Introduction to Statistical Learning Theory

Lecture 2

Reminder: We are given m samples $\{(x_i, y_i)\}_{i=1}^m \sim \mathcal{D}^m$ and a hypothesis space $\mathcal{H}$ and we wish to return $h \in \mathcal{H}$ minimizing $L_{\mathcal{D}}(h) = \mathbb{E}[\ell(h(x), y)]$.

Problem 1: It is unrealistic to hope to find the exact minimizer after seeing only a sample of the data (or even if we had perfect knowledge). We can only reasonably hope for an **approximate** solution:
$L_{\mathcal{D}}(h) \leq \min_{h \in \mathcal{H}} L_{\mathcal{D}}(h) + \epsilon$.

Problem 2: We depend on a random sample. There is always a chance we get a bad sample that doesn't represent $\mathcal{D}$. Our algorithm can only be **probably** correct: there is always some probability δ that we are completely wrong.

We wish to find a **probably approximately correct (PAC)** hypothesis.

Definition (PAC learnable)

A hypothesis class $\mathcal{H}$ is PAC learnable, if there exists a learning algorithm A, satisfying that for any $\epsilon > 0$ and $\delta \in (0, 1)$ there exist $\mathfrak{M}(\epsilon, \delta) = poly(\frac{1}{\epsilon}, \frac{1}{\delta})$ such that for i.i.d samples $S^m = \{(x_i, y_i)\}_{i=1}^m$ drawn from any distribution $\mathcal{D}$ and $m \geq \mathfrak{M}(\epsilon, \delta)$ the algorithm returns a hypothesis $A(S^m) \in \mathcal{H}$ satisfying

$$P_{S^m \sim \mathcal{D}^m}(L_{\mathcal{D}}(A(S)) > \min_{h \in \mathcal{H}} L_{\mathcal{D}}(h) + \epsilon) < \delta$$

Next will show that if $L_S(h) \approx L_{\mathcal{D}}(h)$ for all h then the ERM is a PAC learning algorithm.

We can look at an error of a learning algorithm $A : S^m \to \mathcal{H}$ as

$$L_{\mathcal{D}}(A(S)) = L_{\mathcal{D}}(h^*) + (L_{\mathcal{D}}(A(S) - L_{\mathcal{D}}(h^*))) \tag{1}$$

The first term is the approximation error. If we enlarge $\mathcal{H}$ it will decrease (or not increase).

The second term is the estimation error. In general, the richer $\mathcal{H}$ is the harder it is to find the optimum and this should increase. This is what we will focus on.

Definition (Uniform convergence)

A hypothesis class $\mathcal{H}$ has the uniform convergence property, if for any $\epsilon > 0$ and $\delta \in (0,1)$ there exist $\mathfrak{M}(\epsilon, \delta) = poly(\frac{1}{\epsilon}, \frac{1}{\delta})$ such that for any distribution $\mathcal{D}$ and $m \geq \mathfrak{M}(\epsilon, \delta)$ i.i.d samples $S^m = \{(x_i, y_i)\}_{i=1}^m \sim \mathcal{D}^m$ with probability at least $1 - \delta$, $|L_S^m(h) - L_{\mathcal{D}}(h)| < \epsilon$ for all $h \in \mathcal{H}$.

It is trivial to bound $|L_S^m(h) - L_{\mathcal{D}}(h)|$ for a single h using the Hoeffding inequality (for a bounded loss function). The difficulty is to bound all the $h \in \mathcal{H}$ uniformly.

Theorem (PAC by uniform convergence)

If $\mathcal{H}$ has the uniform convergence with $\mathfrak{M}(\epsilon, \delta)$ then $\mathcal{H}$ is PAC learnable with the ERM algorithm and $\mathfrak{M}(\frac{\epsilon}{2}, \delta)$ samples.

Proof.

By uniform convergence: With probability at least $1 - \delta$ for all $h \in \mathcal{H}$, $|L_S(h) - L_{\mathcal{D}}(h)| \leq \frac{\epsilon}{2}$.

Define $h_{ERM} = \arg\min_{h \in \mathcal{H}} L_S(h)$ and $h^* = \arg\min_{h \in \mathcal{H}} L_{\mathcal{D}}(h)$.

$L_{\mathcal{D}}(h_{ERM}) \leq L_S(h_{ERM}) + \frac{\epsilon}{2} \leq L_S(h^*) + \frac{\epsilon}{2} \leq L_{\mathcal{D}}(h^*) + \epsilon$ □

A first simple example of PAC learnable spaces - finite hypothesis spaces.

Theorem (uniform convergence for finite $\mathcal{H}$)

Let $\mathcal{H}$ be a finite hypothesis space and $\ell : \mathcal{Y} \times \mathcal{Y} \to [0,1]$ be a bounded loss function, then $\mathcal{H}$ has the uniform convergence property with $\mathfrak{M}(\epsilon, \delta) = \frac{\ln\left(\frac{2|\mathcal{H}|}{\delta}\right)}{2\epsilon^2}$ *and is therefore PAC learnable by the ERM algorithm.*

Proof.

For any $h \in \mathcal{H}$, $\ell(h(x_1), y_1), ..., \ell(h(x_m), y_m)$ are i.i.d random variables with expected value $L_{\mathcal{D}}(h)$.

According to the Hoeffding inequality,

$$P(|L_S(h) - L_{\mathcal{D}}(h)| > \epsilon) \leq 2e^{-2\epsilon^2 m} \leq 2e^{-2\epsilon^2 \mathfrak{M}(\epsilon,\delta)} = \frac{\delta}{|\mathcal{H}|} \tag{2}$$

Proof (Cont.)

We can now use the union bound: For all events $A_1, ..., A_n$

$$P(\cup_{i=1}^{n} A_i) \leq \sum_{i=1}^{n} P(A_i) \tag{3}$$

For all $h \in \mathcal{H}$ define A_h as the event that $|L_S(h) - L_{\mathcal{D}}(h)| > \epsilon$. By equation 2 we know that $P(A_h) \leq \frac{\delta}{|\mathcal{H}|}$. With equation 3 we can conclude

$$P(\exists h \in \mathcal{H} : |L_S(h) - L_{\mathcal{D}}(h)| > \epsilon) = P(\cup_{h \in \mathcal{H}} A_h) \leq \sum_{h \in \mathcal{H}} P(A_h)$$

$$\leq \sum_{h \in \mathcal{H}} \frac{\delta}{|\mathcal{H}|} = \delta$$

□

We have seen that finite hypothesis class can be learned, but what about infinite ones like linear predictors?

We can discretize (after all we are working on a finite precision machines), but this is not a great solution. The main problem is with the use of the union bound as similar hypothesis will fail on similar samples.

The solution is the check how many **effective** hypothesis there are on a sample of size m.

We will restrict ourselves (for the time being) to binary classification with $0-1$ loss.

Definition

Let $\mathcal{H}$ be a set of function from $\mathcal{X}$ to $\{\pm 1\}$ and let $C \subset \mathcal{X}$ be a subset of the input space. We denote by $\mathcal{H}|_C$ all the function that can be derived by restricting functions in $\mathcal{H}$ to C.

$$\mathcal{H}|_C = \{h|_C : C \to \{\pm 1\} : \ h \in \mathcal{H}\}$$

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.

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

Notice that $\Pi_{\mathcal{H}}(m) \leq 2^m$.

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

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

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

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

As we can see, even for an infinite hypothesis set it is possible that $\Pi_{\mathcal{H}}(m) \ll 2^m$.

We can now state the main theorem that shows the importance of the growth function.

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

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

Immediate corollary - if $\Pi_{\mathcal{H}}(m)$ grows sub-exponentially then $\mathcal{H}$ is PAC learnable.

This is not a simple proof, so we will go over the main steps first.

We wish to reduce the problem to a finite problem , so we will start by showing the we can replace $L_{\mathcal{D}}$ by $L_{\tilde{S}}$ - the error on another m independent "test" samples.

The next step to show you can fix the samples, and look at the probability of permuting between the train and test sets.

Last part will be to use the union bound and Hoeffding on this reduced case.

We define $Z = \mathcal{X} \times \{\pm 1\}$.

Lemma (1)

Let $Q = \{S \in Z^m : \exists h \in \mathcal{H}\, s.t. \, |L_S(h) - L_{\mathcal{D}}(h)| \geq \epsilon\}$ *and* $R = \{(S_1, S_2) \in Z^{2m} : \exists h \in \mathcal{H}\, s.t. \, |L_{S_1}(h) - L_{S_2}(h)| \geq \frac{\epsilon}{2}\}$. *For* $m \geq \frac{4}{\epsilon^2}$, $P_{S \sim \mathcal{D}^m}(Q) \leq 2P_{S_1 \times S_2 \sim \mathcal{D}^{2m}}(R)$.

Proof.

Let $S_1 \in Q$ and pick $h \in \mathcal{H}$ such that $|L_{S_1}(h) - L_{\mathcal{D}}(h)| \geq \epsilon$.
By the Hoeffding inequality we know that $P_{S_2}(|L_{S_2}(h) - L_{\mathcal{D}}(h)| \leq \frac{\epsilon}{2}) \geq \frac{1}{2}$.
This means that

$$P\left(\exists h \in \mathcal{H} : \, |L_{\mathcal{D}}(h) - L_{S_1}(h)| \geq \epsilon \wedge |L_{\mathcal{D}}(h) - L_{S_2}(h)| \leq \frac{\epsilon}{2}\right)$$
$$\geq \frac{P(Q)}{2}$$

Proof (Cont.)

$$P\left(\exists h \in \mathcal{H} : |L_{\mathcal{D}}(h) - L_{S_1}(h)| \geq \epsilon \wedge |L_{\mathcal{D}}(h) - L_{S_2}(h)| \leq \frac{\epsilon}{2}\right)$$
$$\geq \frac{P(Q)}{2}$$

We now notice that if $|L_{\mathcal{D}}(h) - L_{S_1}(h)| \geq \epsilon$ and $|L_{\mathcal{D}}(h) - L_{S_2}(h)| \leq \frac{\epsilon}{2}$, then by the triangle inequality $|L_{S_2}(h) - L_{S_1}(h)| \geq \frac{\epsilon}{2}$.

This means that the probability above is lesser or equal to $P(R)$ concluding our proof. □

The next step is to bound the probability of R with permutations between "training" and "testing".

Define Γ_m as the set of permutations on $\{1, ..., 2m\}$ that swap between i and $i+m$, i.e. for $\sigma \in \Gamma_m$ and $1 \leq i \leq m$, $\sigma(i) = i$ or $\sigma(i) = i+m$.

Lemma (2)

Let R be any subset of Z^{2m} and $\mathcal{D}$ any distribution on Z. Then

$$P_{S\sim\mathcal{D}^{2m}}(R) = \mathbb{E}_S\left[P_\sigma(\sigma S \in R)\right] \leq \max_{S\in Z^{2m}} P_\sigma(\sigma S \in R)$$

When σ is chosen uniformly from Γ_m.

Proof.

As S is a set of $2m$ i.i.d samples, then the probability of any event is invariant to permutation, i.e. $\forall \sigma \in \Gamma_m$, $P_{S\sim\mathcal{D}^{2m}}(R) = P_{S\sim\mathcal{D}^{2m}}(\sigma S \in R)$.

Based on this we can deduce:

$$P_{S\sim\mathcal{D}^{2m}}(R) = \mathbb{E}_S[\mathbb{1}_R(S)] = \frac{1}{|\Gamma_m|}\sum_{\sigma\in\Gamma_m}\mathbb{E}_S[\mathbb{1}_R(\sigma S)]$$

From the linearity of expectation we get

$$P_{S\sim\mathcal{D}^{2m}}(R) = \mathbb{E}_S\left[\frac{1}{|\Gamma_m|}\sum_{\sigma\in\Gamma_m}\mathbb{1}_R(\sigma S)\right] = \mathbb{E}_S\left[P_\sigma(\sigma S \in R)\right]$$

This proves the first equality, the fact that $\mathbb{E}_S\left[P_\sigma(\sigma S \in R)\right] \leq \max_{S\in Z^{2m}} P_\sigma(\sigma S \in R)$ is trivial. □

We have shown that we just need to bound the probability of permuting a fixed sample.

Lemma (3)

For the set $R = \{(S_1, S_2) \in Z^{2m} : \exists h \in \mathcal{H}\, s.t. \, |L_{S_1}(h) - L_{S_2}(h)| \geq \frac{\epsilon}{2}\}$ *as in lemma 1, and permutation* σ *chosen uniformly from* Γ_m*,*

$$\max_{S \in Z^{2m}} P_\sigma(\sigma S \in R) \leq 2\Pi_{\mathcal{H}}(2m) e^{-\frac{\epsilon^2 m}{8}}$$

Proof.

Let $S = ((x_1, y_1), ..., (x_{2m}, y_{2m}))$ be the maximizing S, and let $C = \{x_1, ..., x_{2m}\}$. By definition $\mathcal{H}|_C = \{h_1, ..., h_t\}$ for $t \leq \Pi_{\mathcal{H}}(2m)$.

Proof (Cont.)

We have $\sigma S \in R$ if and only if for some $h \in \mathcal{H}$,

$$\left| \frac{1}{m} \sum_{i=1}^{m} \ell(h(x_{\sigma(i)}), y_{\sigma(i)}) - \frac{1}{m} \sum_{i=m+1}^{2m} \ell(h(x_{\sigma(i)}), y_{\sigma(i)}) \right| \geq \frac{\epsilon}{2}$$

As $h|_C \equiv h_j|C$ for some $1 \leq j \leq t$, it is enough to look at $h_1, ..., h_t$. We define

$$u_i^j = \begin{cases} 1 & \text{if } h_j(x_i) \neq y_i \\ 0 & \text{otherwise} \end{cases}$$

So $\sigma S \in R$ if and only if for some $1 \leq j \leq t$

$$\left| \frac{1}{m} \sum_{i=1}^{m} u_{\sigma(i)}^j - \frac{1}{m} \sum_{i=m+1}^{2m} u_{\sigma(i)}^j \right| \geq \frac{\epsilon}{2} \tag{4}$$

Proof (Cont.)

Notice that $u^j_{\sigma(i)} - u^j_{\sigma(m+i)} = \pm|u^j_i - u^j_{m+i}|$ with both possibilities equally likely, so

$$P_\sigma\left(\left|\frac{1}{m}\sum_{i=1}^{m}(u^j_{\sigma(i)} - u^j_{\sigma(i+m)})\right| \geq \frac{\epsilon}{2}\right) = P\left(\left|\frac{1}{m}\sum_{i=1}^{m}|u^j_i - u^j_{m+i}|\beta_i\right| \geq \frac{\epsilon}{2}\right)$$

where $\beta_i \in \{\pm 1\}$ uniformly and independently. By the hoeffding inequality this is smaller then $2\exp\left(-\frac{\epsilon^2 m}{8}\right)$ and using the union bound on all $h \in \mathcal{H}|_C$ we can bound it by $2\Pi_{\mathcal{H}}(2m)e^{-\frac{\epsilon^2 m}{8}}$ □

Summery -

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

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

The proof is just the combination of lemmas 1-3.

note: in lemma 1 re required that $m \geq \frac{4}{\epsilon^2}$, this is not a problem because the bound in this theorem is trivial for $m < \frac{4}{\epsilon^2}$.