TTIC 31290: Machine Learning for Algorithm Design (Fall 2025)
Avrim Blum and Dravyansh Sharma

Lecture 6: 10/16/25 Lecturer: Dravyansh Sharma

Outline for today

- Agnostic PAC learning
- Learning from a finite collection of algorithms
- VC-dimension and pseudo-dimension
- Sauer's Lemma

1 Agnostic PAC learning

So far we have assumed that there is a perfect concept $c^* \in C = H$ that correctly labels all the examples. But this may not be true, e.g., if the labels are noisy or if the learner's hypothesis space H does not contain any hypothesis that correctly labels all the examples. We will now revise our definitions to handle this.

Definition 1. *(Agnostic PAC learning). A learner receives a sample*

$$S = \{(x_1, y_1), \ldots, (x_m, y_m)\},$$

where each (x_i, y_i) is drawn i.i.d. from an unknown distribution D over $X \times \{0, 1\}$. The algorithm outputs a hypothesis $h \in H$.

The error *of a hypothesis h with respect to D is*

$$\text{err}_D(h) = \Pr_{(x,y)\sim D}[h(x) \neq y].$$

We say that the hypothesis class H is agnostic PAC (Probably Approximately Correct) learnable *if there exists a polynomial function $m_C(\epsilon, \delta)$ and a learner such that for all $\epsilon, \delta \in (0, 1)$, and for all distributions D on $X \times \{0, 1\}$, when the learner is given $m \geq m_C(\epsilon, \delta)$ i.i.d. labeled examples from D, it outputs $\hat{h} \in H$ satisfying*

$$\Pr\left[\text{err}_D(\hat{h}) - \min_{h\in H}(\text{err}_D(h)) \leq \epsilon\right] \geq 1 - \delta.$$

For finite H we will show that the following natural algorithm is an agnostic PAC learner.

Definition 2. *(Empirical Risk Minimization, ERM). Output $\hat{h} \in \text{argmin}_{h\in H}\text{err}_S(h)$.*

We will now give a bound on the sample complexity of agnostic PAC learning for the ERM learner. We will make use of the concentration inequalities (specifically Hoeffding's bounds) we saw earlier in the course.

Theorem 1. *ERM is an agnostic PAC learner for any finite hypothesis class H, and needs only*

$$\frac{2}{\epsilon^2}\left(\ln(|H|) + \ln\frac{2}{\delta}\right)$$

examples to output a hypothesis of error at most ϵ with probability at least $1-\delta$.

Proof. We will use Hoeffding's inequality applied to indicator random variables $Z_i = \mathbb{I}\{h(x_i) \neq y_i\}$.

Lemma 1 (Hoeffding's bound (special case))**.** *Let $Z_1, \ldots, Z_m$ be i.i.d. Bernoulli random variables with mean $\mu = \mathbb{E}[Z_i]$ and let $\widehat{\mu} = \frac{1}{m}\sum_{i=1}^{m} Z_i$. Then for any $t > 0$,*

$$\Pr\left(|\widehat{\mu} - \mu| > t\right) \leq 2\exp(-2mt^2).$$

For a fixed hypothesis $h \in H$, define $Z_i = \mathbb{I}\{h(x_i) \neq y_i\}$ so that $\mathrm{err}_S(h) = \widehat{\mu}$ and $\mathrm{err}_D(h) = \mu$. Applying Hoeffding's with $t = \epsilon/2$ gives

$$\Pr\Big(|\mathrm{err}_S(h) - \mathrm{err}_D(h)| > \frac{\epsilon}{2}\Big) \leq 2\exp\Big(-2m\frac{\epsilon^2}{4}\Big) = 2\exp\Big(-\frac{m\epsilon^2}{2}\Big).$$

Apply a union bound over all $h \in H$ to control the deviation simultaneously for every hypothesis:

$$\Pr\Big(\exists h \in H:\ |\mathrm{err}_S(h) - \mathrm{err}_D(h)| > \frac{\epsilon}{2}\Big) \leq 2|H|\exp\Big(-\frac{m\epsilon^2}{2}\Big).$$

Choose m so that the right-hand side is at most δ, i.e.

$$2|H|\exp\Big(-\frac{m\epsilon^2}{2}\Big) \leq \delta \quad\Longleftrightarrow\quad m \geq \frac{2}{\epsilon^2}\ln\Big(\frac{2|H|}{\delta}\Big).$$

Hence, when m satisfies this bound, with probability at least $1-\delta$ we have the *uniform convergence* property

$$\forall h \in H: \quad |\mathrm{err}_S(h) - \mathrm{err}_D(h)| \leq \frac{\epsilon}{2}.$$

Condition on this high-probability event. By definition of ERM, for any $h \in H$ and therefore also for $h^* \in \mathrm{argmin}_{h\in H}\mathrm{err}_D(h)$,

$$\mathrm{err}_S(\hat{h}) \leq \mathrm{err}_S(h^*).$$

Using uniform convergence we get

$$\mathrm{err}(\hat{h}) \leq \mathrm{err}_S(\hat{h}) + \frac{\epsilon}{2} \leq \mathrm{err}_S(h^*) + \frac{\epsilon}{2} \leq \mathrm{err}(h^*) + \frac{\epsilon}{2} + \frac{\epsilon}{2} = \mathrm{err}(h^*) + \epsilon.$$

Thus with probability at least $1-\delta$ the ERM hypothesis $\hat{h}$ satisfies the agnostic guarantee $\mathrm{err}(\hat{h}) \leq \min_{h\in H}\mathrm{err}_D(h) + \epsilon$, as claimed. □

2 Learning from a finite collection of algorithms

The above argument also extends to selecting algorithms from a finite collection of algorithms. The key difference is that we need to use the appropriate concentration inequality.

Lemma 2 (Hoeffding's bound)**.** *Let $Z_1, \ldots, Z_m$ be i.i.d. bounded random variables with values in $[0, U]$ with mean $\mu = \mathbb{E}[Z_i]$ and let $\widehat{\mu} = \frac{1}{m}\sum_{i=1}^{m} Z_i$. Then for any $t > 0$,*

$$\Pr\left(|\widehat{\mu} - \mu| > t\right) \leq 2\exp\left(-\frac{2mt^2}{U^2}\right).$$

We have the following analogous result for uniform convergence (and therefore the sample complexity) of selecting the best algorithm from a finite algorithm family.

Theorem 2. *Let $\mathcal{A}$ be a finite set of algorithms, with utility $u(x, A) \in [0, 1]$ for every input $x \in \Pi$ and algorithm $A \in \mathcal{A}$, and D be an input distribution over instances in Π. Fix $\epsilon, \delta \in (0, 1)$. Let $m \geq \frac{U^2}{2\epsilon^2}\left(\ln|\mathcal{A}| + \ln\frac{2}{\delta}\right)$. Then with probability at least $1 - \delta$ over the draw of i.i.d. sample $(x_1, \ldots, x_m) \sim D^m$,*

$$\left|\frac{1}{m}\sum_i u(x_i, A) - \mathbb{E}_{x\sim D}[u(x, A)]\right| < \epsilon$$

for every algorithm $A \in \mathcal{A}$.

Note that uniform convergence is also an interesting result in itself. It implies that even if we don't find the best algorithm on the "training sample" of input instances but just one that achieves small sample error, as long as the sample complexity of uniform convergence is met, we can guarantee that the "test error" on any future instance will be small (within ϵ of the training error). This will be even more important for infinite algorithm/hypothesis classes (where exact ERM can be challenging to implement), which we will look at next.

3 VC-dimension and pseudo-dimension

Definition 3 (Shattering)**.** *Let X be an instance space and $H \subseteq \{0, 1\}^X$ a hypothesis class. We say that H* shatters *a finite set $S = \{x_1, \ldots, x_m\} \subseteq X$ if for every labeling $y = (y_1, \ldots, y_m) \in \{0, 1\}^m$, there exists some $h \in H$ such that*

$$h(x_i) = y_i \quad \textit{for all } i = 1, \ldots, m.$$

Intuitively, this means H is rich enough to realize all possible labelings of S — it can "fit" any assignment of 0s and 1s on those m points.

Example. [Intervals on the real line] Let $X = \mathbb{R}$ and

$$H_{\text{int}} = \{h_{a,b}(x) = \mathbb{I}\{a \leq x \leq b\} : a, b \in \mathbb{R}, a \leq b\}.$$

- H_{int} can shatter any set of 2 points (e.g., pick two distinct real numbers $x_1 < x_2$ and check that all four labelings can be achieved by appropriate intervals).

- However, it *cannot* shatter any set of 3 points, since the labeling $(1, 0, 1)$ is impossible (an interval cannot cover the first and third point but not the middle one).

Hence, the largest set H can shatter has size 2.

Definition 4 (VC Dimension)**.** *The* Vapnik–Chervonenkis (VC) dimension *of a hypothesis class H, denoted* $\mathrm{VCdim}(H)$*, is the size of the largest finite subset $S \subseteq X$ that is shattered by H. If H can shatter arbitrarily large finite sets, we say* $\mathrm{VCdim}(H) = \infty$*.*

Remark 1 (VC dimension of intervals)**.** *For the interval class H_{int} above, we found that H can shatter any set of size* 2 *but not* 3*. Thus,*

$$\mathrm{VCdim}(H_{\text{int}}) = 2.$$

Intuitively, the VC dimension measures the expressive power or flexibility of a hypothesis class. A larger VC dimension means that the class can represent more complex decision boundaries and hence can fit more varied patterns in data. However, a very large VC dimension can lead to overfitting — the model can perfectly fit training data even when it does not generalize.

Definition 5 (Pseudodimension)**.** *Let $\mathcal{F}$ be a class of real-valued functions $f : X \to \mathbb{R}$. We say that $\mathcal{F}$* pseudoshatters *a set $S = \{x_1, \ldots, x_m\}$ if there exist real numbers (thresholds)*

$$r_1, r_2, \ldots, r_m \in \mathbb{R}$$

such that for every labeling $y \in \{0, 1\}^m$, there exists some $f \in \mathcal{F}$ satisfying

$$\forall i \in \{1, \ldots, m\} : \quad \mathbb{I}\{f(x_i) \geq r_i\} = y_i.$$

The pseudodimension *of $\mathcal{F}$, denoted* $\mathrm{Pdim}(\mathcal{F})$*, is the size of the largest set that $\mathcal{F}$ pseudoshatters.*

Intuitively, the pseudodimension generalizes the VC dimension from binary-valued hypotheses to real-valued ones. It captures the richness of a function class in terms of how many independent thresholds it can realize simultaneously.

Remark 2. *When $\mathcal{F}$ takes only binary values $\{0, 1\}$, pseudodimension reduces exactly to VC dimension.*

4 Sauer's Lemma

To reason about generalization for infinite hypothesis classes, we use the number of distinct labelings realizable on a finite sample.

Definition 6 (Growth Function)**.** *For a hypothesis class H, define the* growth function

$$\Pi_H(m) = \max_{S \subseteq \mathcal{X}, |S| = m} |\{(h(x_1), \ldots, h(x_m)) : h \in H\}|.$$

That is, $\Pi_H(m)$ counts the maximum number of distinct labelings H can induce on m points.

Theorem 3 (Sauer's Lemma)**.** *Let $H \subseteq \{0,1\}^{\mathcal{X}}$ with* $\mathrm{VCdim}(H) = d$*. Then for any integer $m \geq d$,*

$$\Pi_H(m) \leq \sum_{i=0}^{d} \binom{m}{i}.$$

In particular, $\Pi_H(m) = O(m^d)$ when $m \geq d$.

Proof. We will proceed by induction on both m, d.

Base cases. (a) $d = 0$, any m. No set of points can be shattered, so all points can be labeled only in one way. $\Pi_H(m) = 1 = \sum_{i=0}^{0} \binom{m}{i}$.

(b) $m = 0$, any d. RHS is 1, which bounds the number of labelings on a set of zero points.

Inductive step. $d > 0, m > 0$. Suppose the claim holds for all d', m' such that $d' + m' < d + m$. Let $S = \{x_1, \ldots, x_m\} \subseteq \mathcal{X}$. Let H_S be a set of hypotheses in H that achieve $\Pi_H(m)$ labelings of S.

Fix arbitrary $x \in S$, say x_m, and let $S' = S \setminus \{x_m\}$. Define H' as the smallest subset of H_S that labels S' maximally. If two hypotheses in H_S label S' in the same way, then exactly one is in H' (chosen to be smallest).

Now, we claim that $\mathrm{VCDim}(H_S \setminus H') \leq d - 1$. Note that if $H_S \setminus H'$ shatters $T' \subseteq S'$, then H_S shatters $T' \cup \{x_m\}$. This is by construction of H' (Think why? Using maximality, if there is an $h \in H_S \setminus H'$ for which there is no corresponding $h' \in H'$ that agrees on S' and differs exactly on x_m, then it could be added to H').

On the other hand, $H' \subseteq H_S$, and so $\mathrm{VCDim}(H') \leq d$.

By the induction hypothesis:

$$\begin{aligned}
\Pi_H(m) = |H_S| = |H'| + |H_S/H'| &\leq \sum_{i=0}^{d} \binom{m-1}{i} + \sum_{i=0}^{d-1} \binom{m-1}{i} \\
&= \binom{m-1}{0} + \sum_{i=1}^{d} \left(\binom{m-1}{i} + \binom{m-1}{i-1} \right) \\
&= \binom{m}{0} + \sum_{i=1}^{d} \binom{m}{i},
\end{aligned}$$

where we have used the Pascal identity $\binom{m}{i} = \binom{m-1}{i} + \binom{m-1}{i-1}$. This completes the induction. □

Sauer's Lemma shows that even though there may be infinitely many hypotheses, the number of distinct labelings on a finite sample grows only polynomially with m once VC dimension is bounded. This combinatorial control is the key to generalization, as we will see in the next lecture.

Additional Resources:

- A. Anthony and P. L. Bartlett, *Neural Network Learning: Theoretical Foundations*, Cambridge University Press, 1999.
- S. Shalev-Shwartz and S. Ben-David, *Understanding Machine Learning: From Theory to Algorithms*, Cambridge University Press, 2014.