LECTURE #16

MARCH 31, 2016

LECTURER: ELAD HAZAN SCRIBES: NASSIM FEDEL AND ESTHER ROLF

Today's lecture is motivated by a fundamental question of statistical learning: *which hypothesis classes are learnable, and with what sample complexity?* So far, we've seen that all finite hypothesis classes are learnable, and some, but not all, infinite ones are. Thus, the cardinality of the hypothesis class does not give a tight characterization of when a problem is learnable. In this lecture, we will define the concept of *VC-dimension*, a quantity which can be determined for both finite and infinite hypothesis classes, and which precisely determines when a problem is learnable. For the purpose of this lecture, we consider boolean learning classification tasks, where the label set $Y = \{0, 1\}$.

1 VC Theory

Define $\mathcal{H}_C$ to be the restriction of $\mathcal{H}$ onto C, i.e.:

$$\mathcal{H}_C := \{h : C \to Y = \{0, 1\} : h \text{ is the restriction of some } \hbar \in \mathcal{H}\}$$

Definition 1.1 (Shattering)**.** Let $C \subseteq X$. We say that C is *shattered* by $\mathcal{H}$ when

$$|\mathcal{H}_C| = 2^{|C|}$$

Note that we always have $|\mathcal{H}_C| \leq 2^{|C|}$. So C is shattered when the hypothesis class can fully represent all functions on C.

Example 1.2. Let $\mathcal{H}$ be the set of intervals on the real line (i.e. inside the interval is classified as 1). Let $X = \mathbb{R}$ be the real line, and let $Y = \{0, 1\}$. Then the set $\{-1, 0\}$ is clearly shattered by $\mathcal{H}$; for example, $\mathcal{H}$ contains the intervals $(-\frac{3}{2}, -\frac{1}{2})$, $(-\frac{1}{2}, \frac{1}{2})$, $(-\frac{3}{2}, \frac{1}{2})$ and $(\frac{1}{2}, \frac{3}{2})$. However, a set of three points, for example, $\{-1, 0, 1\}$ is *not* shattered by $\mathcal{H}$, since any interval containing -1 and 1 must contain 0, so the map sending -1 to 1, 0 to 0, and 1 to 1 is not represented by $\mathcal{H}$.

Definition 1.3 (VC-dimension)**.** The VC-dimension of $\mathcal{H}$ is the maximal cardinality m such that there exists a subset $C \subseteq X$ such that $|C| = m$ and C is shattered.

So for example, the hypothesis class in example 1.2 has VC-dimension 2.

Example 1.4. Let $X = \mathbb{R}^2$ and $\mathcal{H}$ be the set of axis-aligned rectangles (classify inside the rectangle as 1). The VC-dimension is 4: for example, take C to be $\{(-1, 0), (1, 0), (0, -1), (0, 1)\}$.

The definition of VC-dimension applies to all types of hypothesis classes.

Example 1.5 (binary decision trees)**.** Let $\mathcal{H}$ be the set of decision trees of size 3 (depth 3) on n variables. A single hypothesis $h \in \mathcal{H}$ can fully represent functions on 3 of the n variables, but no more, so VC-dimension($\mathcal{H}$)=3.

Example 1.6 (infinite VC-dimension)**.** Let $\mathcal{H}$ be the set of all convex polygons in Euclidean space. This has infinite VC dimension, since the cardinality of shattered subsets is unbounded: for example take equally spaced points on the unit sphere. You can place arbitarily many points on the surface of the sphere, and still find a convex polygon which includes any subset of those points, and excludes the rest.

2 Tight characterization of (boolean) statistical learnability

Theorem 2.1 (Fundamental theorem of statistical/PAC learning). *A learning problem $(X, Y, \mathcal{H}, l)$ is PAC-learnable if and only if VC-dim($\mathcal{H}$) $< \infty$. Furthermore, if this holds, a sufficient sample complexity is $\frac{\text{VC-dim}(\mathcal{H})}{\varepsilon^2} \log \frac{1}{\delta}$. That is, if we have finite VC-Dimension, then for any $\varepsilon, \delta > 0$, $err(h_{ERM}) \leq \varepsilon$ with probability $\geq 1 - \delta$ for a sample of size $\frac{\text{VC-dim}(\mathcal{H})}{\varepsilon^2} \log \frac{1}{\delta}$.*

Note that for finite VC-dim($\mathcal{H}$) $\leq \log |\mathcal{H}|$.

Also, theorem 2.1 implies that if VC-dim($\mathcal{H}$) $= \infty$, then $\mathcal{H}$ is not learnable. This follows the same intuition as the no free lunch theorem, that without restricting $\mathcal{H}$, we cannot learn.

Proof of fundamental theorem of statistical learning:

Proof. Define the growth function:

$$\tau_{\mathcal{H}}(m) := \max_{C \subseteq X, |C| = m} \{|\mathcal{H}_C|\}$$

Let $d :=$ VC-dim($\mathcal{H}$). For $m \leq d$, we have $\tau_{\mathcal{H}}(m) = 2^m$. For $m > d$, we prove Sauer's lemma: that $\tau_{\mathcal{H}}(m) = O(m^d)$.

Lemma 2.2 (Sauer's lemma). *For $m \geq d$, we have*

$$\tau_{\mathcal{H}}(m) \leq \begin{Bmatrix} m \\ d \end{Bmatrix} = \sum_{i=1}^{d} \binom{m}{i} = O(m^d)$$

Proof. We use induction on $m + d$ to prove the left inequality. Consider the base case $m + d = 0$. If $|\mathcal{H}| > 1$, there exists $x \in X$ and $h_1, h_2 \in \mathcal{H}$ such that $h_1(x) \neq h_2(x)$, which means that $\{x\}$ is shattered, which would mean $d \geq 1$. Thus, we have $\tau_{\mathcal{H}}(m) \leq 1$.

Now assume that the statement holds for any $m + d = k$. We show that it holds for any $m + d = k + 1$. Fix such $m, d \geq 0$ such that $m + d = k + 1$, and define $m_0 = m - 1$. Take any $C = \{x_1, \ldots, x_{m_0+1}\} \subseteq X$ of size $m = m_0 + 1$ such that $|\mathcal{H}_C| = \tau_{\mathcal{H}}(m_0 + 1)$. Further, for any $h \in \mathcal{H}_C$, define $h|_{m_0}$ to be the restriction of h to $C \setminus \{x_{m_0+1}\}$. We now define the following two sets of restrictions of hypotheses:

$$\mathcal{H}_1 := \{h|_{m_0} : h \in \mathcal{H}_C\}$$
$$\mathcal{H}_2 := \{h \in \mathcal{H}_C : h(x_{m_0+1}) = 1 \text{ and } \exists \hbar \in \mathcal{H}_C \text{ with } h|_{m_0} = \hbar|_{m_0} \text{ and } \hbar(x_{m_0+1}) = 0\}$$

Conceptually, the idea is to break up the complexity of $\mathcal{H}_C$ into two parts: complexity coming from the first m_0 elements, and the complexity by virtue of including x_{m_0+1} in the set.

We first claim that

$$|\mathcal{H}_C| = |\mathcal{H}_1| + |\mathcal{H}_2|$$

To see this, for each $h \in \mathcal{H}_C$, we are in one of two cases: no other $\hbar \in \mathcal{H}_C$ is equal to h on the first m_0 elements, or exactly one other $\hbar \in \mathcal{H}_C$ is equal to h on the first m_0 elements (i.e. $h|_{m_0} = \hbar|_{m_0}$). In the first case, h is counted exactly once by $\mathcal{H}_1$ and $\mathcal{H}_2$ (it's not in $\mathcal{H}_2$, and its restriction to $C \setminus \{x_{m_0+1}\}$ is counted once by $\mathcal{H}_1$). In the second case, we have $h \neq \hbar$, and this pair of distinct hypotheses from $\mathcal{H}_C$ is counted exactly twice by $\mathcal{H}_1$ and $\mathcal{H}_2$: their common restriction to $C \setminus \{x_{m_0+1}\}$ is counted once by $\mathcal{H}_1$, and whichever one of h and $\hbar$ classifies x_{m_0+1} as 1 is counted once by $\mathcal{H}_2$.

Further, by definition, we have

$$|\mathcal{H}_1| = |\mathcal{H}_{C \setminus \{x_{m_0+1}\}}| \leq \tau_{\mathcal{H}}(m_0) \leq \begin{Bmatrix} m_0 \\ d \end{Bmatrix} \text{ by induction hypothesis}$$

We also have

$$\text{VC-dim}(\mathcal{H}_2) \leq \text{VC-dim}(\mathcal{H}_C) - 1 \leq \text{VC-dim}(\mathcal{H}) - 1$$

The right inequality follows since $\mathcal{H}_C \subseteq \mathcal{H}$. To see the left inequality, let C_2 be a subset of C of maximal cardinality that is shattered by $\mathcal{H}_2$. Since everything in $\mathcal{H}_2$ classifies x_{m_0+1} as 1, clearly $x_{m_0+1} \notin C_2$ and $|C_2 \cup \{x_{m_0+1}\}| = |C_2| + 1$. But $C_2 \cup \{x_{m_0+1}\}$ is clearly shattered by $\mathcal{H}_C$, since everything in $\mathcal{H}_2$ has a related hypothesis in $\mathcal{H}_C$ that agrees with it on the first m_0 elements but maps x_{m_0+1} to 0.

Thus,

$$|\mathcal{H}_2| = \left|(\mathcal{H}_2)_{C\setminus\{x_{m_0+1}\}}\right| \leq |\tau_{\mathcal{H}_2}(m_0)| \leq \begin{Bmatrix} m_0 \\ d-1 \end{Bmatrix} \text{ by induction hypothesis}$$

where the leftmost equality follows since everything in $\mathcal{H}_2$ classifies x_{m_0+1} in the same way, so they're fully determined by their classification on $C \setminus \{x_{m_0+1}\}$.

Putting it all together, we get

$$|\mathcal{H}_C| \leq \begin{Bmatrix} m_0 \\ d \end{Bmatrix} + \begin{Bmatrix} m_0 \\ d-1 \end{Bmatrix} \leq \begin{Bmatrix} m_0+1 \\ d \end{Bmatrix} \quad \leftarrow \text{ see homework}$$

as desired. □

The second part of the proof of the theorem is to show that

$$\text{err}(h_{\text{ERM}}) \sim \frac{\log \tau_{\mathcal{H}}(2m)}{m}$$

To show this, we take two samples $S, S' \sim \mathcal{D}$ over X, where the two samples have size m.

Let A be the event that there exists $h \in \mathcal{H}$ such that $\text{err}_{\mathcal{D}}(h) > \varepsilon$ and $\text{err}_S(h) = 0$, and let B be the event that there exists $h \in \mathcal{H}$ such that $\text{err}_{S'}(h) \geq \frac{\varepsilon}{2}$ and $\text{err}_S(h) = 0$. Then note that

$$\begin{aligned}
\mathbb{P}[A] &= \mathbb{P}[A \mid B]\mathbb{P}[B] + \mathbb{P}[A \mid B^c]\mathbb{P}[B^c] \\
&\geq \mathbb{P}[A \mid B]\mathbb{P}[B] \\
&\geq \frac{1}{2}\mathbb{P}[B] \quad \leftarrow \text{ we will show this next time}
\end{aligned}$$

The proof will be completed in the next lecture.

□