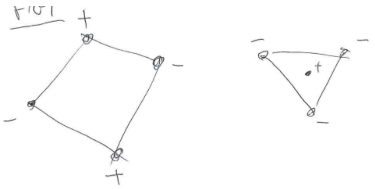
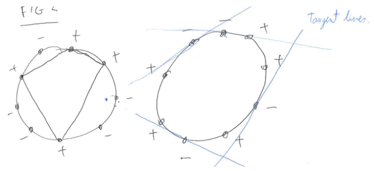
1 Lecture 4: 2017.02.13

Example 1.1 (Hyperplanes in $\mathbb{R}^2$). It is clear that the VCD is at least 3 because given three points, we can always choose a line that separates the points in any way we want. In fact, the VCD is exactly 3.



In general, for hyperplanes in $\mathbb{R}^d$ the VCD dimension is $d + 1$.

Example 1.2 (Convex n-gons in $\mathbb{R}^2$). In this case, $VCD = 2n + 1$. For example, if $n = 4$, $VCD = 9$.



Remark 1.3. *Note that in these geometric situations, the VCD dimension is linearly related to the free parameters in our model. This is generally true (more or less). This supports the idea that we need more samples than parameters to be able to start learning something significant.*

Remark 1.4 (Monotone conjunctions over $\{0,1\}^m$ e.g. $x_1 \wedge x_5 \wedge x_7$)**.** *We claim that* $VCD \geq n$*. Indeed, consider the bit vectors:* $\langle 01\ldots1\rangle, \langle 101\ldots1\rangle, \ldots, \langle 1\ldots10\rangle$*. We can always find a conjunction with the desired label. Namely,* $x_1 \wedge \ldots \wedge x_i^{\vee} \wedge \ldots \wedge x_n$ *or its negation as needed.*

Overloaded Notation

$$\begin{aligned} \Pi_{\mathcal{C}}(m) &:= \max_{S\,:|S|=m} \Big\{ \big|\Pi_{\mathcal{C}}(S)\big| \Big\} && (1.1) \\ &= \text{max. \# of labelings induced by } \mathcal{C} \text{ on m points.} && (1.2) \end{aligned}$$

Note that we have the following:

- $\Pi_{\mathcal{C}}(m) \leq 2^m$.
- If $m \leq VCD(\mathcal{C})$, $\Pi_{\mathcal{C}}(m) = 2^m$.
- If $m > VCD(\mathcal{C})$, $\Pi_{\mathcal{C}}(m) < 2^m$.

Theorem 1.5. *Let* $VCD(\mathcal{C}) = d$*. Then* $\Pi_{\mathcal{C}}(m) = O(m^d)$*.*

Remark 1.6. *As the sample size becomes large compared to the dimension, the growth rate of* $\Pi_{\mathcal{C}}(m)$ *is sub-exponential. In fact, the fraction of possible behaviors grows like* $\frac{\frac{1}{m^d}}{2^m}$ *which tends to 0 as* $m \to \infty$*.*

Proof of Theorem 1.5

Consider $\phi_d(m) := \phi_d(m-1) + \phi_{d-1}(m-1)$ with boundary conditions $\phi_0(m) = \phi_d(0) = 1$. We first prove the following lemma.

Lemma 1.7. *If $VCD(\mathcal{C}) = d$, then for any m we have $\Pi_{\mathcal{C}}(m) \leq \phi_d(m)$.*

Proof. We use a double induction argument. The base cases are clear so assume the lemma is true for any $d' \leq d$ and $m' \leq m$ with at least one the inequalities being $<$. Consider the set $S = \{x_1, \ldots, x_m\}$. Recall that $\Pi_{\mathcal{C}}(S)$ is just the possible labelings of S that can be realized with concepts in $\mathcal{C}$. Now consider all the possible labelings that are possible for $\{x_1, \ldots, x_{m-1}\} = S \setminus \{x_m\}$. For each of these, there are exactly 2 potential labelings for S depending on whether $x_m = 0$ or $x_m = 1$. Then, by the induction hypothesis, we have

$$\left|\Pi_{\mathcal{C}}(S \setminus \{x_m\})\right| \leq \Pi_{\mathcal{C}}(m-1) \leq \phi_d(m-1). \tag{1.3}$$

Next consider the labelings of $S \setminus \{x_m\}$ that appear twice in the labelings of S. That is, those such that both $x_m = 0$ and $x_m = 1$ are possible labelings. Then we can shatter at most $d-1$ points with $S \setminus \{x_m\}$ because otherwise we would violate $VCD(\mathcal{C}) = d$. Finally, by induction hypothesis we have that

$$\#\text{ of labelings of } S \setminus \{x_m\} \;\; \leq \phi_{d-1}(m-1). \tag{1.4}$$

Combining these two bound we obtain the result. □

Claim 1.8. $\phi_d(m) = \sum\limits_{i=0}^{d} \binom{m}{i}$.

Proof. By induction,

$$\begin{aligned}
\phi_d(m) = \phi_d(m-1) + \phi_{d-1}(m-1) &= \sum_{i=0}^{d} \binom{m-1}{i} + \sum_{i=0}^{d-1} \binom{m-1}{i} && (1.5)\\
&= \sum_{i=0}^{d} \binom{m-1}{i} + \sum_{i=0}^{d} \binom{m-1}{i-1} && (1.6)\\
&= \sum_{i=0}^{d} \left[\binom{m-1}{i} + \binom{m-1}{i} \right] = \sum_{i=0}^{d} \binom{m}{i}. && (1.7)
\end{aligned}$$

□

Now the proof of Theorem 1.5 follows by noting that $\sum\limits_{i=0}^{d} \binom{m}{i} = O(m^d)$.

1.1 Combinatorial → Probabilistic

Fix target $c \in \mathcal{C}$ and a distribution D. Thinking of c as a set, we can consider $\mathcal{H} \triangle c = \{h \triangle c : h \in \mathcal{H}\}$, which are the *error regions*.

Definition 1.9. We say that S of m points is a *ε-net* for $\mathcal{H}\triangle c$ if for all $r \in \mathcal{H}\triangle c$ such that $D[r] \geq \varepsilon$, then we have $S \cap r \neq \emptyset$.

Remark 1.10. *$VCD(\mathcal{H}\triangle c) = VCD(\mathcal{H})$. This is because taking the symmetric difference corresponds to a XOR operation.*

Consider the auxiliary function on $S = \{x_1, \dots, x_m\}$

$$A(S) \coloneqq \begin{cases} 1 & \text{if } S \text{ is not an } \varepsilon\text{-net for } \mathcal{H}\triangle c \\ 0 & \text{else} \end{cases}, \tag{1.8}$$

$$B\left(S, S'\right) \coloneqq \begin{cases} 1 & \text{if } \exists r \in \mathcal{H}\triangle c \text{ st. } D[r] \geq \varepsilon,\, r \cap S = \emptyset, |r \cap S'| \geq \frac{\varepsilon m}{2} \\ 0 & \text{else} \end{cases}. \tag{1.9}$$

Remark 1.11. *Event $B\left(S, S'\right)$ occurs if $\exists r \in \mathcal{H}\triangle c$ st. r is "hit":*

- *$l \geq \frac{\varepsilon m}{2}$ times in $S \cup S'$ but* all *l fall in S'.*

$$\begin{aligned} \mathbb{P}_{S,S'}[B(S,S')] &= \mathbb{P}_{S,S'}[B(S,S') \mid A(S)] \cdot \mathbb{P}_{S,S'}[A(S)] & (1.10) \\ &\geq \frac{1}{2}\mathbb{P}_{S,S'}[A(S)], & (1.11) \end{aligned}$$

where we used Chebyshev's inequality on the last line.

Draw $\langle S, S' \rangle$ of $2m$ points in 2 steps:

1. Draw $2m$ points randomly from the distribution D. $x_1, \dots, x_m, x_{m+1}, \dots, x_{2m}$

   - # possible labelings of the $2m$ points by $\mathcal{H}\triangle c$ fixed $\leq \phi_d(2m)$.

2. Split the $2m$ points randomly into S and S'.

Claim 1.12. $\mathbb{P}_{permutation}[B\left(S, S'\right)] = \frac{\binom{m}{l}}{\binom{2m}{l}} \leq 2^{-l}$.

Proof. Simple exercise in combinatorics. □

Then we have that

$$\mathbb{P}[B\left(S, S'\right)] \leq \phi_d(2m)2^{-\frac{\varepsilon m}{2}} \leq C_0(2m)^d 2^{-\frac{\varepsilon m}{2}} \leq e^{C_1[d\log(m) - \varepsilon m]}. \tag{1.12}$$

We have just proven the following theorem.

Theorem 1.13. *As long as $m \geq C_2\left(\frac{1}{\varepsilon}\log\left(\frac{1}{\delta}\right) + \frac{d}{\varepsilon}\log\left(\frac{1}{\varepsilon}\right)\right)$, where $VCD(\mathcal{H}) = d$, we have that with probability $\geq 1 - \delta$, any* consistent *$h \in \mathcal{H}$ has error $\leq \varepsilon$ with respect to c and D.*