CS 7545: Machine Learning Theory **Fall 2021**

Lecture 11: VC Dimension

Instructor: Santosh Vempala *Lecture date: 10/18/2021*

Disclaimer: *These notes have not been subjected to the usual scrutiny reserved for formal publications.*

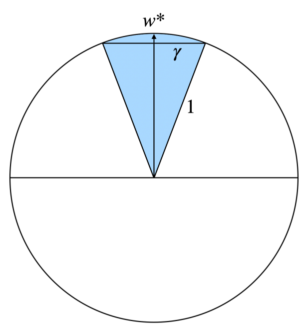
We have seen PAC and mistake bound algorithms for many concept classes. In the case of learning halfspaces, the number of mistakes made by Perceptron or Winnow have a $1/\gamma^2$ dependence on the margin γ. In fact, one can make this $\log(1/\gamma)$.

Suppose we predict the majority of all surviving w, i.e., suppose after examples $x^1, x^2, \ldots, x^\ell$, we have $W = \{w : \|w\| \leq 1, w^\top x^i \geq 0 \text{ (or } w^\top x^i < 0) \ \forall i \leq \ell\}$ as candidates and we consider which of $|W \cap \{w : w^\top x^{\ell+1} \geq 0\}|$, $|W \cap \{w : w^\top x^{\ell+1} < 0\}|$ is larger. Predict according to that.

Then in each step we eliminate 1/2 of the volume. Suppose B is the unit ball of dimension n. $\mathrm{Vol}(W)$ starts at $\mathrm{Vol}(B)$. After m mistakes,

$$\mathrm{Vol}(W) \leq \frac{1}{2^m}\mathrm{Vol}(B). \tag{11.1}$$

At the end, $\mathrm{Vol}(W)$ is at least the volume of the γ-cone.



$$\mathrm{Vol}(B) = \int_0^1 \left(\sqrt{1-t^2}\right)^{n-1} \mathrm{Vol}(B^{n-1})\, dt$$
$$\mathrm{Vol}(\gamma\text{-cap}) = \int_{\sqrt{1-\gamma^2}}^1 \left(\sqrt{1-t^2}\right)^{n-1} \mathrm{Vol}(B^{n-1})\, dt$$

Then we have for some constant c,

$$\frac{\mathrm{Vol}(\gamma\text{-cap})}{\mathrm{Vol}(B)} = \frac{\int_{\sqrt{1-\gamma^2}}^1 \left(\sqrt{1-t^2}\right)^{n-1}\, dt}{\int_0^1 \left(\sqrt{1-t^2}\right)^{n-1}\, dt} \geq c\gamma^n$$
$$\mathrm{Vol}(W) \geq c\gamma^n \mathrm{Vol}(B). \tag{11.2}$$

By (11.1) and (11.2),

$$c\gamma^n \mathrm{Vol}(B) \leq \frac{1}{2^m}\mathrm{Vol}(B)$$
$$m \leq cn\log(1/\gamma).$$

Thus, the number of mistakes is $m = O(n\log(1/\gamma)$.

11.1 VC Dimension

Let $x^{(1)}, x^{(2)}, \ldots, x^{(\ell)}$ be i.i.d. samples from a distribution $\mathcal{D}$. Let $\mathcal{H}$ be a hypothesis class and $h^* \in \mathcal{H}$. Suppose $h \in \mathcal{H}$ such that $h(x^{(i)}) = h^*(x^{(i)})$ for $i = 1, \ldots, m$.

- How many samples m are needed such that $\Pr_D(h(x) \neq h^*(x)) \leq \epsilon$ with probability $1 - \delta$?
- For a sample set S with m points, how many distinct ways can concepts in $\mathcal{H}$ partition (label) S?

Definition 11.1 (VC-dimension)**.** *The VC-dimension of a concept class $\mathcal{H}$ is the largest integer m such that there exists a set of m points that can be shattered by concepts in $\mathcal{H}$. We say that a set S of size m is shattered by $\mathcal{H}$ if S can be labelled in 2^m ways by concepts in $\mathcal{H}$.*

Example 11.2.

- *Intervals in $\mathbb{R}$: VC-dim $= 2$*
- *Axis-paralleled rectangles in $\mathbb{R}^2$: VC-dim $= 4$*
- *Half-spaces in $\mathbb{R}^d$: VC-dim $= d + 1$.*

Theorem 11.3 (Sauer's Lemma)**.** *For a concept class $\mathcal{H}$ with VC-dim d, let $\mathcal{H}(m)$ be the number of distinct ways to label m points using $h \in \mathcal{H}$. Then $\mathcal{H}(m) \leq m^d$.*

Proof. We will show the following by induction on m,

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

The base case is when $m \leq d$. The above is true by the definition of VC-dimension.

Let S be a set of m points and let $x \in S$. Consider the set $S \setminus \{x\}$. Let $\mathcal{H}(S)$ denote the number of ways to split S by concepts in $\mathcal{H}$. By the induction hypothesis, $\mathcal{H}(S \setminus \{x\}) \leq \binom{m-1}{\leq d}$. Note that

$$\binom{m}{\leq d} = \binom{m-1}{\leq d} + \binom{m-1}{\leq d-1}.$$

So it suffices to show that

$$\mathcal{H}(S) - \mathcal{H}(S \setminus \{x\}) \leq \binom{m-1}{\leq d-1}. \tag{11.3}$$

Let $H|_S$ be the concept class of restriction of concepts in $\mathcal{H}$ on S. How can $\mathcal{H}(S)$ be large? There must be labellings h and h' such that they agree on all points in S except x. Let $\mathcal{T} = \{h \in \mathcal{H}|_S : h(x) = 1, \exists h' \in \mathcal{H}|_S \text{ s.t. } h'(x) = 0 \text{ and } h(y) = h'(y), \forall y \in S \setminus \{x\}\}$. Then $\mathcal{H}(S) - \mathcal{H}(S \setminus \{x\}) \leq \mathcal{T}(S \setminus \{x\})$. Suppose VC-dimension of $\mathcal{T}$ is d'. So $2^{d'}$ points can be shattered by $\mathcal{T}$. Then $d' + 1$ points can be shattered by $\mathcal{H}$. So $d' + 1 \leq d$. Then by induction hypothesis, $\mathcal{T}(S \setminus \{x\}) \leq \binom{m-1}{\leq d-1}$, which proves (11.3). □

11.2 Bounding sample complexity by VC dimension

The following concentration inequalities will be helpful in this section.

Theorem 11.4 (Multiplicative Chernoff bound)**.** *Suppose $X_1, X_2, \ldots, X_m$ are independent $0/1$ random variables. Let $X = \sum_{i=1}^{m} X_i$. Then*

$$\Pr(X \geq (1 + \delta)\mathbb{E}X) \leq e^{-\frac{\delta^2}{2+\delta}\mathbb{E}X}$$

$$\Pr(X \leq (1 - \delta)\mathbb{E}X) \leq e^{-\frac{\delta^2}{2}\mathbb{E}X}.$$

Theorem 11.5 (Hoeffding's inequality)**.** *Suppose $X_1, X_2, \dots, X_m$ are independent random variables bounded by $a_i \le X_i \le b_i$. Let $X = \sum_{i=1}^m X_i$. Then*

$$\Pr(X \ge \mathbb{E}X + t) \le e^{-\frac{2t^2}{\sum_{i=1}^m (a_i-b_i)^2}}$$
$$\Pr(X \le \mathbb{E}X - t) \le e^{-\frac{2t^2}{\sum_{i=1}^m (a_i-b_i)^2}}.$$

Theorem 11.6. *The number of examples needed to (ϵ,δ)-PAC learn hypothesis class $\mathcal{H}$ with VC-dim d is at most*

$$\frac{2}{\epsilon}\left(\log(2\mathcal{H}(2m)) + \log(1/\delta)\right) = O\left(\frac{1}{\epsilon}(d\log(1/\epsilon) + \log(1/\delta)\right).$$

Let $\ell(x)$ be the unknown labeling function. The error of $h \in \mathcal{H}$ is defined as

$$\mathrm{err}_D(h) = \Pr_{x\sim\mathcal{D}}(h(x) \neq \ell(x))$$
$$\mathrm{err}_S(h) = \frac{|\{x \in S : h(x) \neq \ell(x)\}|}{|S|}.$$

Theorem 11.7. *If S is a set of i.i.d. samples from $\mathcal{D}$ of size*

$$m \ge \frac{8}{\epsilon^2}\left(\log(2\mathcal{H}(2m)) + \log(1/\delta)\right) = O\left(\frac{1}{\epsilon^2}(d\log(1/\epsilon) + \log(1/\delta)\right),$$

then with probability $1-\delta$, for all $h \in \mathcal{H}$,

$$|\mathrm{err}_S(h) - \mathrm{err}_{\mathcal{D}}(h)| \le \epsilon.$$

Proof of Theorem 11.6. We find a hypothesis h_S that correctly classifies m points. We want to show that with probability at least $1-\delta$.

$$Pr_{\mathcal{D}}(h_S(x) \neq h^*(x)) \le \epsilon.$$

Let A be the event that $\mathrm{err}_S(h) = 0$ and $\mathrm{err}_{\mathcal{D}}(h) > \epsilon$. Consider a different setting where we pick 2 subsets of size m, say S and S'. Let B be the event that $\mathrm{err}_S(h) = 0$ and $\mathrm{err}_{S'}(h) > \epsilon/2$.

Claim 11.8.

$$\Pr(B) \ge \frac{1}{2}\Pr(A).$$

Let $\Pr(B|A)$ be the probability that h has at least $\epsilon/2$ error on m points given that h has at least ϵ error on D. For $i \in [m]$, let X_i be 0/1 random variables such that

$$X_i = \begin{cases} 1 & \text{if } h \text{ makes an error on the } i\text{'th point of } S' \\ 0 & \text{otherwise} \end{cases}$$

So, $\Pr(B|A) = \Pr(\sum_{i=1}^m X_i \ge \epsilon m/2)$. By Chernoff bound,

$$\Pr\left(\sum_{i=1}^m X_i < \mathbb{E}(\sum_{i=1}^m X_i) - \frac{\epsilon m}{2}\right) \le e^{-\frac{\epsilon m}{8}}.$$

For $m \ge 8/\epsilon$, we have

$$\Pr(B) \ge \Pr(A)\Pr(B|A) \ge \frac{1}{2}\Pr(A).$$

By the claim, it suffices to show that $\Pr(B) \le \delta/2$. For this, we pick $2m$ points S''. We partition them S'' into two subsets S, S' of m points each in the following manner. Pair up the $2m$ points $(a_1, b_1), \dots, (a_m, b_m)$

randomly and assign a_i to S and b_i to S' with probability 1/2, and with the remaining 1/2, assign a_i to S' and b_i to S. Now we want to bound $\Pr(\mathrm{err}_S(h) = 0 \text{ and } \mathrm{err}_{S'}(h) > \epsilon/2)$ for a fixed hypothesis h. If h makes error on both a_i and b_i for some index i, then $\Pr(B) = 0$ since no error allowed on S. Also if B occurs, then h must make an error on exactly one of a_i or b_i for at least $\epsilon m/2$ indices i. For an i with error on a_i, the probability that a_i is assigned to S' is 1/2. So,

$$\Pr(B) \leq \Pr(\text{all } \epsilon m/2 \text{ errors fall in } S') \leq \frac{1}{2^{\epsilon m/2}}.$$

Since the number of possible distinct labelling for S'' is at most $\mathcal{H}(2m)$, it suffices to have

$$2^{-\epsilon m/2}\mathcal{H}(2m) \leq \frac{\delta}{2},$$

i.e., $m \geq \frac{2}{\epsilon}\left(\log(2\mathcal{H}(2m)) + \log(1/\delta)\right)$. □

Proof of Theorem 11.7. For a fixed hypothesis $h \in \mathcal{H}$, let A be the bad event that $|\mathrm{err}_S(h) - \mathrm{err}_{\mathcal{D}}(h)| > \epsilon$. Let B be the bad event that $|\mathrm{err}_S(h) - \mathrm{err}_{S'}(h)| > \epsilon/2$ for random subsets S and S' of size m each. By an argument similar to the proof of Claim 11.8, we have $\Pr(B) \geq \frac{1}{2}\Pr(A)$. So it suffices to show that $\Pr(B) \leq \delta/2$. For this, we pick $2m$ points S'', pair up the $2m$ points $(a_1, b_1), \ldots, (a_m, b_m)$, and again partition them into two subsets S, S' of m points each following the same process. If $|\mathrm{err}_S(h) - \mathrm{err}_{S'}(h)| > \epsilon/2$, then for at least $\epsilon m/2$ indices i such that h makes an error on exactly one of a_i or b_i. Now, we want to bound $\Pr(|\mathrm{err}_S(h) - \mathrm{err}_{S'}(h)| > \epsilon/2)$. For indices with exactly one error, define X_i to be random variables such that

$$X_i = \begin{cases} 1 & \text{if the error goes to } S \\ -1 & \text{if the error goes to } S' \end{cases}$$

By Hoeffding's inequality,

$$\Pr(B) \leq \Pr\left(\left|\sum_{i=1}^{m} X_i\right| > \frac{\epsilon m}{2}\right) \leq 2e^{-\frac{\epsilon^2 m}{8}}.$$

Since the number of possible distinct labelling of S'' is at most $\mathcal{H}(2m)$, it suffices to have

$$2e^{-\frac{\epsilon^2 m}{8}}\mathcal{H}(2m) \leq \frac{\delta}{2},$$

i.e., $m \geq \frac{8}{\epsilon^2}\left(\log(2\mathcal{H}(2m)) + \log(1/\delta)\right)$. □