8803 Machine Learning Theory

Maria-Florina Balcan Lecture 11: September 27th, 2011

---

1 VC-dimension and Learnability

Definition 1 *The* **Vapnik-Chervonenkis dimension** *of C, denoted as $VCdim(C)$, is the cardinality of the largest set S shattered by C. If arbitrarily large finite sets can be shattered by C, then $VCdim(C) = \infty$.*

Given a class H, define the class $\text{MAJ}_k(H)$ to be the class of functions achievable by taking majority votes over k functions in H. For example, if H is the class of conjunctions and $k = 3$ then a typical function in $\text{MAJ}_k(H)$ might be "$f(x) = 1$ if x satisfies at least two out of three of $x_1x_4x_5$, $x_2x_3x_4$, and x_3x_7." Let's say we allow repetitions.

Claim 1 *Let $MAJ_k(H)$ is the class of functions achievable by taking majority votes over k functions in H. If the hypothesis class H has VC-dimension d, then the class $MAJ_k(H)$ has VC-dimension $O(kd \log kd)$.*

Proof: Let D be the VC-dimension of $\text{MAJ}_k(H)$, so by definition, there must exist a set S of D points shattered by $\text{MAJ}_k(H)$. We know by Sauer's lemma that there are at most D^d ways of partitioning the points in S using functions in H.

Now, since each function h in $\text{MAJ}_k(H)$ is determined by some k functions $h_1, h_2, \ldots, h_k$ in H, this means that the partitioning of S induced by h is determined by the partitioning of S induced by $h_1, \ldots, h_k$. Since there are at most $(D^d)^k = D^{dk}$ ways of selecting k partitions of S consistent with H (possibly with repetitions), this means there are at most D^{kd} ways of partitioning the points in S using functions in $\text{MAJ}_k(H)$.

On the other hand, since S is shattered by $\text{MAJ}_k(H)$, we know all 2^D partitionings are possible. We therefore must have $2^D \leq D^{kd}$, and so $D \leq 2kd \log(kd)$ (for $kd \geq 4$). ■

A General Upper Bound on the Sample Complexity

In previous lectures we have shown that the VC-dimension of a concept class gives an upper bound on the number of samples needed to learn concepts from the class.

For example, we have shown:

Theorem 1 *Let C be an arbitrary hypothesis space of VC-dimension d. Let D be an arbitrary unknown probability distribution over the instance space and let c^* be an arbitrary unknown target function. For any ϵ, $\delta > 0$, if we draw a sample S from D of size m satisfying*

$$m \geq \frac{8}{\epsilon}\left[d \ln\left(\frac{16}{\epsilon}\right) + \ln\left(\frac{2}{\delta}\right)\right].$$

then with probability at least $1-\delta$*, all the hypotheses in* C *with* $err_D(h) > \epsilon$ *are inconsistent with the data, i.e.,* $err_S(h) \neq 0$.

So it is possible to PAC-learn a class C of VC-dimension d with parameters δ and ϵ given that the number of samples m is at least $m \geq c\left(\frac{d}{\epsilon}\log\frac{1}{\epsilon} + \frac{1}{\epsilon}\log\frac{1}{\delta}\right)$ where c is a fixed constant. So, as long as $VCdim(C)$ is finite, it is possible to PAC-learn concepts from C even though C might be infinite.

A Lower Bound on the Sample Complexity

We show that this sample complexity result is tight within a factor of $O(\log(1/\epsilon))$.

Theorem 2 *Any algorithm for PAC-learning a concept class of VC dimension* d *with parameters* ϵ *and* $\delta \leq 1/15$ *must use more than* $(d-1)/(64\epsilon)$ *examples in the worst case.*

Proof: Consider a concept class C with VC dimension d. Let $X = \{x_1, \ldots, x_d\}$ be shattered by C. To show a lower bound we construct a particular distribution that forces any PAC algorithm to take that many examples. The support of this probability distribution is X, so we can assume WLOG that $C = C(X)$, so C is a finite class, $|C| = 2^d$. Note that we have arranged things such that for all possible labelings of the points in X, there is exactly one concept in C that induces that labeling. Thus, choosing the target concept uniformly at random from C is equivalent to flipping a fair coin d times to determine the labeling induced by c on X.

Let $m = (d-1)/(64\epsilon)$, and A be an algorithm that uses at most m i.i.d. examples and then produces a hypothesis h. We need to show that there exist a distribution D on X and a concept $c \in C$ such that the $er(h) > \epsilon$ with probability at least 1/15.

We first define D independently of A:

$$p(x_1) = 1 - 16\epsilon$$
$$p(x_2) = p(x_3) = \cdots = p(x_d) = \frac{16\epsilon}{d-1}$$

In the following we assume that S is a random i.i.d sample from D of size m. We want to establish that there is a c so that $\Pr_S[er(h) > \epsilon] > \frac{1}{15}$.

Let $X' = \{x_2, \ldots, x_d\}$. For any fixed $c \in C$ and hypothesis h, let

$$er'(h) = \Pr[c(x) \neq h(x) \wedge x \in X'].$$

For technical reasons, it is easier to prove that $\Pr_S[er'(h) > \epsilon] > 1/15$, which is enough since $er'(h) \leq er(h)$.

We pick a random $c \in C$ and show that with positive probability c is hard to learn for A, thereby showing that there must be some fixed c that is hard to learn for A.

Let us now define the event:

$B:$ $\quad S$ contains less than $(d-1)/2$ points in X'.

We have:

$$\Pr_S[B] \;\; \geq \;\; 1/2 \tag{1}$$

To see this, let Z be the number of points in S that are from X'. Clearly, $\mathrm{E}[Z] = 16\epsilon m = (d-1)/4$. We have $\Pr_S[B] \geq 1 - \Pr[Z \geq (d-1)/2] \geq 1/2$, since by Markov's inequality we have $\Pr[Z \geq (d-1)/2] \leq 1/2$.

We can also show:

$$\mathrm{E}_{c,S}[\, er'(h) \;|\; B\,] \;\; > \;\; 4\epsilon \tag{2}$$

Let S be the set of points that A gets. Choosing a random c is equivalent to flipping a fair coin for each point in X to determine its label. Since h is independent of the labeling of $X' - S$, the contribution to $er'(h)$ is expected to be $16\epsilon/(2(d-1))$ for each point in $X' - S$. When B occurs, we have $|X' - S| > (d-1)/2$; thus the expected value of $er'(h)$ given B is strictly greater than 4ϵ.

Using (1) and (2) we get a lower bound on $\mathrm{E}_{c,S}[er'(h)]$.

$$\mathrm{E}_{c,S}[er'(h)] \geq \Pr_S[B] \cdot \mathrm{E}_{c,S}[\, er'(h) \;|\; B\,] > \frac{1}{2} \cdot 4\epsilon = 2\epsilon.$$

So there must exist some $c^* \in C$ such that $\mathrm{E}_S[er'(h)] > 2\epsilon$. We take c^* as the target concept and show that A is likely to produce a hypothesis with high error rate.

Using the fact that for any h we have $er'(h) \leq \Pr[x \in X'] = 16\epsilon$ we note that

$$\mathrm{E}_S[\, er'(h) \;|\; er'(h) > \epsilon\,] \;\; \leq \;\; 16\epsilon \;\; \text{for any fixed } c. \tag{3}$$

We have:

$$\begin{aligned} 2\epsilon &< \mathrm{E}_S[er'(h)] \\ &= \Pr_S[er'(h) > \epsilon] \cdot \mathrm{E}_S[\, er'(h) \;|\; er'(h) > \epsilon\,] \\ &\quad + (1 - \Pr_S[er'(h) > \epsilon]) \cdot \mathrm{E}_S[\, er'(h) \;|\; er'(h) \leq \epsilon\,]. \end{aligned}$$

Next we apply (3) to get

$$\begin{aligned} 2\epsilon < \mathrm{E}_S[er'(h)] &\leq \Pr_S[er'(h) > \epsilon] \cdot 16\epsilon + (1 - \Pr_S[er'(h) > \epsilon]) \cdot \epsilon \\ &= 15\epsilon \Pr_S[er'(h) > \epsilon] + \epsilon, \end{aligned}$$

which implies $\Pr_S[er'(h) > \epsilon] > 1/15$, as desired. ■