Learning Theory: Lecture Notes

Lecturer: Kamalika Chaudhuri Scribe: Qiushi Wang

October 9, 2015

1 The Agnostic PAC Model

Recall that one of the constraints of the PAC model is that the data distribution D has to be separable with respect to the hypothesis class $\mathcal{H}$. The Agnostic PAC model removes this restriction. That is, there no longer exists a $h \in \mathcal{H}$ with $\mathsf{err}_D(h) = 0$.

Definition 1 (Agnostic PAC Model) *A hypothesis class $\mathcal{H}$ is said to be Agnostic PAC-Learnable if there is an algorithm A with the following property. For all ϵ, δ, $0 \leq \epsilon, \delta \leq \frac{1}{2}$, all distributions D over $\mathcal{X} \times \mathcal{Y}$, if A is given ϵ, δ and $m_{\mathcal{H}}(\epsilon, \delta)$ examples from D, then with probability $\geq 1 - \delta$, it outputs a $h \in H$ with:*

$$\underset{D}{\mathsf{err}}(h) \leq \epsilon + \inf_{h^* \in \mathcal{H}} \underset{D}{\mathsf{err}}(h^*)$$

The learning procedure in the PAC model is to find a hypothesis in $\mathcal{H}$ which is consistent with all the input examples. In the Agnostic PAC model, there is no such hypothesis. Instead, a common learning procedure is to find a hypothesis h that minimizes the *empirical error*, or the error on the training examples.

Suppose that given a set of samples S drawn from a data distribution D, h^* minimizes the empirical error $\mathsf{err}(h, S)$ while h_{opt} minimizes the true error $\mathsf{err}_D(h)$.

$$h^* = \arg\min_{h \in \mathcal{H}} \mathsf{err}(h, S) \qquad \text{and} \qquad h_{\text{opt}} = \arg\min_{h \in \mathcal{H}} \underset{D}{\mathsf{err}}(h).$$

Our goal is to find the condition under which $\mathsf{err}_D(h^*) \leq \varepsilon + \mathsf{err}_D(h_{\text{opt}})$.

Lemma 1 *For a fixed $h \in \mathcal{H}$ and m samples S drawn from D,*

$$\mathbb{P}\left(|\underset{D}{\mathsf{err}}(h) - \mathsf{err}(h, S)| \geq \varepsilon \right) \leq 2e^{-m\varepsilon^2}.$$

PROOF: Let $S = \{(x_1, y_1), \ldots, (x_m, y_m)\}$ be the sample set, and let $Z_i = \mathbb{1}(h(x_i) \neq y_i)$ for any $h \in \mathcal{H}$. Then,

$$\mathbb{E}[Z_i] = \underset{D}{\mathsf{err}}(h) \qquad \text{and} \qquad \mathsf{err}(h, S) = \frac{1}{m} \sum_i Z_i.$$

The bound then follows directly from applying Hoeffding's Inequality. □

Theorem 1 *For a finite hypothesis class $|\mathcal{H}|$,*

$$\mathbb{P}\left(\underset{D}{\mathsf{err}}(h^*) - \underset{D}{\mathsf{err}}(h_{opt}) \geq \varepsilon \right) \leq 2|\mathcal{H}|e^{-m\varepsilon^2/4}.$$

Proof: First observe that $\mathsf{err}_D(h^*) - \mathsf{err}_D(h_{\text{opt}})$ can be split into three terms

$$\underset{D}{\mathsf{err}}(h^*) - \underset{D}{\mathsf{err}}(h_{\text{opt}}) = \left(\underset{D}{\mathsf{err}}(h^*) - \mathsf{err}(h^*, S)\right) + (\mathsf{err}(h^*, S) - \mathsf{err}(h_{\text{opt}}, S)) + \left(\mathsf{err}(h_{\text{opt}}, S) - \underset{D}{\mathsf{err}}(h_{\text{opt}})\right).$$

The middle term, $(\mathsf{err}(h^*, S) - \mathsf{err}_D(h_{\text{opt}})) \le 0$, because h^* minimizes $\mathsf{err}(h, S)$. Thus

$$\underset{D}{\mathsf{err}}(h^*) - \underset{D}{\mathsf{err}}(h_{\text{opt}}) \le 2 \sup_{h \in \mathcal{H}} |\underset{D}{\mathsf{err}}(h) - \mathsf{err}(h, S)|.$$

The theorem then results from combining this with the previous lemma, and applying an Union Bound over all $h \in \mathcal{H}$:

$$\mathbb{P}\left(\sup_{h \in \mathcal{H}} |\underset{D}{\mathsf{err}}(h) - \mathsf{err}(h, S)| \ge \frac{\varepsilon}{2}\right) \le \sum_{h \in \mathcal{H}} \mathbb{P}\left(|\underset{D}{\mathsf{err}}(h) - \mathsf{err}(h, S)| \ge \frac{\varepsilon}{2}\right) \le 2|\mathcal{H}|e^{-m\varepsilon^2/4}.$$

□

For failure probability $\le \delta$, the bound in Theorem 1 can be re-written as:

$$\varepsilon(m) \le 2 \cdot \sqrt{\frac{\ln(2|\mathcal{H}|/\delta)}{m}} \tag{1}$$

Contrast this with the analogous bound for PAC learning:

$$\varepsilon(m) \le \frac{\ln(|\mathcal{H}|/\delta)}{m} \tag{2}$$

Thus, Agnostic PAC learning is statistically harder than PAC learning. Usually it is also computationally harder as well.

2 Bounds for Infinite Hypothesis Classes

The generalization bounds we have proved so far apply to finite hypothesis classes, because the union bound step breaks down when $\mathcal{H}$ is infinite. We will now see how we can exploit the structure of a hypothesis class to show generalization bounds which apply infinite classes as well.

What kind of structure can we exploit? In cases where a hypothesis class is infinite, many different hypotheses can produce the same labeling so often the set of meaningful hypotheses is much smaller. We will measure the complexity a hypothesis class by the richness of the labelings it can produce.

This notion can be made formal by the *VC dimension*. Assuming binary classification, that is $\mathcal{Y} = \{0, 1\}$, for a hypothesis class $\mathcal{H}$, and a set of examples $S = \{x_1, \ldots, x_m\}$, we define:

$$\Pi_{\mathcal{H}}(S) = \{(h(x_1), \ldots, h(x_m)) \mid h \in \mathcal{H}\}.$$

Here $\mathcal{H}$ may be infinite but $\Pi_{\mathcal{H}}(S)$ has at most 2^m possible elements, and under certain conditions on $\mathcal{H}$, $\Pi_{\mathcal{H}}(S)$ may have even less.

Definition 1 *We say a hypothesis class $\mathcal{H}$* shatters *S if $\Pi_{\mathcal{H}}(S) = \{0, 1\}^m$.*

Definition 2 *The* VC dimension *of $\mathcal{H}$ is the size of the largest set of examples that can be shattered by $\mathcal{H}$. The VC dimension is infinite if for all m, there is a set of m examples shattered by $\mathcal{H}$.*

Example 1: Bidirectional Thresholds. Let $\mathcal{X} = \mathbb{R}$ with $\mathcal{H} = \mathbb{R} \times \{+, -\}$. Here each example is a point on a line, and has a binary label. Each hypothesis in $\mathcal{H}$ corresponds to a threshold t and a sign ($+$ or $-$), and can be written as $h_{\{t,+\}}$ or $h_{\{t,-\}}$, defined as follows:

$$\begin{aligned} h_{\{t,+\}}(x) &= +, \qquad x \geq t \\ &= -, \qquad \text{otherwise} \end{aligned}$$

In other words, $h_{\{t,+\}}$ labels everything to the right of t as $+$ and everything else as $-$, and $h_{\{t,-\}}$ is defined correspondingly. Since t can take on any real value, $\mathcal{H}$ is infinite.

Note that on any fixed set of points $S = \{x_1, x_2, \ldots, x_m\}$ of size m, $|\Pi_{\mathcal{H}}(S)| \leq 2m$.Consider the following $m + 1$ intervals:

$$(-\infty, x_1), (x_1, x_2), (x_2, x_3), \ldots, (x_{m-2}, x_{m-1}), (x_{m-1}, x_m), (x_m, \infty) \tag{3}$$

Two thresholds t and t' placed in the same interval and with the same sign would result in the same labeling; moreover the pairs $h_{\{-\infty,+\}}$ and $h_{\{\infty,-\}}$ as well as $h_{\{-\infty,-\}}$ and $h_{\{\infty,+\}}$ result in the same labelling. Thus there are $\leq 2m$ distinct labelings.

What is the VC dimension of this class? Thresholds can produce all possible labels on a set of two distinct points. However on a sequence of three points, they cannot label the sequence $+, -, +$ or $-, +, -$. Thus no sets of size 3 are shattered, and the VC dimension of this hypothesis class is 2.

Example 2: Intervals on the line. Let $\mathcal{X} = \mathbb{R}$ with $\mathcal{H} = \mathbb{R} \times \mathbb{R}$. Samples again label points on the line and each hypothesis corresponds to two real values defining an interval; points inside the interval are labeled $+$ and everything else is labeled $-$. Formally, for each interval $[a, b]$, $h_{[a,b]}(x) = +$ for $a \leq x \leq b$, and $-$ otherwise.

For any set $S = \{x_1, \ldots, x_m\}$ of m points,$|\Pi_{\mathcal{H}}(S)| = \binom{m+1}{2} + 1$. Any two hypotheses $h_{[a,b]}$ and $h_{[a',b']}$ where a and a' (or b and b') lie in the same interval in the sequence in Equation 3 produce the same labeling of S. Thus there are $\leq \binom{m+1}{2}$ distinct labelings of S where not all data points are labeled $-$, corresponding to hypotheses $h_{[a,b]}$ where a and b lie in different intervals in the sequence in Equation 3. Finally, we add the all $-$ labelling which is achieved by $h_{[a,a]}$ for any a.

What is the VC dimension of intervals? Intervals can label any sequence of two distinct points but cannot label a sequence of three distinct points $+, -, +$. Thus the VC dimension of $\mathcal{H}$ is 2. If $\mathcal{H}$ is expanded to allow bidirectional intervals, the previous sequence could then be labeled but sequences such as $+, -, +, -$ could not be, giving a VC dimension of 3.

Example 3: Linear Classifiers. Let $\mathcal{X} = \mathbb{R}^2$ with $\mathcal{H} = \{\text{linear classifiers over } \mathbb{R}^2\}$. Consider a set S of 3 points in general position. Figure 2 shows that all possible labelings of S are achievable by $\mathcal{H}$. Thus there exists a set of 3 points that can be shattered by $\mathcal{H}$.

On the other hand, it can be shown that no set of 4 distinct points on the plane can be shattered by $\mathcal{H}$. Thus the VC dimension of $\mathcal{H}$ is 3. Note that a set of 3 collinear points on the plane cannot be shattered by $\mathcal{H}$ because the labeling $+, -, +$ is not achievable by $\mathcal{H}$; but this does not change the VC dimension calculation because there is a set of size 3 that can be shattered.

In general, the VC dimension for the hypothesis class of linear classifiers in $\mathbb{R}^d$ is $d + 1$.

Theorem 2 *For any finite hypothesis class $\mathcal{H}$,* VC-dim$(\mathcal{H}) \leq \log_2 |\mathcal{H}|$.

PROOF: If $\mathcal{H}$ shatters S then $|\mathcal{H}|$ is at least 2^m meaning the VC dimension can be at most $\log_2 |\mathcal{H}|$. $\square$

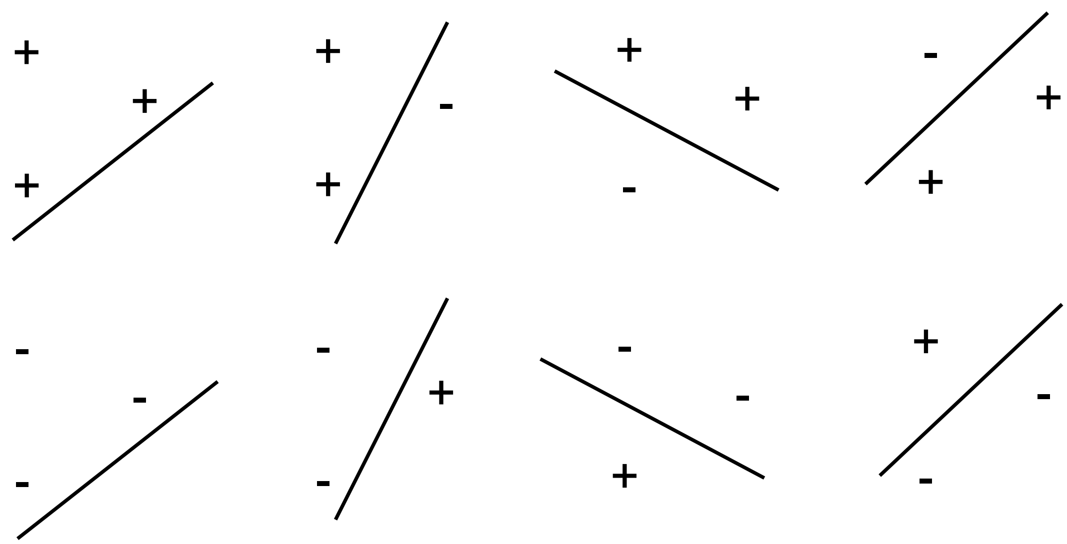

Figure 1: All possible labelings of S are achievable by the class of linear classifiers on the plane.

Example 3: Infinite VC dimension. Let $\mathcal{X} = \mathbb{R}$ and $\mathcal{H} = \mathbb{R}$. For $w \in \mathbb{R}$ a hypothesis is given by

$$h_w(x) = \text{sign}(\sin(wx)).$$

For all m, the set $S = \{2^1, 2^2, \ldots, 2^m\}$ is shattered by h. To see this, let $w = -\pi * (0.y_1 y_2 \ldots y_m)$ be a decimal binary encoding of a set of desired labels, converting -1 to 0. Essentially each x_i bit shifts w to produce the desired label as a result of the fact that $\text{sign}(\sin(\pi z)) = (-1)^{\lfloor z \rfloor}$. Thus the VC dimension of this hypothesis class is infinite.

2.1 Sauer's Lemma

Sauer's Lemma formally relates the VC dimension of a hypothesis class $\mathcal{H}$ and the size of $\Pi_{\mathcal{H}}(S)$ for any set S of examples of size m.

Lemma 2 *If the VC dimension for a hypothesis class $\mathcal{H}$ is d then for a set of m samples S, where $m \geq d$,*

$$|\Pi_{\mathcal{H}}(S)| \leq \sum_{i=0}^{d} \binom{m}{i} \leq \left(\frac{em}{d}\right)^d \in O(m^d)$$

Proof: We will prove this by induction over m and d. Let $\Phi_d(m) = \sum_{i=0}^{d} \binom{m}{i}$. The two base cases:

- When $m = 0$, S is the empty set so $|\Pi_{\mathcal{H}}(S)| \leq 1$ and $\Phi_d(0) = 1$.
- When $d = 0$, $\mathcal{H}$ cannot even shatter one point so only one labeling is possible and $|\Pi_{\mathcal{H}}(S)| = \Phi_0(m) = 1$.

Then, assuming Sauer's Lemma holds for $(m-1, d)$ and $(m-1, d-1)$, we wish to show $|\Pi_{\mathcal{H}}(S)| \leq \Phi_d(m)$.

Let $S = \{x_1, \ldots, x_m\}$. In what follows, we restrict ourselves to the sample space S. Restriction to S can only decrease the VC dimension of $\mathcal{H}$, so it does not affect the theorem statement.

We start by splitting $|\Pi_{\mathcal{H}}(S)|$ through introducing two new hypothesis classes $\mathcal{H}_1$ and $\mathcal{H}_2$ defined on samples $S' = \{x_1, \ldots, x_{m-1}\}$. $\mathcal{H}_1$ is identical to $\mathcal{H}$ but ignores the last example x_m while $\mathcal{H}_2$ consists of only those hypotheses where duplicates differing only on x_m would occur in $\mathcal{H}$. A sample split could be as follows:

$$
\begin{array}{c|ccccc|c|cccc|c|cccc}
 & & & \mathcal{H} & & & & & \mathcal{H}_1 & & & & & \mathcal{H}_2 & & \\
 & x_1 & x_2 & x_3 & x_4 & x_5 & & x_1 & x_2 & x_3 & x_4 & & x_1 & x_2 & x_3 & x_4 \\
h_1 & 0 & 1 & 1 & 0 & 0 & \rightarrow & 0 & 1 & 1 & 0 & & & & & \\
h_2 & 0 & 1 & 1 & 0 & 1 & & & & & & \rightarrow & 0 & 1 & 1 & 0 \\
h_3 & 0 & 1 & 1 & 1 & 0 & \rightarrow & 0 & 1 & 1 & 1 & & & & & \\
h_4 & 1 & 0 & 0 & 1 & 0 & \rightarrow & 1 & 0 & 0 & 1 & & & & & \\
h_5 & 1 & 0 & 0 & 1 & 1 & & & & & & \rightarrow & 1 & 0 & 0 & 1 \\
h_6 & 1 & 1 & 0 & 0 & 1 & \rightarrow & 1 & 1 & 0 & 0 & & & & &
\end{array}
$$

If a set is shattered by $\mathcal{H}_1$, it is also shattered by $\mathcal{H}$. Thus

$$\text{VC-dim}(\mathcal{H}_1) \leq \text{VC-dim}(\mathcal{H}) = d.$$

If S' is shattered by $\mathcal{H}_2$, then $S' \cup \{x_m\}$ is shattered by $\mathcal{H}$ implying

$$\text{VC-dim}(\mathcal{H}_1) \leq \text{VC-dim}(\mathcal{H}) - 1 = d - 1.$$

With this split, $|\Pi_{\mathcal{H}}(S)| = |\Pi_{\mathcal{H}_1}(S')| + |\Pi_{\mathcal{H}_2}(S')|$. Let ℓ be any labeling of $S \setminus \{x_m\}$ achievable by $\mathcal{H}$; if $(\ell, +)$ and $(\ell, -)$ both occur in $\Pi_{\mathcal{H}}(S)$, then ℓ occurs in both $\mathcal{H}_1$ and $\mathcal{H}_2$; otherwise, ℓ occurs only in $\mathcal{H}_1$.

So by the inductive hypothesis,

$$
\begin{aligned}
|\Pi_{\mathcal{H}}(S)| \leq \Phi_d(m-1) + \Phi_{d-1}(m-1) &= \sum_{i=0}^{d} \binom{m-1}{i} + \sum_{i=0}^{d-1} \binom{m-1}{i} \\
&= \sum_{i=0}^{d} \binom{m-1}{i} + \sum_{i=1}^{d} \binom{m-1}{i-1} = \sum_{i=1}^{d} \binom{m}{i} = \Phi_d(m).
\end{aligned}
$$

Finally, from Sterling's approximation, for when $m \geq d$,

$$\Phi_d(m) = \sum_{i=0}^{d} \binom{m}{i} \leq \left(\frac{m}{d}\right)^d \sum_{i=0}^{d} \binom{m}{d} \left(\frac{d}{m}\right)^i = \left(\frac{m}{d}\right)^d \left(1 + \frac{d}{m}\right)^m \leq \left(\frac{em}{d}\right)^d.$$

$\square$