Computational Learning Theory

- Brief Overview of Machine Learning
- Consistency Model
- Probably Approximately Correct Learning
- **Sample Complexity and Occam's Razor**
- Dealing with Noises
- ...

Valiant's Theorem

Basic question on sample complexity

Say we want to PAC-learn $\mathcal{C}$ using $\mathcal{H}$, how many examples are sufficient?

Theorem

If learner can produce a hypothesis $h \in \mathcal{H}$ consistent with

$$m \geq \frac{1}{\epsilon} \log \left(\frac{|\mathcal{H}|}{\delta} \right)$$

examples, then

$$\text{Prob}[\text{err}_{\mathcal{D}}(h) \leq \epsilon] \geq 1 - \delta.$$

i.e., it is a PAC-learner

A Proof of Valiant's Theorem

- Call a hypothesis h *bad* if $\text{err}_{\mathcal{D}}(h) > \epsilon$
- Let h be any bad hypothesis, then

$$\text{Prob}[h \text{ consistent with } m \text{ i.i.d. examples}] < (1-\epsilon)^m$$

- Noting that the hypothesis produced by learner is consistent with m i.i.d. examples, thus by union bound

$$\begin{aligned}
& \text{Prob}[\text{Learner outputs a bad hypothesis}] \\
\leq\ & \text{Prob}[\text{some } h \in \mathcal{H} \text{ is bad and is consistent with } m \text{ i.i.d. examples}] \\
\leq\ & |\mathcal{H}|(1-\epsilon)^m \\
\leq\ & \delta
\end{aligned}$$

last inequality holds because

$$m \geq \frac{1}{\epsilon} \log\left(\frac{|\mathcal{H}|}{\delta}\right)$$

Some Consequences of Valiant's Theorem

Corollary

Learning BOOLEAN CONJUNCTIONS *only need* $\frac{1}{\epsilon}\log\left(\frac{3^n}{\delta}\right)$ *samples. (Thus, the learner is an efficient PAC-learner!)*

Corollary

If learner can produce a hypothesis $h \in \mathcal{H}$ *consistent with* m *examples, then*

$$\text{Prob}\left[\text{err}_{\mathcal{D}}(h) \leq \frac{1}{m}\log\left(\frac{|\mathcal{H}|}{\delta}\right)\right] \geq 1-\delta$$

Interpretation:

- $\text{err}_{\mathcal{D}}(h)$ gets smaller when m gets larger, because there's more data to learn from
- $\text{err}_{\mathcal{D}}(h)$ gets smaller when $|\mathcal{H}|$ gets smaller. The more we know about the concept, the smaller the hypothesis class becomes, thus the better the learning error

Occam's Razor

Theorem (Occam's Razor, Roughly stated)

If a learner always produce a hypothesis $h \in \mathcal{H}$ *with* $|h| = O((n|c|)^\alpha m^\beta)$ *for some fixed* α *(arbitrary) and* $0 < \beta < 1$, *then it is an efficient PAC-learner.*

Proof.

The set of all hypotheses that the learner can possibly output is relatively "small" since each such hypothesis has small size.
Apply Valiant's theorem. □

What Happens if $\mathcal{H}$ is Infinite?

Natural question

What if $|\mathcal{H}|$ is more than exponential or even infinite? How many (i.i.d.) samples from $\mathcal{D}$ do we need given ϵ, δ?

V. Vapnik and A. Chervonenkis. *"On the uniform convergence of relative frequencies of events to their probabilities.* Theory of Probability and its Applications, 16(2):264280, 1971.

gave a very original and important answer.

VC-Dimension Intuitively

- VC-Dimension of a function class measure how "complex" and "expressive" the class is
- Roughly, $\text{VCD}(\mathcal{H})$ is the maximum number of data points for which no matter how we label them, there's always $h \in \mathcal{H}$ consistent with them
- VC used this to derive bounds for expected loss given empirical loss
- Since VCD is defined in terms of model fitting and number of data points, the concept applies to almost all imaginable models
- It's a much better indicator of models' ability than number of parameters

VC-Dimension Rigorously

- Since $h : \Omega \to \{0, 1\}$, h can be viewed as a subset of Ω
- For any finite $S \subseteq \Omega$, let $\Pi_{\mathcal{H}}(S) = \{h \cap S : h \in \mathcal{H}\}$
- We call $\Pi_{\mathcal{H}}(S)$ the *projection* of $\mathcal{H}$ on S
- Equivalently, suppose $S = \{x_1, \dots, x_m\}$, let

$$\Pi_{\mathcal{H}}(S) = \{[h(x_1), \dots, h(x_m)] \mid h \in \mathcal{H}\}$$

  we call $\Pi_{\mathcal{H}}(S)$ the set of all *dichotomies* (also called *behaviors*) on S *realized by* (or *induced by*) $\mathcal{H}$
- S is *shattered* by $\mathcal{H}$ if $|\Pi_{\mathcal{H}}(S)| = 2^{|S|}$

Definition (VC-dimension)

$\mathrm{VCD}(\mathcal{H}) = \max\{|S| : S \text{ shattered by } \mathcal{H}\}$.

VC-Dimension: Examples

- Set of all positive half lines on $\mathbb{R}$ has $\text{VCD} = 1$
- Set of all intervals on $\mathbb{R}$ has $\text{VCD} = 2$
- Set of all half-planes on $\mathbb{R}^2$ has $\text{VCD} = 3$
- Set of all half-spaces on $\mathbb{R}^d$ has $\text{VCD} = d + 1$
- Set of all balls on $\mathbb{R}^d$ has $\text{VCD} = d + 1$
- Set of all axis-parallel rectangles on $\mathbb{R}^2$ has $\text{VCD} = 4$
- Set of all d-vertex convex polygons on $\mathbb{R}^2$ has $\text{VCD} = 2d + 1$
- Set of all sets of intervals on $\mathbb{R}$ has $\text{VCD} = \infty$

VC-Dimension: Sauer's Lemma

Lemma (Sauer 1972, Perles & Shelah 1972)

Suppose $\text{VCD}(\mathcal{H}) = d < \infty$. *Define*

$$\Pi_{\mathcal{H}}(m) = \max\{|\Pi_{\mathcal{H}}(S)| : S \subseteq \Omega, |S| = m\}$$

($\Pi_{\mathcal{H}}(m)$ *is the maximum size of a projection of* $\mathcal{H}$ *on an* m*-subset of* Ω*.)*
Then,

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

(Note that, if $\text{VCD}(\mathcal{H}) = \infty$, then $\Pi_{\mathcal{H}}(m) = 2^m, \forall m$)

A Proof of Sauer's Lemma

- Induct on $m + d$. For $h \in \mathcal{H}$, let $h_S = h \cap S$
- $m = 0$ is obvious. When $d = 0$, $|\Pi_{\mathcal{H}}(S)| = 1 = \Phi_0(m)$
- Consider $m > 0, d > 0$. Fix arbitrary $s \in S$.
- Define

$$\mathcal{H}' = \{h_S \in \Pi_{\mathcal{H}}(S) \mid s \notin h_S,\ h_S \cup \{s\} \in \Pi_{\mathcal{H}}(S)\}$$

- Then,

$$|\Pi_{\mathcal{H}}(S)| = |\Pi_{\mathcal{H}}(S - \{s\})| + |\mathcal{H}'| = |\Pi_{\mathcal{H}}(S - \{s\})| + |\Pi_{\mathcal{H}'}(S)|$$

- Since $\text{VCD}(\mathcal{H}') \le d - 1$,

$$|\Pi_{\mathcal{H}}(S)| \le \Phi_d(m-1) + \Phi_{d-1}(m) = \Phi_d(m).$$

Vapnik-Chervonenkis Theorem

Theorem

Suppose $\text{VCD}(\mathcal{H}) = d < \infty$. *There's a constant* $c_0 > 0$ *such that, if a learner can produce a hypothesis* $h \in \mathcal{H}$ *consistent with*

$$m \geq \frac{c_0}{\epsilon}\left(\log\left(\frac{1}{\delta}\right) + d\log\left(\frac{1}{\epsilon}\right)\right)$$

i.i.d. examples, then it is a PAC-learner, i.e.

$$\text{Prob}[\text{err}_{\mathcal{D}}(h) \leq \epsilon] \geq 1 - \delta.$$

Proof of Vapnik-Chervonenkis Theorem

- Consider a concept c and a hypothesis class $\mathcal{H}$
- Suppose our algorithm outputs a hypothesis consistent with c on m i.i.d. examples $S = \{x_1, \dots, x_m\}$
- Let $h\Delta c$ denote the *symmetric difference* between h and c,

$$\begin{aligned} \Delta(c) &= \{h\Delta c \mid h \in \mathcal{H}\} \\ \Delta_\epsilon(c) &= \{r \mid r \in \Delta(c),\ \operatorname*{Prob}_{x \leftarrow \mathcal{D}}[x \in r] > \epsilon\} \end{aligned}$$

- Then, for any $h \in \mathcal{H}$, $\text{err}_{\mathcal{D}}(h) > \epsilon$ (i.e. h is "bad") iff $h\Delta c \in \Delta_\epsilon(c)$
- S is called an *ϵ-net* if $S \cap r \neq \emptyset$ for every $r \in \Delta_\epsilon(c)$
- If S is an ϵ-net, then the output hypothesis is good! Thus,

$$\begin{aligned} & \text{Prob}[\text{Algorithm outputs a bad hypothesis}] \\ \leq\ & \text{Prob}[S \text{ is not an } \epsilon\text{-net}] \\ =\ & \text{Prob}[\exists r \in \Delta_\epsilon(c) \text{ s.t. the } m \text{ i.i.d. examples } S \text{ does not "hit" } r] \end{aligned}$$

Proof of Vapnik-Chervonenkis Theorem

- Let A be the event that there some *region* $r \in \Delta_\epsilon(c)$ which S does not hit. We want to upper bound Prob$[A]$
- Suppose we draw m more i.i.d. examples $T = \{y_1, \dots, y_m\}$ (for analytical purposes, the learner does not really draw T)
- Let B be the event that there some *region* $r \in \Delta_\epsilon(c)$ which S does not hit but T does hit r at least $\epsilon m/2$ times
- Now, for any $r \in \Delta_\epsilon(c)$ that S does not hit, Prob$[y_i \in r] > \epsilon$. Hence, by Chernoff bound, when $m \geq 8/\epsilon$, the probability that at least $\epsilon m/2$ of the y_i belong to r is at least $1/2$.
- Consequently, Prob$[B \mid A] \geq 1/2$.
- Thus, Prob$[A] \leq 2\,$Prob$[B]$.
- We can thus upper bound Prob$[B]$ instead!

Proof of Vapnik-Chervonenkis Theorem

Why is upper-bounding B easier?

- B is the event that, after drawing $2m$ i.i.d. examples $S \cup T = \{x_1, \dots, x_m, y_1, \dots, y_m\}$, there exists some region $r \in \Delta_\epsilon(c)$ which S does not hit but T hits $\geq \epsilon m/2$ times.
- Equivalently, B is the event that, after drawing $2m$ i.i.d. examples $S \cup T = \{x_1, \dots, x_m, y_1, \dots, y_m\}$, there exists some region $r \in \Pi_{\Delta_\epsilon(c)}(S \cup T)$ for which $S \cap r = \emptyset$ and $|T \cap r| \geq \epsilon m/2$.
  It is not difficult to see that
  $$|\Pi_{\Delta_\epsilon(c)}(S \cup T)| \leq |\Pi_{\Delta(c)}(S \cup T)| = |\Pi_{\mathcal{H}}(S \cup T)| \leq \left(\frac{2me}{d}\right)^d$$

- Prob$[B]$ remains the same if we draw $2m$ examples $U = \{u_1, \dots, u_{2m}\}$ first, and then partition U randomly into $U = S \cup T$.

Proof of Vapnik-Chervonenkis Theorem

- Fix U and $r \in \Pi_{\Delta_\epsilon(c)}(U)$. Let $p = |U \cap r|$.
- Let F_r be the event that $S \cap r = \emptyset, |T \cap r| \geq \epsilon m/2$. We can assume $\epsilon m/2 \leq p \leq m$. Then

$$
\begin{aligned}
\text{Prob}[F_r \mid U] &= \frac{\binom{2m-p}{m}}{\binom{2m}{m}} \\
&= \frac{(2m-p)(2m-p-1)\cdots(m-p+1}{2m(2m-1)\cdots(m+1)} \\
&= \frac{m(m-1)\cdots(m-p+1}{2m(2m-1)\cdots(2m-p+1)} \\
&\leq \left(\frac{1}{2}\right)^p \\
&\leq \frac{1}{2^{\epsilon m/2}}
\end{aligned}
$$

Proof of Vapnik-Chervonenkis Theorem

(In the following, if the underlying distribution $\mathcal{D}$ on Ω is continuous, replace the sum by the corresponding integral, and the probability by the density function.)

$$\begin{aligned}
\text{Prob}[B] &= \text{Prob}\left[\exists r \in\in \Delta_\epsilon(c) \text{ such that } F_r \text{ holds}\right] \\
&= \sum_U \text{Prob}\left[\exists r \in \Delta_\epsilon(c) \text{ such that } F_r \text{ holds} \mid U\right] \text{Prob}[U] \\
&= \sum_U \text{Prob}\left[\exists r \in \Pi_{\Delta_\epsilon(c)}(U) \text{ such that } F_r \text{ holds} \mid U\right] \text{Prob}[U] \\
&\leq \sum_U \left(\frac{2me}{d}\right)^d 2^{-\epsilon m/2} \text{Prob}[U] \\
&= \left(\frac{2me}{d}\right)^d 2^{-\epsilon m/2}
\end{aligned}$$

Proof of Vapnik-Chervonenkis Theorem

$$\mathsf{Prob}[A] \leq 2\,\mathsf{Prob}[B] \leq 2\left(\frac{2me}{d}\right)^d 2^{-\epsilon m/2} \leq \delta.$$

When

$$m \geq \frac{c_0}{\epsilon}\left(\log\left(\frac{1}{\delta}\right) + d\log\left(\frac{1}{\epsilon}\right)\right)$$

(We will need ϵ bounded away from 1, say $\epsilon < 3/4$, for c_0 to not be dependent on ϵ, but that's certainly desirable!)

The Lower Bound

Theorem

For any sample space Ω and any concept class $\mathcal{C}$ with $\text{VCD}(\mathcal{C}) = d$, there exist a distribution $\mathcal{D}$ on it, and a concept $c \in \mathcal{C}$ such that, **any** *learning algorithm which takes $\leq d/2$ samples will* **not** *be a PAC-learner with $\epsilon = 1/8, \delta = 1/7$. such that*

The Proof

- Suppose $X \subseteq \Omega$ is shattered by $\mathcal{C}$, $|X| = d$
- Let $\mathcal{D}$ be the uniform distribution on X, thus $\mathcal{D}$ is 0 on $\Omega - X$.
- Without loss of generality, we can assume $\mathcal{C} = 2^X$

Proof idea

Use the argument from expectation!

Pick $c \in \mathcal{C}$ at random, show that the expected performance of the learner (over the random choice c) is "bad," which implies that there exists a $c \in \mathcal{C}$ for which the performance is bad.

- Let S denote a random sample of $\leq d/2$ examples
- Let x denote a random example
- Let h_S denote the hypothesis output by the learner if its examples are S

The Proof

$$\operatorname*{Prob}_{c,S,x}[h_S(x) \neq c(x)] \geq \operatorname*{Prob}_{c,S,x}[h_S(x) \neq c(x) \mid x \notin S] \operatorname*{Prob}_{c,x,S}[x \notin S] \geq \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}$$

- Marginalizing over c, we have

$$\operatorname*{Prob}_{c,S,x}[h_S(x) \neq c(x)] = \mathsf{E}_c \left[\operatorname*{Prob}_{S,x}[h_S(x) \neq c(x) \mid c] \right].$$

  Thus, there exists a $c \in \mathcal{C}$ such that $\text{Prob}_{S,x}[h_S(x) \neq c(x) \mid c] \geq \frac{1}{4}$.
- For this fixed c, we have $\text{Prob}_{S,x}[h_S(x) \neq c(x)] \geq \frac{1}{4}$.
- Now, marginalizing over S, we have

$$\frac{1}{4} \leq \operatorname*{Prob}_{S,x}[h_S(x) \neq c(x)] = \mathsf{E}_S \left[\operatorname*{Prob}_{S,x}[h_S(x) \neq c(x) \mid S] \right] = \mathsf{E}_S[\text{err}(h_S)]$$

The Proof

Thus,

$$\mathsf{E}_S[1 - \mathsf{err}(h_S)] = 1 - \mathsf{E}_S[\mathsf{err}(h_S)] \leq 3/4.$$

By Markov's inequality,

$$\mathop{\mathsf{Prob}}_S[1 - \mathsf{err}(h_S) \geq 7/8] \leq \frac{\mathsf{E}_S[1 - \mathsf{err}(h_S)]}{7/8} \leq \frac{3/4}{7/8} = \frac{6}{7}.$$

Thus,

$$\mathop{\mathsf{Prob}}_S\left[\mathsf{err}(h_S) < \frac{1}{8}\right] \leq \frac{6}{7},$$

as desired.