are equal that a binomial $B(m, (1-\epsilon)/2)$ has a value greater then $m/2$. This can be bounded using Slud's inequality:

Theorem (Slud's inequality)

Let $X \sim B(m, (1-\epsilon)/2)$ *then*

$$P(X \geq m/2) \geq \frac{1}{2}\left(1 - \sqrt{1 - \exp(-m\epsilon^2/(1-\epsilon^2))}\right)$$

So the error probability is greater or equal to $\frac{1}{2}\left(1 - \sqrt{1 - \exp(-m\epsilon^2/(1-\epsilon^2))}\right) \geq \frac{1}{2}\left(1 - \sqrt{1 - \exp(-2m\epsilon^2)}\right)$ using the $\epsilon^2 < 1/2$ assumption. We can conclude that for $m < 0.5\ln(1/(4\delta))/\epsilon^2$

$$\max_b P\left(L_{\mathcal{D}_b}(A(Y)) - \min_h L_{\mathcal{D}_b}(h) \geq \epsilon\right) \geq \frac{1}{2}(1 - \sqrt{1-4\delta}) \geq \delta$$

Where the last inequality is simple algebra. This finishes the proof.

We now need to bound the dependence in $d = VC(\mathcal{H})$

Lemma (2)

Under the previous conditions, $\mathcal{M}(\epsilon, \delta) \geq \frac{d}{8^3 \epsilon^2}$ *for* $\epsilon < 1/8\sqrt{2}$.

The proof is similar to the previous proof. Define $\rho = 8\epsilon$. Pick $c_1, ..., c_d$ that $\mathcal{H}$ shatters. for any $b \in \{\pm 1\}^d$ define a distribution $\mathcal{D}_b$ that first picks $x = c_i$ uniformly out of $c_1, ..., c_d$ then picks y with probability $(1 + yb_i\rho)/2$.

The next step is to prove that the ERM is optimal algorithm when considering *worst case*. The proof is very similar to what we did earlier (using independence and the same tricks) but a bit more cumbersome so we will skip it.

For any function f

$$L_{\mathcal{D}_b}(f) = \frac{1+\rho}{2} \cdot \frac{|\{i \in [d] : \ f(c_i) \neq b_i\}|}{d} + \frac{1-\rho}{2} \cdot \frac{|\{i \in [d] : \ f(c_i) = b_i\}|}{d}$$

So $L_{\mathcal{D}_b}(f) - \min_h L_{\mathcal{D}_b}(h) = \rho \cdot \frac{|\{i \in [d] : f(c_i) \neq b_i\}|}{d}$.

We will bound $\mathbb{E}_{S \sim \mathcal{D}_b^M}[L_{\mathcal{D}_b}(ERM(S)) - \min_{h \in \mathcal{H}} L_{\mathcal{D}_b}(h)]$ next:

$$\mathbb{E}_S[L_{\mathcal{D}_b}(ERM(S)) - \min_{h \in \mathcal{H}} L_{\mathcal{D}_b}(h)] = \frac{\rho}{d} \mathbb{E}_S[|\{i \in [d] : \ ERM(c_i) \neq b_i\}|]$$

We can look at the sampling as first sampling the c_i index $K \sim U([d])^m$ and then sampling the labels $y_i \sim b_{K_i}$ (with some abuse of notation).

We define for each $K \in [d]^m$, $n_i(K)$ the number of times the index i appears in K. Then

$$\frac{\rho}{d}\mathbb{E}_S[|\{i \in [d] : ERM(c_i) \neq b_i\}|] = \frac{\rho}{d}\sum_{i=1}^{d}\mathbb{E}_K\mathbb{E}_{y_j \sim b_{K_j}}[\mathbb{1}(ERM(S)(c_i) \neq b_i]$$

$$\overset{1}{\geq} \frac{\rho}{2d}\sum_{i=1}^{d}\mathbb{E}_K(1-\sqrt{1-\exp(-2\rho^2 n_i(K))}) \overset{2}{\geq} \frac{\rho}{2d}\sum_{i=1}^{d}\mathbb{E}_K(1-\sqrt{2\rho^2 n_i(K)})$$

$$\overset{3}{\geq} \frac{\rho}{2d}\sum_{i=1}^{d}\left(1-\sqrt{2\rho^2\mathbb{E}_K[n_i(K)]}\right) = \frac{\rho}{2d}\sum_{i=1}^{d}\left(1-\sqrt{2\rho^2 m/d}\right)$$

$$= \frac{\rho}{2}\left(1-\sqrt{2\rho^2 m/d}\right)$$

Where (1) is Slut's inequality as before (using $\rho^2 < 1/2$), (2) if from the inequality $1 - e^{-a} \geq a$ and (3) is Jensen's inequality.

In summery we have shown so far that for every algorithm A, there exists a distribution such that
$\mathbb{E}_S[L_{\mathcal{D}_b}(A(S)) - \min_{h\in\mathcal{H}} L_{\mathcal{D}_b}(h)] \geq \frac{\rho}{2}\left(1 - \sqrt{2\rho^2 m/d}\right) \geq \frac{\rho}{4}$
for $m < \frac{d}{8^3\epsilon^2} = \frac{d}{8\rho^2}$.

To finish we will use a version of the Markov inequality
$P(X > a) \geq \mathbb{E}[X] - a$, for $X \in [0,1]$, $a \in (0,1)$. Define
$\Delta = \frac{1}{\rho}\left(L_{\mathcal{D}_b}(A(S)) - \min_{h\in\mathcal{H}} L_{\mathcal{D}_b}(h)\right)$ and notice that $\Delta \in [0,1]$.

$$P(L_{\mathcal{D}_b}(A(S)) - \min_{h\in\mathcal{H}} L_{\mathcal{D}_b}(h) > \epsilon) = P(\Delta > \epsilon/\rho) \geq \mathbb{E}[\Delta] - \frac{\epsilon}{\rho} \geq \frac{1}{4} - \frac{\epsilon}{\rho} = \frac{1}{8}$$

finishing the proof of the lemma. With both lemmas, the theorem is straightforward.

We have seen that learning is possible with $\mathcal{M}(\epsilon,\delta) = \mathcal{O}\left(\frac{d\ln(1/\epsilon)+\ln(1/\delta)}{\epsilon^2}\right)$ using the ERM algorithm, and that $\mathcal{M}(\epsilon,\delta) = \Omega\left(\frac{d+\ln(1/\delta)}{\epsilon^2}\right)$ for any learning algorithm.

We have seen (and it can be extended) that the ERM is optimal when it comes to minimizing the worst case scenario.

It is important to note, that under further assumptions (such as smoothness, etc.) other algorithms may perform much better.