Vapnik-Chervonenkis Theorem and the Double Sampling Trick

Results in this lecture are from [1, 2].

1 Sample complexity for infinite hypothesis classes

The next theorem is an analog of Valiant's theorem for infinite hypothesis classes.

Theorem 1.1 (Sample complexity for infinite hypothesis classes.)**.** *Suppose* $\text{VCD}(\mathcal{H}) = d < \infty$. *There is a universal constant* c_0 *such that, if a learner can always produce a hypothesis 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 the learner is a PAC-learner.

Proof. When the hypothesis class is (potentially) infinite, the union bound is useless. The VC proof resolves this problem by "projecting down" to a finite case.

Define

$$\Delta(c) = \{h\Delta c \;:\; h \in \mathcal{H}\},$$

and

$$\Delta_\epsilon(c) = \left\{ r \in \Delta(c) \;:\; \operatorname*{Prob}_{x \leftarrow \mathcal{D}}[x \in r] > \epsilon \right\}.$$

In words, $\Delta_\epsilon(c)$ consists of all "error regions" r which are denser than ϵ. As in the proof of Valiant's theorem, if our sample set S "hits" all regions in $\Delta_\epsilon(c)$ and the learner outputs a hypothesis h_S consistent with S, then h_S is a good hypothesis. Such a sample set S is called an *ϵ-net*. In summary,

$$\begin{aligned} \operatorname*{Prob}_S[h_S \text{ is a good hypothesis}] &\geq \operatorname*{Prob}_S[S \text{ is an } \epsilon\text{-net}] \\ &= \operatorname*{Prob}_S[S \text{ hits every region in } \Delta_\epsilon(c)]. \end{aligned}$$

So we will bound the probability that S forms an ϵ-net. Remember that S consists of m i.i.d. examples taken according to the (unknown) distribution $\mathcal{D}$. Here's a very important trick, called the *double sampling trick*. Suppose we take m more i.i.d. examples T. (These are examples taken for the analytical purposes, the learner does not take them in practice.)

- Let B be the event that S is not an ϵ-net, namely B is the event that S misses some region $r \in \Delta_\epsilon(c)$
- Let C be the event that S misses some region $r \in \Delta_\epsilon(c)$ but T hits that region r more than $\epsilon m/2$ times. To be a little more precise, C is the event that there exists some $r \in \Delta_\epsilon(c)$ for which S misses entirely and T hits $> \epsilon m/2$ times.

Since C cannot happen without B, we have

$$\begin{aligned}\operatorname*{Prob}_{S,T}[C] &= \operatorname*{Prob}_{S,T}[C|B]\operatorname*{Prob}_{S,T}[B] + \operatorname*{Prob}_{S,T}[C|\bar{B}]\operatorname*{Prob}_{S,T}[\bar{B}] \\ &= \operatorname*{Prob}_{S,T}[C|B]\operatorname*{Prob}_{S,T}[B] \\ &= \operatorname*{Prob}_{S,T}[C|B]\operatorname*{Prob}_{S}[B].\end{aligned}$$

We estimate $\text{Prob}_{S,T}[C|B]$ first. Conditioned on B, let r be any region in $\Delta_\epsilon(c)$ that S misses. Then,

$$\operatorname*{Prob}_{S,T}[C|B] \geq \operatorname*{Prob}_{T}[T \text{ hits } r \text{ more than } \epsilon m/2 \text{ times}]$$

We know that the probability that an arbitrary sample in T hits r is more than ϵ. Hence, let X be the number of times T hits r, by Chernoff bound we have

$$\operatorname*{Prob}_{S,T}[X \leq \epsilon m/2] \leq e^{-\epsilon m/8}.$$

For $m \geq 8/\epsilon$ the right hand side is at most $1/2$. Thus, for $m \geq 8/\epsilon$ we conclude that $\text{Prob}_{S,T}[C|B] \geq 1/2$, which means

$$\operatorname*{Prob}_{S,T}[C] \geq \frac{1}{2}\operatorname*{Prob}_{S}[B].$$

Thus, instead of trying to upper-bound $\text{Prob}_S[B]$, we can try to upper-bound $\text{Prob}_{S,T}[C]$. Why is upper-bounding $\text{Prob}_{S,T}[C]$ any easier? Well, C is the event that there is some region $r \in \Pi_{\Delta_\epsilon(c)}(S \cup T)$ for which S misses entirely and T hits more than $\epsilon m/2$ times. Note now that the number of regions r in $\Pi_{\Delta_\epsilon(c)}(S \cup T)$ is *finite*. We can apply the union bound!

Now, suppose we take $S = \{s_1, \cdots, s_m\}, T = \{t_1, \cdots, t_m\}$, and then swap s_i, t_i with probability $1/2$. Call the resulting pair S', T'. Then the probability that event C holds with respect to S', T' is the same as the probability that C holds with respect to S, T because all examples are i.i.d..

Fix a region $r \in \Pi_{\Delta_\epsilon(c)}(S \cup T)$ which $S \cup T$ hits $l \geq \epsilon m/2$ times *and* $|r \cap \{s_i, t_i\}| \leq 1$. Then, the probability that S' doesn't hit r and T' hits r l times is exactly $1/2^l \leq 1/2^{\epsilon m/2}$. Thus, by the union bound

$$\begin{aligned}\operatorname*{Prob}_{S',T'}[C] &\leq |\Pi_{\Delta_\epsilon(c)}(S \cup T)|\frac{1}{2^{\epsilon m/2}} \\ &\leq |\Pi_{\Delta(c)}(S \cup T)|\frac{1}{2^{\epsilon m/2}} \\ &= |\Pi_{\mathcal{H}}(S \cup T)|\frac{1}{2^{\epsilon m/2}} \\ &\leq |\Pi_{\mathcal{H}}(2m)|\frac{1}{2^{\epsilon m/2}} \\ &\leq \left(\frac{2em}{d}\right)^d \frac{1}{2^{\epsilon m/2}}\end{aligned}$$

The equality follows because there is a bijection between $\Pi_{\mathcal{H}}(X)$ and $\Pi_{\Delta(c)}(X)$ for any $X \subseteq \Omega$: we map $h \cap X$ to $(h\Delta c) \cap X$ and vice versa. The last inequality follows from Sauer lemma. Overall, we need to pick m such that

$$\text{Prob}[B] \leq 2\text{Prob}[C] \leq 2\frac{(2em/d)^d}{2^{\epsilon m/2}} \leq \delta,$$

which would be satisfied if we pick

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

for sufficiently large c_0. □

Exercise 1. Show that for any $X \subset \Omega$, $|\Pi_{\delta(c)}(X)| = |\Pi_{\mathcal{H}}(X)|$.

2 A lower bound on sample complexity

We will show that $m = \Omega(d/\epsilon)$ samples must be taken in order to PAC-learn a concept class $\mathcal{C}$ with VC-dimension d, where ϵ is any given error parameter, and a constant confidence level $\delta \leq 1/16$. In order to illustrate the main idea, let us prove a slightly weaker result, leaving the general result as an exercise.

Theorem 2.1. *For any sample space Ω and any concept class $\mathcal{C}$ with $\text{VCD}(\mathcal{C}) = d$, there exist a distribution $\mathcal{D}$ on Ω, and a concept $c^* \in \mathcal{C}$ such that, any learning algorithm which takes $\leq d/2$ samples is not a PAC-learner (of $\mathcal{C}$ using $\mathcal{C}$) with parameters $\epsilon = 1/8, \delta = 1/8$.*

Proof. Suppose $X \subseteq \Omega$ is shattered by $\mathcal{C}$, where $|X| = d$. 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. We use the argument from expectation! Pick $c \in \mathcal{C}$ uniformly at random, we will show that the expected performance of the learner (over the random target concept c) is "bad," which implies that there exists a $c \in \mathcal{C}$ for which the performance is bad. Let S denote a random set of examples of $m \leq d/2$ examples, let x denote a random sample, and h_S denote the hypothesis output by the learner if its examples are S. The proof has three steps.

1. Show that $\text{E}_c\left[\text{E}_S[\text{err}(h_S) \mid c]\right] \geq 1/4$.
2. By the argument from expectation, there exists a target concept c^* for which $\text{E}_S[\text{err}(h_S)] \geq 1/4$.
3. Then, by a simple application of Markov's inequality we conclude that $\text{Prob}_S[\text{err}(h_S) \leq 1/8] \leq 6/7 < 1 - \delta$.

Let's implement the above ideas. Note that

$$\text{Prob}_{c,S,x}[h_S(x) \neq c(x) \mid x \notin S] = \text{E}_S\left[\text{E}_x\left[\underbrace{\text{Prob}_c[h_S(x) \neq c(x)]}_{\geq 1/2} \mid x \notin S\right] \mid S\right] \geq 1/2.$$

Hence, from the law of total probability and the fact that $|S| \leq d/2$ we have

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

Now, marginalizing over c we have

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

Thus, there exists a target concept $c^* \in \mathcal{C}$ such that $\text{Prob}_{S,x}[h_S(x) \neq c^*(x)] \geq \frac{1}{4}$. Now, marginalizing over S, we obtain

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

Thus, by linearity of expectation

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

By Markov's inequality,

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

Equivalently, $\text{Prob}_S\left[\text{err}(h_S) \leq \frac{1}{8}\right] \leq \frac{6}{7}$ as desired. $\square$

Exercise 2. In this exercise, we prove a more general lower bound: if the learner only takes $\Omega(d/\epsilon)$ i.i.d. examples then it can not PAC-learn a concept class $\mathcal{C}$ with VC-dimension d and error parameter ϵ, confidence parameter $\delta = 1/15$.

Fix a subset $X \subset \Omega$ of size $|X| = d$ such that X is shattered by $\mathcal{C}$. Let $X = \{\omega_0, \omega_1, \ldots, \omega_{d-1}\}$. Fix $\epsilon \in (0, 1/16)$ and $\delta = 1/15$. Define the distribution $\mathcal{D}$ on X where $\mathcal{D}$ assigns a mass of $1 - 16\epsilon$ to ω_0 and and a mass of $16\epsilon/(d-1)$ to *each* of $\omega_1, \ldots, \omega_{d-1}$. Clearly $\mathcal{D}$ is a distribution on X which is also a distribution on Ω. Without loss of genearlity, we can also assume that $\mathcal{C} = 2^X$.

We will show that there exists a target concept $c^* \in \mathcal{C}$ such that if a learner only takes $m = \frac{d-1}{64\epsilon}$ i.i.d. examples. then it cannot PAC-learn $\mathcal{C}$ under the data distribution $\mathcal{D}$ and parameters $\epsilon, \delta = 1/15$.

Let S denote the multiset of sample points taken from $\mathcal{D}$, where $|S| = m = \frac{d-1}{64\epsilon}$. Random concepts $c \in \mathcal{C}$ are taken uniformly. Let $X' = \{\omega_1, \ldots, \omega_{d-1}\}$.

(a) Prove that $\text{Prob}_{x,S}[x \in X' \setminus S] \geq 4\epsilon$.

(**Hint:** Let T_S denote the number of times S hits X'. Observing the following:

$$\text{Prob}_{x,S}[x \in X' \setminus S] = \text{Prob}_{x,S}[x \in X' \setminus S \mid T_S \leq (d-1)/2]\text{Prob}[T_S \leq (d-1)/2].$$

Use Markov's inequality to show that $\text{Prob}[T_S > (d-1)/2] \leq 1/2$.)

(b) Let h_S denote the hypothesis the learner outputs given the examples S. Show that

$$\text{Prob}_{x,S,c}[h_S(x) \neq c(x) \wedge x \in X'] \geq 2\epsilon.$$

(c) Define $\text{err}'(h) = \text{Prob}_x[h(x) \neq c(x) \wedge x \in X']$. Show that,

$$2\epsilon \leq \text{E}_S[\text{err}'(h_S)]$$

(d) By writing

$$\text{E}_S[\text{err}'(h_S)] = \text{E}_S[\text{err}'(h_S) \mid \text{err}'(h_S) > \epsilon]\text{Prob}[\text{err}'(h_S) > \epsilon] + \text{E}_S[\text{err}'(h_S) \mid \text{err}'(h_S) \leq \epsilon]\text{Prob}[\text{err}'(h_S) \leq \epsilon],$$

prove that

$$2\epsilon \leq 16\epsilon\text{Prob}[\text{err}'(h_S) > \epsilon] + \epsilon$$

from which we conclude that $\text{Prob}[\text{err}'(h_S) > \epsilon] \geq 1/15$.

(e) Finally, show that

$$\text{Prob}[\text{err}(h_S) > \epsilon] \geq \text{Prob}[\text{err}'(h_S) > \epsilon]$$

to finish the proof.

References

[1] V. N. Vapnik and A. Y. Chervonenkis, *On the uniform convergence of relative frequencies of events to their probabilities*, Doklady Akademii Nauk USSR, 181 (1968).

[2] ———, *On the uniform convergence of relative frequencies of events to their probabilities*, Theory of Probability and its Applications, 16 (1971), pp. 264–280.