CS 596: Theoretical Machine Learning

Lecturer: Pranjal Awasthi
Scribe: Ana Echavarria

Lecture # 3
September 27, 2016

1 Probably Approximately Correct (PAC) Model

Let $\mathcal{X}$ be an instance space, D be a distribution over $\mathcal{X}$ and H be a class of functions $f : \mathcal{X} \mapsto \{0, 1\}$. For a function $h^* \in H$, the goal of the PAC model the goal is to design an algorithm A that learns h^* given a training set $S = \{(x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m)\}$ where $x_i \in X$ and $y_i = h^*(x_i)$.
We define the true error of a function $h \in H$ to be $err(h) = Pr_{x\sim D}[h(x) \neq h^*(x)]$ and the empirical error of a h with respect to a training set S as $err_S(h) = \frac{1}{|S|}\sum_{i=1}^{|S|} I(h(x_i) \neq y_i)$.
ERM selects $h = \arg\min_{f\in H} err_S(f)$ and we proved last lecture that given a finite function class H, ERM PAC-learns H provided the size of the training set $m \geq \frac{1}{\epsilon}\log\left(\frac{|H|}{\delta}\right)$.
However, the following two cases are not covered in this theorem:

- What happens if $h^* \notin H$?
- What happens if H is an infinite function class?

We will address these issues in the next two subsections.

1.1 PAC learning any function

Theorem 1. *Given a finite function class H, for any $\epsilon > 0$, $\delta > 0$, any distribution D over $\mathcal{X}$ and **any target function** h^*. With probability $\geq 1 - \delta$ we have that for all $h \in H$ $|err_S(h) - err(h)| \leq \epsilon$ provided the size of the training set $m = |S|$ is at least $\frac{1}{2\epsilon^2}\log\left(\frac{2|H|}{\delta}\right)$*

Proof. We want to bound the probability that there is a function in the class that overfits, that is we want $Pr(\exists h \in H : |err_S(h) - err(h)| > \epsilon) \leq \delta$.
Let us first find a bound for $Pr(|err_S(h) - err(h)| > \epsilon)$ for a fixed $h \in H$.

$$Pr(|err_S(h) - err(h)| > \epsilon) = Pr(|\frac{1}{m}\sum_{i=1}^{m} Z_i - err(h)| > \epsilon)$$

Where $m = |S|$ and $Z_i = I(h(x_i) \neq y_i)$ is a random variable such that $E[Z_i] = Pr(h(x_i) \neq y_i) = err(h)$. Applying Hoeffding's Inequality

$$Pr(|err_S(h) - err(h)| > \epsilon) \leq 2e^{-2m\epsilon^2}$$

Now

$$\begin{aligned} Pr(\exists h \in H : |err_S(h) - err(h)| > \epsilon) &\leq \sum_{h\in|H|} Pr(|err_S(h) - err(h)| > \epsilon) \\ &\leq \sum_{h\in|H|} 2e^{-2m\epsilon^2} \\ &= |H|2e^{-2m\epsilon^2} \end{aligned}$$

Making this last quantity be $\leq \delta$ we have $m \geq \frac{1}{2\epsilon^2}\log\left(\frac{2|H|}{\delta}\right)$ □

Corollary 1.1. *Let h be the output of ERM and let $f_{best} = \arg\min_{f \in H} err(f)$. With probability $\geq 1 - \delta$ we have that $err(h) \leq err(f_{best}) + 2\epsilon$*

Proof. Let h be the output of ERM, then we have that

$$err(h) \leq err_S(h) + \epsilon \leq err_S(f_{best}) + \epsilon \leq err(f_{best}) + 2\epsilon$$

Where the first and last inequalities come from theorem 1 and the second inequality comes from the fact that $h = \arg\min_{f \in H} err_S(f)$. □

Theorem 1 can also be rewritten as

Theorem 2. *Given a finite function class H, for any $\epsilon > 0$, $\delta > 0$, any distribution D over $\mathcal{X}$ and any target function h^*. With probability $\geq 1 - \delta$ we have that*

$$\forall h \in H \ |err_S(h) - err(h)| \leq \sqrt{\frac{1}{2m} \log\left(\frac{2|H|}{\delta}\right)}$$

where the m is the size of the training set S.

This is known as *uniform convergence.*

1.2 PAC learning infinite function classes

So far, all the theorems we have presented rely on the fact that the size of the function class H is finite since we use a union bound argument in the proofs. We want to see what happens if the size of the function class H is infinite.
For example, consider $H = \{\text{sgn}(x - a) : a \in \mathbb{R}\}$, $h^* \in H$ and
$S = \{(x_1, h^*(x_1)), (x_2, h^*(x_2)), \ldots, (x_m, h^*(x_m))\}$ where $x_1 < x_2 < \ldots < x_m$. Clearly, the size of H is infinite, however notice that for any $a \in (x_i, x_i + 1)$ the empirical error of $h_a(x) = \text{sgn}(x - a)$ is going to be the same. That is, we only need to bound the failure probability of one function per interval which makes the search space for a function in H to be finite.
Following the above example, we introduce the following definition

Definition 1.1. *For a class of functions H, we define $C[m]$ to be the maximum number of different labeling of any set S of size m using functions from H.*

For example, if H is the class of sign functions defined above then $C[1] = 2$ and $C[2] = 3$. This definition is used in the following two theorems, which we will prove in the next lecture.

Theorem 3. *Given any function class H, for any $\epsilon > 0$, $\delta > 0$, any distribution D over $\mathcal{X}$ and any target function $h^* \in H$. With probability $\geq 1 - \delta$ we have that for all $h \in H$ if $errs(h) = 0$ then $err(h) \leq \epsilon$ provided the size of the training set $m = |S|$ is at least $\frac{2}{\epsilon} \log\left(\frac{2C[2m]}{\delta}\right)$*

Theorem 4. *For any function classes H we have that $C[m] = 2^m$ always or,*

$$C[m] = \begin{cases} 2^m & m \leq d \\ m^d & m > d \end{cases}$$

Where $d = VC\text{-}dim(H)$ is the VC-dimension of H.